

STATISTICAL ANALYSIS AND DESIGN OF CROWDSOURCING APPLICATIONS

Adam Kapelner

A DISSERTATION

in

Statistics

For the Graduate Group in
Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

Abba Krieger
Robert Steinberg Professor of
Statistics and Operations Research

Co-Supervisor of Dissertation

Ed George
Universal Furniture
Professor of Statistics

Graduate Group Chairperson

Eric Bradlow
K.P. Chao Professor, Marketing,
Statistics and Education

Dissertation Committee

Abba Krieger, Professor
Ed George, Professor

Larry Brown, Professor
Dean Foster, Professor

UMI Number: 3622073

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3622073

Published by ProQuest LLC (2014). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

STATISTICAL ANALYSIS AND DESIGN OF CROWDSOURCING APPLICATIONS

COPYRIGHT © 2014

Adam Kapelner

Acknowledgments

I would like to first and foremost thank my parents and sisters. Without their constant support and love, none of this would have been possible. I thank my close friends Philip Ernst and Dana Chandler for being there for me. I thank my graduate student colleagues most notably Justin Bleich and Alex Goldstein, not only for coauthoring many of these thesis chapters, but for all the fun we've had over the years. I also thank my advisors Abba Krieger, Ed George and Dean Foster for their support and guidance. I also want to thank Larry Brown and Paul Rosenbaum who have been a tremendous help over the years on different stages of my thesis. I would also like to thank Andrea Troxel and Kate Propert from the UPenn Biostatistics department for their collaboration and mentoring. I also acknowledge the National Science Foundation for support from their Graduate Research Fellowship Program.

ABSTRACT

STATISTICAL ANALYSIS AND DESIGN OF CROWDSOURCING APPLICATIONS

Adam Kapelner

Abba Krieger

Ed George

This thesis develops methods for the analysis and design of crowdsourced experiments and crowdsourced labeling tasks. Much of this document focuses on applications including running natural field experiments, estimating the number of objects in images and collecting labels for word sense disambiguation. Observed shortcomings of the crowdsourced experiments inspired the development of methodology for running more powerful experiments via matching on-the-fly. Using the label data to estimate response functions inspired work on non-parametric function estimation using Bayesian Additive Regression Trees (BART). This work then inspired extensions to BART such as incorporation of missing data as well as a user-friendly R package.

Contents

1	Introduction	1
2	Motivation in Crowdsourcing Markets	7
2.1	Introduction	8
2.2	Mechanical Turk and its potential for field experimentation	11
2.3	Experimental Design	15
2.4	Experimental Results and Discussion	22
2.5	Conclusion	31
3	Preventing Satisficing in Surveys	33
3.1	Introduction	34
3.2	Methods	36
3.3	Results	46
3.4	Discussion	55
3.5	Future directions	57
4	Detecting Heterogeneous Effects via Crowdsourcing	62
4.1	Introduction	62
4.2	Experimental Methods and Design	66
4.3	Results	74
4.4	Conclusions and Future Directions	77
5	Matching on-the-fly in Sequential Experiments	78
5.1	Introduction	78
5.2	The Algorithm, Estimation, and Testing	82
5.3	Simulation Studies	91
5.4	Demonstration Using Real Data	97
5.5	Discussion	101

6	Estimating the Number of Objects in Images	107
6.1	Introduction	107
6.2	Crowdsourcing Object Identification in Images	108
6.3	Engineering	110
6.4	Statistical Model and Implementation	116
6.5	Experiment and Results	121
6.6	Conclusion	134
7	Collecting labels for Word Sense Disambiguation	136
7.1	Introduction	137
7.2	Methods and data collection	138
7.3	Results and data analysis	141
7.4	Conclusion	148
8	Bayesian Additive Regression Trees Implementation	149
8.1	Introduction	149
8.2	Overview of BART	151
8.3	The <code>bartMachine</code> package	158
8.4	Regression Features	164
8.5	Classification Features	185
8.6	Discussion	189
9	Incorporating Missingness into BART	191
9.1	Introduction	192
9.2	Background	194
9.3	Missing Incorporated in Attributes within BART	203
9.4	Generated Data Simulations	206
9.5	Real Data Example	212
9.6	Discussion	216
A	Appendices	218
A.1	Supplement for Chapter 2	218
A.2	Supplement for Chapter 4	230
A.3	Supplement for Chapter 5	235
A.4	Supplement for Chapter 8	245
A.5	BART for Panel Data	256
	Bibliography	279

List of Tables

2.1	Summary statistics for response variables and demographics	23
2.2	A heatmap illustration of our results	24
2.3	Results for the treatment effects on quantity of images	25
2.4	Results for the treatment effects on quality of labeling	29
3.1	Overview of treatments and how they improve data quality	38
3.2	Summary statistics by treatment	47
3.3	Attrition by treatment	48
3.4	IMC pass rate by treatment (no covariates)	49
3.5	IMC pass rate by treatment (with covariates)	50
3.6	Results for Question A by treatment	52
3.7	Results for Question B by treatment	56
4.1	Power by number of duplicates and interaction effect size	73
4.2	Experimental sample sizes	74
4.3	Experimental estimates of ATE	75
5.1	Response models for the three scenarios proposed	91
5.2	Balance and relative sample efficiency results	95
5.3	Simulated sizes for all scenarios, competitors, and tests	97
5.4	Results for the sequential matching procedure for the behavioral data	99
5.5	Results for the sequential matching procedure for the clinical trial data	102
6.1	Averages and standard errors of $F1$ scores over all workers	124
6.2	Results for true positives being beta-distributed	126
6.3	The raw features used to discriminate a cluster	129
7.1	The OntoNotes sample of 89 words used in this study	140
7.2	Regression results of correctness on word, context, and senses features	143
7.3	Accuracy using plurality voting for different numbers of Turkers	146

9.1	MDM models in the context of statistical learning	197
9.2	Missingness scenarios for the BHD simulations	213
A.1	Average absolute bias of sequential matching and post matching	236
A.2	Out-of-sample RMSE values for 9 datasets	255

List of Figures

2.1	Main task portal for a subject in the meaningful treatment	18
2.2	Example preferences as a function of task meaningfulness	21
3.1	Display of Question B for all treatments	42
3.2	Screenshot of the <i>instructional manipulation check</i>	43
4.1	Illustration of samples being drawn from the population	63
4.2	External validity versus internal validity tradeoff	64
4.3	Splash page of the experimental HIT	67
4.4	Part of the survey page in the experimental HIT	67
4.5	Diagram of the multi-stage experiment layout	69
4.6	The screen shown to subjects in the “S” wing post survey	71
4.7	The design matrix for each of the k studies.	72
4.8	Box and whisker plots for the framing study	75
4.9	Sample proportions of cooperate in the priming study	76
4.10	Box and whisker plots for the sunk cost	76
5.1	Results for power at $\alpha = 0.05$ with matching propensity $\lambda = 10\%$	94
5.2	Illustration of the sequential procedure on historical data	101
6.1	The <code>DistributeEyes</code> Components and their interactions	110
6.2	<code>DistributeEyes</code> Project Settings Dialog Window	111
6.3	Worker submission check window	112
6.4	The training video and example quiz questions	114
6.5	The HTML training application within MTurk	115
6.6	Cropped selections from the project images	122
6.7	Putative beta fit and Q-Q for the beta distribution	125
6.8	Posterior of N for each project (truth assumed unknown)	127
6.9	Posterior of N for each project (truth assumed known)	131
6.10	Accuracy regressed on training path length and time spent training	132

7.1	An example of the WSD task that appears inside an MTurk HIT . . .	141
7.2	Predicted accuracy vs. number of senses for a sample of the words . . .	144
7.3	Accuracy of all 595 Turkers in this study	147
8.1	Model creation times as a function of sample size	161
8.2	Summary for the <code>bartMachine</code> model on the automobile data	166
8.3	Out-of-sample predictive performance by number of trees	167
8.4	Test of normality of errors	170
8.5	Convergence diagnostics for the cross-validated <code>bartMachine</code> model . . .	171
8.6	Fitted versus actual response values for the automobile dataset	173
8.7	Average variable inclusion proportions in the automobile data	174
8.8	Tests of covariate importance in the automobile dataset	176
8.9	Partial Dependence Plots for two features in the automobile dataset . . .	178
8.10	Visualization of the three variable selection procedures	182
8.11	The top 10 average variable interaction counts	185
8.12	Test of covariate importance for a Diabetes predictor	188
8.13	Partial dependence plot for a Diabetes predictor	189
9.1	Posterior draws for a variety of new observations	207
9.2	Simulation results of the response model for the three MDM's	211
9.3	Simulations of oosRMSE by different probabilities of missingness	215
A.1	The HIT as initially encountered on MTurk	218
A.2	The colorblindness test	219
A.3	Opening screen of training video	220
A.4	Examples of meaningful cues	221
A.5	Describing the training process	221
A.6	The quiz after watching the training video	223
A.7	The training interface in all treatments	224
A.8	The landing page after a labeling task is completed	225
A.9	The survey after completion of the HIT	226
A.10	Power illustrated for matching parameter $\lambda = 1\%$	237
A.11	Power illustrated for matching parameter $\lambda = 2.5\%$	238
A.12	Power illustrated for matching parameter $\lambda = 5\%$	239
A.13	Power illustrated for matching parameter $\lambda = 7.5\%$	240
A.14	Power illustrated for matching parameter $\lambda = 10\%$	241
A.15	Power illustrated for matching parameter $\lambda = 20\%$	242
A.16	Power illustrated for matching parameter $\lambda = 35\%$	243
A.17	Power illustrated for matching parameter $\lambda = 50\%$	244

Introduction

The four parts of this thesis were inspired by the work I have done in the modern phenomenon known as “crowdsourcing.” What is crowdsourcing? Imagine a world where you can release a small task to an anonymous person located anywhere in the world. This concept has experienced exponential growth since its inception in both industry and in academia with the largest platform being Amazon’s Mechanical Turk (MTurk). In my observations over the past seven years, academic use of crowdsourcing falls into two main categories:

- (a) *Experimentation* — Using the crowd to run an experiment to elicit a causal effect of an experimental condition on a prespecified outcome measure. With proper Institutional Review Board approval, these experiments can be natural field experiments, where the participants do not know they are part of an experiment with a randomized experimental condition. A worker is only allowed to complete one experimental task.
- (b) *Label Collection* — Using the crowd to do a prespecified task such as mark an image, or determine the meaning of a word in a paragraph, write about a musical melody, etc. In this task, a worker is allowed to provide as many labels as he wishes.

This thesis and related research is motivated by my interest in understanding these two uses of crowdsourcing. As a result of the various topics explored, it is split into four parts:

- (I) sequential experiments, Chapters 2, 3 and 4
- (II) methodology that improves sequential experiments, Chapter 5
- (III) label collection, Chapters 6 and 7
- (IV) applied machine learning research, Chapters 8 and 9.

My contribution to *Experimentation* is covered in parts I and II of this thesis; *Label Collection* is addressed in parts III and IV.

The majority of the work in this document is joint with my colleagues and each chapter's contributors are appropriately marked at the beginning of the chapter and project-specific acknowledgements are appropriately marked at the close of each chapter.

Part I

Here, I write about the work I did in *sequential experiments*, meaning that subjects enter over time and must be immediately assigned a treatment.

Chapter 2 (published in Chandler and Kapelner, 2013) conducts the first natural field experiment to explore the relationship between the “meaningfulness” of a task and worker effort. We employed about 2,500 workers from Amazon’s Mechanical Turk (MTurk, an online labor market which we introduce later in the introduction) to label medical images. Although given an identical task, we experimentally manipulated how the task was framed. Subjects in the *meaningful* treatment were told that they were labeling tumor cells in order to assist medical researchers, subjects in the *zero-context* condition (the control group) were not told the purpose of the task, and, in

stark contrast, subjects in the *shredded* treatment were not given context and were additionally told that their work would be discarded. We found that when a task was framed more meaningfully, workers were more likely to participate. We also found that the meaningful treatment increased the quantity of output (with an insignificant change in quality) while the shredded treatment decreased the quality of output (with no change in quantity). We believe these results will generalize to other short-term labor markets. This work also discusses MTurk as an exciting platform for running natural field experiments in economics.

The experiment above made use of surveys for its post manipulation check. People are known to cheat or “satisfice” on surveys, especially when not being watched such as the case over the Internet on MTurk. Without adequate controls, researchers should be concerned that respondents may fill out surveys haphazardly in the unsupervised environment of the Internet. Chapter 3 (published in Kapelner and Chandler, 2010) presents a question-presentation method, called *Kapcha*, which fades in words slowly. This experiment demonstrates technology that reduces satisficing, thereby improving the quality of survey results. This work also presents an open-source platform for further survey experimentation on MTurk.

I then decided to use Mechanical Turk to attempt to detect heterogeneous treatment effects between an experimental subject that chooses to be in a study versus those who are randomly assigned to a study. We call this “selection-into-experiment bias” and the work found in Chapter 4 tests its effects in three experiments run on MTurk. We find no differential average treatment effects in the studies, an unexpected result. We believe the flaw in the study was lack of power: the differences in effects is very small, so even with a large sample size, no result was found.

Part II

The last experiment in Part I inspired me to think about methodology that increases power and efficiency in sequential experiments. This became part II, which consists of Chapter 5 (published in Kapelner and Krieger, 2014). Here, we propose a dynamic allocation procedure that increases power and efficiency when measuring an average treatment effect in fixed sample size sequential randomized trials. Subjects arrive iteratively and are either randomized *or* paired via a matching criterion to a previously randomized subject and administered the alternate treatment. We develop estimators for the average treatment effect that combine information from both the matched pairs and unmatched subjects as well as an exact test. Simulations illustrate the method’s higher efficiency and power over competing allocation procedures in both controlled scenarios and historical clinical trial data.

Part III

Experimentation is not common on MTurk; quick and inexpensive *label collection* is by far the dominant application.

Chapter 6 outlines a method that estimates the count of objects or features in an image such as tallying the number of birds in a photograph or the the number of cells in a microscopic image. The approach has two novel steps. We first develop software that records the labelings of many naive workers via MTurk. We then view each worker’s training as a “capture” in a set of capture-recapture experiments. We use statistical learning to eliminate falsely-trained objects and then take a Bayesian approach and use a Gibbs sampler to estimate the true number of objects.

Chapter 7 (published in Kapelner et al., 2012) demonstrates the use of MTurk to disambiguate 1000 words from among coarse-grained senses, the most extensive investigation to date. Ten unique participants disambiguate each example, and, using

regression, we find surprising features which drive differential wordsense disambiguation accuracy: (a) the number of rephasings within a sense definition is associated with higher accuracy; (b) as word frequency increases, accuracy decreases even if the number of senses is kept constant; and (c) spending more time is associated with a decrease in accuracy. We also observe that all participants are about equal in ability, practice (without feedback) does not seem to lead to improvement, and that having many participants label the same example provides a partial substitute for more expensive annotation.

Part IV

While working on projects involving label collection in Part III, my interests shifted to using labels to make predictions and I began thinking about applied machine learning research.

Chapter 8 (published in Kapelner and Bleich, 2013a) presents `bartMachine` a new package in R implementing Bayesian Additive Regression Trees (BART, Chipman et al., 2010) The package introduces many new features for data analysis using BART such as variable selection, interaction detection, model diagnostic plots, incorporation of missing data and the ability to save trees for future prediction. It is significantly faster than the current R implementation, parallelized, and capable of handling both large sample sizes and high-dimensional data.

Chapter 9 (published in Kapelner and Bleich, 2013b) presents a method for incorporating missing data into general forecasting problems that use BART. We focus on enhancing BART using “Missingness Incorporated in Attributes,” an approach recently proposed for incorporating missingness into decision trees. This procedure extends the native partitioning mechanisms found in tree-based models and does not require imputation. Simulations on Rgenerated models and real data indicate that this pro-

cedure offers promise for both selection model and pattern mixture frameworks as measured by out-of-sample predictive accuracy. We also illustrate BART's abilities to incorporate missingness into uncertainty intervals. The implementation outlined here has been incorporated into the R package `bartMachine`.

I now outline further work which I have collaborated on within the domain of applied machine learning which is not included in the body of this document.

An important problem when building forecasting engines using labeling data is to understand which of the collected features drive the engine. Bleich et al. (2014) outline a procedure based on permutation testing to find features that drive the response function. Further, Bleich and Kapelner (2014) outline a procedure that generalizes BART to fit heteroskedastic models by incorporating linear models of heteroskedasticity. Many times in crowdsourcing applications, workers complete more than one labeling task. This induces a correlation structure in the design matrix. Extending BART to accommodate this type of panel data is a work in progress. The current implementation details are found in Appendix A.5.

It became frustrating to be a user of BART, a black-box algorithm, without being able to “look inside” and understand the inner workings of the model fit. To this end, I worked on a visualization procedure which highlight the variation in the fitted values across the range of a covariate, suggesting where and to what extent heterogeneities might exist. This work and its implementation in R as the package “ICEbox” is found in Goldstein et al. (2014).

Motivation in Crowdsourcing Markets*

Abstract

We conduct the first natural field experiment to explore the relationship between the “meaningfulness” of a task and worker effort. We employed about 2,500 workers from Amazon’s Mechanical Turk (MTurk), an online labor market, to label medical images. Although given an identical task, we experimentally manipulated how the task was framed. Subjects in the *meaningful* treatment were told that they were labeling tumor cells in order to assist medical researchers, subjects in the *zero-context* condition (the control group) were not told the purpose of the task, and, in stark contrast, subjects in the *shredded* treatment were not given context and were additionally told that their work would be discarded. We found that when a task was framed more meaningfully, workers were more likely to participate. We also found that the meaningful treatment increased the quantity of output (with an insignificant change in quality) while the shredded treatment decreased the quality of output (with no change in quantity). We believe these results will generalize to other short-term labor markets. Our study also discusses MTurk as an exciting platform for running natural field experiments in economics.

*Joint work with Dana Chandler

2.1 Introduction

Economists, philosophers, and social scientists have long recognized that non-pecuniary factors are powerful motivators that influence choice of occupation. For a multidisciplinary literature review on the role of meaning in the workplace, we recommend Rosso et al. (2010). Previous studies in this area have generally been based on ethnographies, observational studies, or laboratory experiments. For instance, Wrzesniewski et al. (1997) used ethnographies to classify work into jobs, careers, or callings. Using an observation study, Preston (1989) demonstrated that workers may accept lower wages in the non-profit sector in order to produce goods with social externalities. Finally, Ariely et al. (2008) showed that labor had to be both recognizable and purposeful to have meaning. In this paper, we limit our discussion to the role of meaning in economics, particularly through the lens of competing differentials. We perform the first *natural field experiment* (Harrison and List, 2004) in a real effort task that manipulates levels of meaningfulness. This method overcomes a number of shortcomings of the previous literature, including: interview bias, omitted variable bias, and concerns of external validity beyond the laboratory.

We study whether employers can deliberately alter the perceived “meaningfulness” of a task in order to induce people to do more and higher quality work and thereby work for a lower wage. We chose a task that would appear meaningful for many people if given the right context — helping cancer researchers mark tumor cells in medical images. Subjects in the *meaningful* treatment were told the purpose of their task is to “help researchers identify tumor cells;” subjects in our *zero-context* group were not given any reason for their work and the cells were instead referred to as mere “objects of interest” and laborers in the *shredded* group were given zero context but also explicitly told that their labelings would be discarded upon submission. Hence, the pay structure, task requirements, and working conditions were identical, but we

added cues to alter the perceived meaningfulness of the task.

We recruited workers from the United States and India from Amazon’s Mechanical Turk (MTurk), an online labor market where people around the world complete short, “one-off” tasks for pay. The MTurk environment is a spot market for labor characterized by relative anonymity and a lack of strong reputational mechanisms. As a result, it is well-suited for an experiment involving the meaningfulness of a task since the variation we introduce regarding a task’s meaningfulness is less affected by desires to exhibit pro-social behavior or an anticipation of future work (career concerns). We ensured that our task appeared like any other task in the marketplace and was comparable in terms of difficulty, duration, and wage.

Our study is representative of the kinds of natural field experiments for which MTurk is particularly suited. Section 2.2.2 explores MTurk’s potential as a platform for field experimentation using the framework proposed in Levitt and List (2007, 2009).

We contribute to the literature on compensating wage differentials (Rosen, 1986) and the organizational behavioral literature on the role of meaning in the workplace (Rosso et al., 2010). Within economics, Stern (2004) provides quasi-experimental evidence on compensating differentials within the labor market for scientists by comparing wages for academic and private sector job offers among recent Ph.D. graduates. He finds that “scientists pay to be scientists” and require higher wages in order to accept private sector research jobs because of the reduced intellectual freedom and a reduced ability to interact with the scientific community and receive social recognition. Ariely et al. (2008) use a laboratory experiment with undergraduates to vary the meaningfulness of two separate tasks: (1) assembling Legos and (2) finding 10 instances of consecutive letters from a sheet of random letters. Our experiment augments experiment 1 in Ariely et al. (2008) by testing whether their results extend to the field. Additionally, we introduce a richer measure of task effort, namely *task*

quality. Where our experiments are comparable, we find that our results parallel theirs.

We find that the main effects of making our task more meaningful is to induce a higher fraction of workers to complete our task, hereafter dubbed as “induced to work.” In the meaningful treatment, 80.6% of people labeled at least one image compared with 76.2% in the zero-context and 72.3% in the shredded treatments.

After labeling their first image, workers were given the opportunity to label additional images at a declining piecerate. We also measure whether the treatments increase the quantity of images labeled. We classify participants as “high-output” workers if they label five or more images (an amount corresponding to roughly the top tercile of those who label) and we find that workers are approximately 23% more likely to be high-output workers in the meaningful group.

We introduce a measure of task quality by telling workers the importance of accurately labeling each cell by clicking as close to the center as possible. We first note that MTurk labor is high quality, with an average of 91% of cells found. The meaning treatment had an ambiguous effect, but the shredded condition in both countries lowered the proportion of cells found by about 7%.

By measuring both quantity and quality we are able to observe how task effort is apportioned between these two “dimensions of effort.” Do workers work “harder” or “longer” or both? We found an interesting result: the meaningful condition seems to increase quantity without a corresponding increase in quality and the shredded treatment decreases quality without a corresponding decrease in quantity. Investigating whether this pattern generalizes to other domains may be a fruitful future research avenue.

Finally, we calculate participants’ average hourly waged based on how long they spent on the task. We find that subjects in the meaningful group work for \$1.34 per hour, which is 6 cents less per hour than zero context participants and 14 cents less

per hour than shredded condition participants.

We expect our findings to generalize to other short-term work environments such as temporary employment or piecework. In these environments, employers may not consider that non-pecuniary incentives of meaningfulness matter; we argue that these incentives do matter, and to a significant degree.

Section 2.2 provides background on MTurk and discusses its use as a platform for conducting economic field experiments. Section 2.3 describes our experimental design. Section 3.3 presents our results and discussion and Section 2.5 concludes. Appendices in section A.1 provide full details on our experimental design and suggestions for conducting experiments using the MTurk platform.

2.2 Mechanical Turk and its potential for field experimentation

Amazon’s Mechanical Turk (MTurk) is the largest online, task-based labor market and is used by hundreds of thousands of people worldwide. Individuals and companies can post tasks (known as Human Intelligence Tasks, or “HITs”) and have them completed by an on-demand labor force. Typical tasks include image labeling, audio transcription, and basic internet research. Academics also use MTurk to outsource low-skilled resource tasks such as identifying linguistic patterns in text (Sprouse, 2011) and labeling medical images (Chapter 6 of this document). The image labeling system from the latter study, known as “DistributeEyes,” was originally used by breast cancer researchers and was modified for our experiment.

Beyond simply using MTurk as a source of labor, academics have also begun using MTurk as a way to conduct online experiments. The remainder of the section highlights some of the ways this subject pool is used and places special emphasis on the suitability of the environment for natural field experiments in economics.

2.2.1 General use by social scientists

As Henrich et al. (2010) argue, many findings from social science are disproportionately based on what he calls “W.E.I.R.D.” subject pools (**W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic) and as a result it is inappropriate to believe the results generalize to larger populations. Since MTurk has users from around the world, it is also possible to conduct research across cultures. For example, Eriksson and Simpson (2010) use a cross-national sample from MTurk to test whether differential preferences for competitive environments are explained by females’ stronger emotional reaction to losing, hypothesized by Croson and Gneezy (2009).

It is natural to ask whether results from MTurk generalize to other populations. Paolacci et al. (2010a) assuage these concerns by replicating three classic framing experiments on MTurk: The Asian Disease Problem, the Linda Problem and the Physician Problem; Horton et al. (2011) provide additional replication evidence for experiments related to framing, social preferences, and priming. Berinsky et al. (2012) argues that the MTurk population has “attractive characteristics” because it approximates gold-standard probability samples of the US population. All three studies find that the direction and magnitude of the effects line up well compared with those found in the laboratory.

An advantage of MTurk relative to the laboratory is that the researcher can rapidly scale experiments and recruit hundreds of subjects within only a few days and at substantially lower costs.²

²For example, in our study we paid 2,471 subjects \$789 total and they worked 701 hours (equating to 31 cents per observation). This includes 60 subjects whose data were not usable.

2.2.2 Suitability for natural field experiments in Economics

Apart from general usage by academics, the MTurk environment offers additional benefits for experimental economists and researchers conducting natural field experiments. We analyze the MTurk environment within the framework laid out in Levitt and List (2007, 2009).

In the ideal natural field experiment, “the environment is such that the subjects naturally undertake these tasks and [do not know] that they are participants in an experiment.” Additionally, the experimenter must exert a high degree of control over the environment without attracting attention or causing participants to behave unnaturally. MTurk’s power comes from the ability to construct customized and highly-tailored environments related to the question being studied. It is possible to collect very detailed measures of user behavior such as precise time spent on a webpage, mouse movements, and positions of clicks. In our experiment, we use such data to construct a precise quality measure.

MTurk is particularly well-suited to using experimenter-as-employer designs (Gneezy and List, 2006) as a way to study worker incentives and the employment relationship without having to rely on cooperation of private sector firms.³ For example, Barankay (2010) posted identical image labeling tasks and varied whether workers were given feedback on their relative performance (i.e., ranking) in order to study whether providing rank-order feedback led workers to return for a subsequent work opportunity. For a more detailed overview of how online labor markets can be used in experiments, see Horton et al. (2011).

Levitt and List (2007) enumerate possible complications that arise when experimental findings are extrapolated outside the lab: *scrutiny*, *anonymity*, *stakes*, *selection*, and *artificial restrictions*. We analyze each complication in the context of our

³Barankay (2010) remarks that “the experimenter [posing] as the firm [gives] substantial control about the protocol and thereby eliminates many project risks related to field experiments.

experiment and in the context of experimentation using MTurk in general.

Scrutiny and anonymity. In the lab, experimenter effects can be powerful; subjects behave differently if they are aware their behavior is being watched. Relatedly, subjects frequently lack anonymity and believe their choices will be scrutinized after the experiment. In MTurk, interaction between workers and employers is almost non-existent; most tasks are completed without any communication and workers are only identifiable by a numeric identifier. Consequently, we believe that MTurk experiments are less likely to be biased by these complications.

Stakes. In the lab or field, it's essential to "account properly for the differences in stakes across settings" (Levitt and List, 2007). We believe that our results would generalize to other short-term work environments, but would not expect them to be generalizable to long-term employment decisions such as occupational choice. Stakes must also be chosen adequately for the environment and so we were careful to match wages to the market average.

Selection. Experiments fail to be generalizable when "participants in the study differ in systematic ways from the actors engaged in the targeted real-world setting." We know that within MTurk, it is unlikely that there is selection into our experiment since our task was designed similar in appearance to real tasks. The MTurk population also seems representative along a number of observable demographic characteristics (Berinsky et al., 2012); however, we acknowledge that there are potentially unobservable differences between our subject pool and the broader population. Still, we believe that MTurk subject behavior would generalize to workers' behavior in other short-term labor markets.

Artificial restrictions. Lab experiments place unusual and artificial restrictions on the actions available to subjects and they examine only small, non-representative windows of time because the experimenter typically doesn't have subjects and time horizons for an experiment. In structuring our experiment, workers had substan-

tial latitude in how they performed their task. In contrast with the lab, subjects could “show-up” to our task whenever they wanted, leave at will, and were not time-constrained. Nevertheless, we acknowledge that while our experiment succeeded in matching short-term labor environments like MTurk, that our results do not easily generalize to longer-term employment relationships.

Levitt and List (2009) highlight two limitations of field experiments vis-a-vis laboratory experiments: the *need for cooperation* with third parties and the difficulty of *replication*. MTurk does not suffer from these limitations. Work environments can be created by researchers without the need of a private sector partner, whose interests may diverge substantially from that of the researcher. Further, MTurk experiments can be replicated simply by downloading source code and re-running the experiment. In many ways, this allows a push-button replication that is far better than that offered in the lab.

2.3 Experimental Design

2.3.1 Subject recruitment

In running our randomized natural field experiment, we posted our experimental task so that it would appear like any other task (image labeling tasks are among the most commonly performed tasks on MTurk). Subjects had no indication they were participating in an experiment. Moreover, since MTurk is a market where people ordinarily perform one-off tasks, our experiment could be listed inconspicuously.

We hired a total of 2,471 workers (1,318 from the US and 1,153 from India). Although we tried to recruit equally from both countries, there were fewer Indians in our sample since attrition in India was higher. We collected each worker’s age and gender during a “colorblindness” test that we administered as part of the task. These

and other summary statistics can be found in Table 3.2. By contracting workers from the US and India, we can also test whether workers from each country respond differentially to the meaningfulness of a task.

Our task was presented so that it appeared like a one-time work opportunity (subjects were barred from doing the experiment more than once) and our design sought to maximize the amount of work we could extract during this short interaction. The first image labeling paid \$0.10, the next paid \$0.09, etc, leveling off at \$0.02 per image. This wage structure was also used in Ariely et al. (2008) and has the benefit of preventing people from working too long.

2.3.2 Description of experimental conditions

Upon accepting our task, workers provided basic demographic information and passed a color-blindness test. Next, they were randomized into either the *meaningful*, the *zero-context*, or the *shredded* condition. Those in the shredded condition were shown a warning message stating that their labeling will not be recorded and we gave them the option to leave. Then, all participants were forced to watch an instructional video which they could not fast-forward. See Appendix A.1.1 for the full script of the video as well as screenshots.

The video for the meaningful treatment began immediately with cues of meaning. We adopt a similar working definition of “meaningfulness” as used in Ariely et al. (2008): “Labor [or a task] is meaningful to the extent that (a) it is recognized and/or (b) has some point or purpose.”

We varied the levels of meaningfulness by altering the degree of recognition and the detail used to explain the purpose of our task. In our meaningful group, we provided “recognition” by thanking the laborers for working on our task. We then explained the “purpose” of the task by creating a narrative explaining how researchers

were inundated with more medical images than they could possibly label and that they needed the help of ordinary people. In contrast, the zero-context and shredded groups were not given recognition, told the purpose of the task, or thanked for participating; they were only given basic instructions. Analyzing the results from a post-manipulation check (see section 2.4.4), we are confident that these cues of meaning induced the desired affect.

Both videos identically described the wage structure and the mechanics of how to label cells and properly use the task interface (including zooming in/out and deleting points, which are metrics we analyze). However, in the meaningful treatment, cells were referred to as “cancerous tumor cells” whereas in the zero-context and shredded treatments, they were referred to as nondescript “objects of interest.” Except for this phrase change, both scripts were identical during the instructional sections of the videos. To emphasize these cues, workers in the meaningful group heard the words “tumor,” “tumor cells,” “cells,” etc. 16 times before labeling their first image and similar cues on the task interface reminded them of the purpose of the task as they labeled.

2.3.3 Task interface, incentive structure, and response variables

After the video, we administered a short multiple-choice quiz testing workers’ comprehension of the task and user interface. In the shredded condition, we gave a final question asking workers to again acknowledge that their work will not be recorded.

Upon passing the quiz, workers were directed to a task interface which displayed the image to be labeled and allowed users to mark cancerous tumor cells (or “objects of interest”) by clicking (see figure 2.1). The image shown was one of ten look-alike photoshopped images displayed randomly. We also provide the workers with controls — *zoom functionality* and the ability to *delete points* — whose proper use would allow

them to produce high-quality labelings.

During the experiment, we measured three response variables: (1) induced to work, (2) quantity of image labelings, and (3) quality of image labelings.

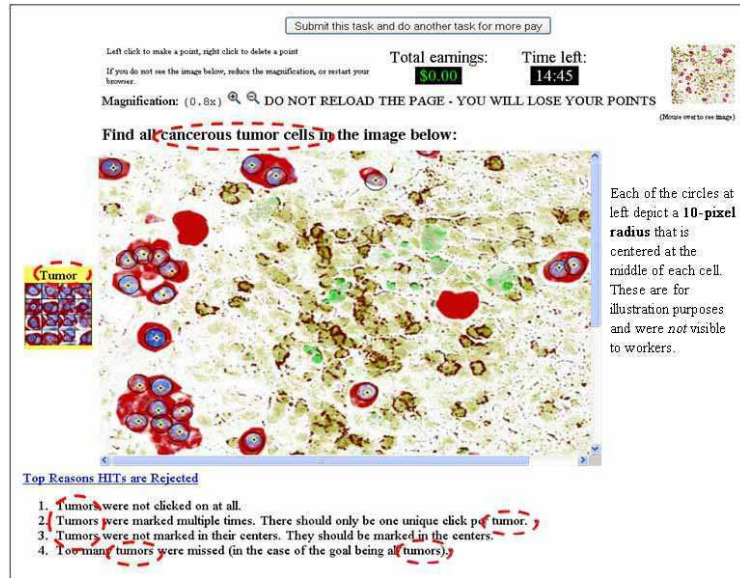


Figure 2.1: Main task portal for a subject in the meaningful treatment. Workers were asked to identify all tumors in the image. Each image had 90 cells and took 5 minutes on average. Our interface reminds the workers in 8 places that they are identifying tumor cells. The black circles around each point were *not* visible to participants. We display them to illustrate the size of a 10-pixel radius.

Many subjects can – and – do stop performing a task even after agreeing to complete it. While submitting bad work on MTurk is penalized, workers can abandon a task with only nominal penalty. Hence, we measure attrition with the response variable *induced to work*. Workers were only counted as induced to work if they watched the video, passed the quiz, and completed one image labeling. Our experimental design deliberately encourages attrition by imposing an upfront and unpaid cost of watching a three-minute instructional video and passing a quiz before moving on to the actual task.

Workers were paid \$0.10 for the first image labeling. They were then given an option to label another image for \$0.09, and then another image for \$0.08, and so on.⁴ At \$0.02, we stopped decreasing the wage and the worker was allowed to label images at this pay rate indefinitely. After each image, the worker could either collect what they had earned thus far, or label more images. We used the *quantity of image labelings* for our second response variable.

In our instructional video, we emphasized the importance of marking the exact center of each cell. When a worker labeled a cell by clicking on the image, we measured that click location to the nearest pixel. Thus, we were able to detect if the click came “close” to the actual cell. Our third response variable, *quality of image labelings* is the proportion of objects identified based on whether a worker’s click fell within a pixel radius from the object’s true center. We will discuss the radii we picked in the following section.

After workers chose to stop labeling images and collect their earnings, they were given a five-question PMC survey which asked whether they thought the task (a) was enjoyable (b) had purpose (c) gave them a sense of accomplishment (d) was meaningful (e) made their efforts recognized. Responses were collected on a five-point Likert scale. We also provided a text box to elicit free-response comments.⁵

2.3.4 Hypotheses

Hypothesis 1 We hypothesize that at equal wages, the meaningful treatment will have the highest proportion of workers induced to work and the shredded condition will have the lowest proportion. In the following section, we provide theoretical justification for this prediction.

⁴Each image was randomly picked from a pool of ten look-alike images.

⁵About 24% of respondents left comments (no difference across treatments).

Hypothesis 2 As in Ariely et al. (2008), we hypothesize that *quantity* of images labeled will be increasing in the level of meaningfulness.

Hypothesis 3 In addition to quantity, we measure the *quality* of image labelings and hypothesize that this is increasing in the level of meaningfulness.

Hypothesis 4 Based upon prior survey research on MTurk populations, we hypothesize that *Indian workers are less responsive to meaning*. Ipeirotis (2010) finds that Indians are more likely to have MTurk as a primary source of income (27% vs. 14% in the US). Likewise, people in the US are nearly twice as likely to report doing tasks because they are fun (41% vs. 20%). Therefore, one might expect financial motivations to be more important for Indian workers.⁶

2.3.5 Theoretical Framework

Hypothesis 1 is justified by modeling workers' labor participation decisions using the modified compensating differentials framework in Rosen (1986). Instead of analyzing a worker's choice between two jobs, a "dirty" and "clean" one, the worker has preferences over jobs with different levels of meaningfulness as characterized by our three treatments.⁷ Each worker is given a take-it-or-leave-it offer that includes a sequence of wages and a meaning level and workers subsequently decide whether to participate (defined as labeling one or more images).

We abstract from the multi-stage decision process and focus on the worker's initial decision. We suppose that workers have quasi-concave preferences $U(C, T)$ that are strictly increasing in the continuous variables of consumption C and the degree of

⁶Although Horton et al. (2011) find that workers of both types are strongly motivated by money.

⁷Unlike (Rosen, 1986), workers are not presented with all three options at once and we do not model the market environment since each worker only receives a take-it-or-leave-it offer.

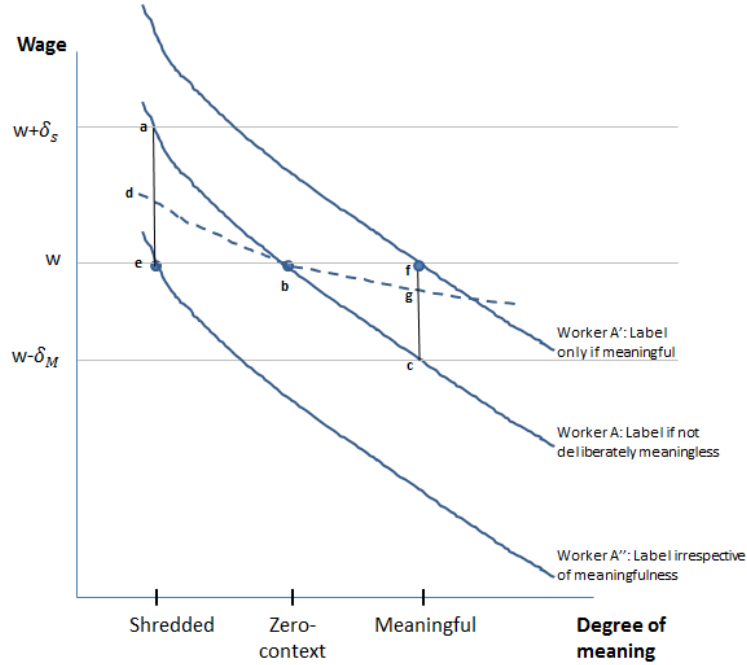


Figure 2.2: Example preferences as a function of task meaningfulness.

meaning associated with a task T . Depending on their treatment, workers are presented with one of three wage-meaning tuples (w, T) (points e , b , or f) that correspond to identical wage offers for our task with three levels of meaning ($T \in \{S, ZC, M\}$, with $M > ZC > S$). Figure 2.2 plots indifference curves representing the total utility obtained from our offer. The three indifference curves plotted for workers A , B , and C depict marginal participants in each group whose IR constraint would stop binding if the wage or meaning levels were any lower.⁸ Consequently, workers' preferences for meaningfulness induce differential participation rates.

To illustrate, Worker A (whose indifference curve intersects points a , b , and c) would participate in all but the shredded treatment since the points b and f are above and to the right of the indifference curve. The line segments ae and fc represent hypothetical compensating differentials that worker A would require to participate in either the zero-context or meaningful treatment. We also link this framework

⁸If a worker is indifferent between working or not, we assume he works.

to hypothesis 4, which discusses whether some workers may be less responsive to meaning. The flatter indifference curve represented by arc dbg shows a worker who is less responsive to meaning.

The data from our experiment help estimate the proportion of people who fall between marginal workers A , B , and C . However, we are unable to estimate the curvature of the utility function, changing costs of effort, or the extent of learning since we don't introduce additional treatments that vary wage and meaning simultaneously.⁹ In future experiments, it would be worthwhile to create new treatments that target these elements.

2.4 Experimental Results and Discussion

We ran the experiment on $N = 2,471$ subjects (1,318 from the United States and 1,153 from India). Table 3.2 shows summary statistics for our response variables (induced to work, number of images, and quality), demographic variables, and hourly wage.

Broadly speaking, as the level of meaning increases, subjects are more likely to participate and they label more images and with higher quality. Across all treatments, US workers participate more often, label more images, and mark points with greater accuracy. Table 2.2 uses a heatmap to illustrate our main effect sizes and their significance levels by treatment, country, and response variable. Each cell indicates the size of a treatment effect relative to the control (i.e., zero context condition). Statistically significant *positive* effects are indicated using green fill where darker green indicates higher levels of significance. Statistically significant *negative* effects are indicated using red fill where darker red indicates higher levels of significance. Black text without fill indicates effects that are marginally significant ($p < 0.10$).

⁹For this reason, we omit these important factors from our model.

	Shredded	Zero Context	Meaningful	US only	India only
% Induced to Work	.723	.762	.806	.85	.666
# Images (if ≥ 1)	5.94 \pm 6.8	6.11 \pm 6.9	7.12 \pm 7.6	5.86 \pm 6.1	7.17 \pm 8.3
Did ≥ 2 labelings	.696	.706	.75	.797	.627
Did ≥ 5 labelings	.343	.371	.456	.406	.373
Avg Hourly Wage	\$1.49	\$1.41	\$1.34	\$1.50	\$1.29
% Male	.616	.615	.58	.483	.743
Age	29.6 \pm 9.3	29.6 \pm 9.5	29.3 \pm 9.1	31.8 \pm 10.5	26.9 \pm 6.8
<i>N</i>	828	798	845	1318	1153
Coarse quality	.883 \pm .21	.904 \pm .18	.930 \pm .14	.924 \pm .15	.881 \pm .21
Fine quality	.614 \pm .22	.651 \pm .21	.676 \pm .18	.668 \pm .19	.621 \pm .26
PMC Meaning	3.44 \pm 1.3	3.54 \pm 1.2	4.37 \pm 0.9	3.67 \pm 1.3	3.98 \pm 1.1

Table 2.1: Summary statistics for response variables and demographics by treatment and country. The statistics for the quality metrics are computed by averaging each worker’s average quality (only for workers who labeled one or more images). The statistics for the PMC meaning question only include workers who finished the task and survey.

Light gray text indicates significance levels above 0.10.

Overall, we observe that the meaningful condition induces an increase in quantity without significantly increasing quality, and the shredded condition induces a quality decrease with quantity remaining constant. This “checkerboard effect” may indicate that meaning plays a role in moderating how workers trade quantity for quality i.e. how their energy is channeled in the task.

We now investigate each response variable individually.

	Induced to work	Did \geq 5 labelings	Fine Quality	Average Hourly Wage
Meaningful	↑ 4.6%*	↑ 8.5%***	↑ 0.7%	↓ 4.5%
Meaningful (US)	↑ 5.1%*	↑ 8.9%**	↑ 3.9%	↓ 7.7%
Meaningful (India)	↓ 2.3%	↑ 7.0%*	↓ 3.1%	↑ 0.5%
Shredded	↓ 4.0%	↓ 2.8%	↓ 7.2%***	↑ 5.6%
Shredded (US)	↓ 2.3%	↓ 5.0%	↓ 6.1%*	↑ 9.5%
Shredded (India)	↓ 6.8%	↓ 1.6%	↓ 8.7%**	↓ 1.4%

* $p < .05$, ** $p < .01$, *** $p < .001$, black text indicates $p < .10$ and grey text indicates $p > 0.10$

Table 2.2: A heatmap illustration of our results. Rows 1 and 4 consider data from both America and India combined. Columns 1, 2, 3 show the results of regressions and column 4 shows the result of two-sample t-tests. Results reported are from regressions without demographic controls.

2.4.1 Labor Participation Results: “Induced to work”

We investigate how treatment and country affects whether or not subjects chose to do our task. Unlike in a laboratory environment, our subjects were workers in a relatively anonymous labor market and were not paid a “show-up fee.” On MTurk, workers frequently start but do not finish tasks; attrition is therefore a practical concern for employers who hire from this market. In our experiment, on average, 25% of subjects began, but did not follow-through by completing one full labeling.

Even in this difficult environment, we were able to increase participation among workers by roughly 4.6% by framing the task as more meaningful (see columns 1 and 2 of table 2.3). The effect is robust to including various controls for age, gender, and time of day effects. As a subject in the meaningful treatment told us, “It’s always

nice to have [HITs] that take some thought and mean something to complete. Thank you for bringing them to MTurk.” The shredded treatment discouraged workers and caused them to work 4.0% less often but the effect was less significant ($p = 0.057$ without controls and $p = 0.082$ with controls). Thus, hypothesis 1 seems to be correct.

	Induced	Induced	Did ≥ 2	Did ≥ 2	Did ≥ 5	Did ≥ 5
Meaningful	0.046*	0.046*	0.047*	0.050*	0.085***	0.088***
	(0.020)	(0.020)	(0.022)	(0.022)	(0.024)	(0.024)
Shredded	-0.040	-0.037	-0.012	-0.005	-0.028	-0.023
	(0.021)	(0.021)	(0.022)	(0.022)	(0.024)	(0.024)
India	-0.185***	-0.183***	-0.170***	-0.156***	-0.035	-0.003
	(0.017)	(0.018)	(0.018)	(0.019)	(0.019)	(0.021)
Male		0.006		-0.029		-0.081***
		(0.018)		(0.019)		(0.021)
Constant	0.848***	0.907***	0.785***	0.873***	0.387***	0.460***
Controls						
Age		0.23		0.29		0.92
Time of Day		0.16		0.06		0.46
Day of Week		0.08		0.00**		0.55
R^2	0.05	0.06	0.04	0.05	0.01	0.02
N	2471	2471	2471	2471	2471	2471

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 2.3: Robust linear regression results for the main treatment effects on quantity of images. Columns 1, 3 and 5 only include treatments and country. Columns 2, 4, and 6 control for gender, age categories, time of day, and day of week. Rows 6-8 show p -values for the partial F -test for sets of different types of control variables.

Irrespective of treatment, subjects from India completed an image 18.5% less often ($p < 0.001$) than subjects from the US. We were interested in interactions between country and treatment, so we ran the separate induced-to-work regression results by country (unshown). We did not find significant effects within the individual countries because we were underpowered to detect this effect when the sample size was halved. We find no difference in the treatment effect for induced to work between India and the United States ($p = 0.97$). This is inconsistent with hypothesis 4 where we predicted Indian subjects to respond more strongly to pecuniary incentives.

It is also possible that the effects for induced to work were weak because subjects could have still attributed meaning to the zero context and shredded conditions, a problem that will affect our results for quantity and quality as well. This serves to bias our treatment effects downward suggesting that the true effect of meaning would be larger. For instance, one zero-context subject told us, “I assumed the ‘objects’ were cells so I guess that was kind of interesting.” Another subject in the zero-context treatment advised us, “you could put MTurkers to good use doing similar work with images, e.g. in dosimetry or pathology ... and it would free up medical professionals to do the heavier work.”

2.4.2 Quantity Results: Number of images labeled

Table 3.2 shows that the number of images increased with meaning. However, this result is conditional on being induced to work and is therefore contaminated with selection bias. We follow Angrist (2001) and handle selection by creating a dummy variable for “did two or more labelings” and a dummy for “did five or more labelings” and use them as responses (other cutoffs produced similar results).

We find mixed results regarding whether the the level of meaningfulness affects the quantity of output. Being assigned to the meaningful treatment group *did* have

a positive effect, but assignment to the shredded treatment did not result in a corresponding decrease in output.

Analyzing the outcome “two or more labelings,” column 3 of table 2.3 shows that the meaningful treatment induced 4.7% more subjects to label two or more images ($p < 0.05$). The shredded treatment had no effect. Analyzing the outcome “five or more labelings” (column 5), which we denote as “high-output workers,”¹⁰ the meaningful treatment was highly significant and induced 8.5% more workers ($p < 0.001$ with and without controls), an increase of nearly 23 percent, and the shredded treatment again has no effect.

Hypothesis 2 (quantity increases with meaningfulness) seems to be correct only when comparing the meaningful treatment to the zero-context treatment. An ambiguous effect of the shredded treatment on quantity is also reported by Ariely et al. (2008).

We didn’t find differential effects between the United States and India. In an unshown regression, we found that Americans were 9.5% more likely to label five or more images ($p < 0.01$) and Indians were 8.4% more likely to label five or more ($p < 0.05$). These two effects were not found to be different ($p = 0.84$) which is inconsistent with hypothesis 4 that Indians are more motivated by pecuniary incentives than Americans.

Interestingly, we also observed a number of “target-earners” who stopped upon reaching exactly one dollar in earnings. A mass of 16 participants stopped at one dollar, while one participant stopped at \$1.02 and not one stopped at \$0.98, an effect also observed by Horton et al. (2011). The worker who labored longest spent 2 hours and 35 minutes and labeled 77 images.

¹⁰Labeling five or more images corresponds to the top tercile of quantity among people who were induced to work.

2.4.3 Quality Results: Accuracy of labeling

Quality was measured by the fraction of cells labeled at a distance of five pixels (“coarse quality”) and two pixels (“fine quality”) from their true centers. In presenting our results (see table 2.4), we analyze the treatment effects using our fine quality measure. The coarse quality regression results were similar, but the fine quality had a much more dispersed distribution.¹¹

Our main result is that fine quality was 7.2% lower in the shredded treatment, but there wasn’t a large corresponding increase in the meaningful treatment.¹² This makes sense; if the workers knew their labelings weren’t going to be checked, there is no incentive to mark points carefully. This result was not different across countries (regression unshown). The meaningful treatment has a marginally significant effect only in the United States, where fine quality increased by 3.9% ($p = 0.092$ without controls and $p = 0.044$ with controls), but there was no effect in India. Thus, hypothesis 3 (quality increases with meaningfulness) seems to be correct *only* when comparing the shredded to the zero context treatment which is surprising.

Although Indian workers were less accurate than United States workers and had 5.3% lower quality ($p < 0.001$ and robust to controls), United States and Indian workers did not respond differentially to the shredded treatment ($p = 0.53$). This again is inconsistent with hypothesis 4.

Experience matters. Once subjects had between 6 and 10 labelings under their belt, they were 1.8% less accurate ($p < 0.01$), and if they had done more than 10 labelings, they were 14% less accurate ($p < 0.001$). This result may reflect negative

¹¹The inter-quartile range of coarse quality overall was [93.3%, 97.2%] whereas the IQR of fine quality was overall [54.7%, 80.0%].

¹²One caveat with our quality results is that we only observe quality for people who were induced to work and selected into our experiment (we have “attrition bias”). Attrition was 4% higher in the shredded treatment and we presume that the people who opted out of labeling images would have labeled them with far worse quality had they remained in the experiment.

	Fine Quality					
	Both Countries		United States		India	
Meaningful	0.007 (0.017)	0.014 (0.014)	0.039 (0.023)	0.039* (0.019)	-0.031 (0.025)	-0.013 (0.021)
Shredded	-0.072*** (0.021)	-0.074*** (0.017)	-0.061* (0.027)	-0.066** (0.023)	-0.087** (0.031)	-0.073** (0.023)
India	-0.053*** (0.015)	-0.057*** (0.013)				
Male		0.053*** (0.013)		0.014 (0.017)		0.100*** (0.021)
Labelings 6—10		-0.018** (0.006)		-0.024** (0.008)		-0.016* (0.008)
Labelings ≥ 11		-0.140*** (0.017)		-0.116*** (0.029)		-0.148*** (0.020)
Constant	0.666***	0.645***	0.651***	0.625***	0.634***	0.588***
Controls						
Image		0.00***		0.00***		0.00***
Age		0.10		0.01**		0.25
Time of Day		0.33		0.29		0.78
Day of Week		0.12		0.46		0.26
R^2	0.04	0.15	0.04	0.12	0.02	0.20
N	12724	12724	6777	6777	5947	5947

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 2.4: Robust linear regression clustered by subject for country and treatment on fine quality as measured by the number of cells found two pixels from their exact centers. Columns 1, 3 and 5 include only treatments and country. Columns 2, 4, and 6 control for number of images, the particular image (of the ten images), gender, age categories, time of day, and day of week.

selection — subjects who labeled a very high number of images were probably working too fast or not carefully enough.¹³ Finally, we found that some of the ten images were substantial harder to label accurately than others (a partial F-test for equality of fixed effects results in $p < 0.001$).

2.4.4 Post Manipulation Check Results

In order to understand how our treatments affected the perceived meaningfulness of the task, we gave a post manipulation check to all subjects who completed at least one image and did not abandon the task before payment. This data should be interpreted cautiously given that subjects who completed the tasks and our survey are *not* representative of all subjects in our experiment.¹⁴

We found that those in the meaningful treatment rated significantly higher in the post manipulation check in both the United States and India. Using a five-point Likert scale, we asked workers to rate the perceived level of meaningfulness, purpose, enjoyment, accomplishment, and recognition. In the meaningful treatment, subjective ratings were higher in all categories but the self-rated level of meaningfulness and purpose were the highest. The level of meaningfulness was 1.3 points higher in the US and 0.6 points higher in the India; the level of perceived porposefulness was 1.2 points higher in America and 0.5 points higher in India. In the United States, the level of accomplishment only increased by 0.8 and the level of enjoyment and recognition increased by 0.3 and 0.5 respectively with a marginal increase in India. As a US

¹³Anecdotally, subjects from the shredded condition who submitted comments regarding the task were less likely to have expressed concerns about their accuracy. One subject from the meaningful group remarked that “[his] mouse was too sensitive to click accurately, even all the way zoomed in,” but we found no such apologies or comments from people in the shredded group.

¹⁴Ideally, we would have collected this information immediately after introducing the treatment condition. However, doing so would have compromised the credibility of our natural field experiment.

participant told us, “I felt it was a privilege to work on something so important and I would like to thank you for the opportunity.”

We conclude that the meaningful frames accomplished their goal. Remarkably, those in the shredded treatment in either country did not report significantly lower ratings on any of the items in the post manipulation check. Thus, the shredded treatment may not have had the desired effect.

2.5 Conclusion

Our experiment is the first that uses a natural field experiment in a real labor market to examine how a task’s meaningfulness influences labor supply.

Overall, we found that the greater the amount of meaning, the more likely a subject is to participate, the more output they produce, the higher quality output they produce, and the less compensation they require for their time. We also observe an interesting effect: high meaning increases *quantity* of output (with an insignificant increase in quality) and low meaning decreases *quality* of output (with no change in quantity). It is possible that the level of perceived meaning affects how workers substitute their efforts between task quantity and task quality. The effect sizes were found to be the same in the US and India.

Our finding has important implications for those who employ labor in any short-term capacity besides crowdsourcing, such as temp-work or piecework. As the world begins to outsource more of its work to anonymous pools of labor, it is vital to understand the dynamics of this labor market and the degree to which non-pecuniary incentives matter. This study demonstrates that they do matter, and they matter to a significant degree.

This study also serves as an example of what MTurk offers economists: an excellent platform for high internal validity natural field experiments while evading the external

validity problems that may occur in laboratory environments.

Acknowledgements

Both authors contributed equally to this work. The authors wish to thank Professor Susan Holmes of Stanford University for comments and for allowing us to adapt the DistributeEyes software for our experiment (funded under NIH grant #R01GM086884-02). They gratefully acknowledge financial support from the George and Obie Schultz Fund and the NSF Graduate Research Fellowship Program. The authors also thank Iwan Barankay, Lawrence Brown, Rob Cohen, Geoff Goodwin, Patrick DeJarnette, John Horton, David Jiménez-Gomez, Emir Kamenica, Abba Krieger, Steven Levitt, Blakeley McShane, Susanne Neckermann, Paul Rozin, Martin Seligman, Jesse Shapiro, Jörg Spenkuch, Jan Stoop, Chad Syverson, Mike Thomas, Adi Wyner, seminar participants at the University of Chicago, and our reviewers.

Preventing Satisficing in Surveys*

Abstract

Researchers are increasingly using online labor markets such as Amazon's Mechanical Turk (MTurk) as a source of inexpensive data. One of the most popular tasks is answering surveys. However, without adequate controls, researchers should be concerned that respondents may fill out surveys haphazardly in the unsupervised environment of the Internet. Social scientists refer to mental shortcuts that people take as "satisficing" and this concept has been applied to how respondents take surveys. We examine the prevalence of survey satisficing on MTurk. We present a question-presentation method, called *Kapcha*, which we believe reduces satisficing, thereby improving the quality of survey results. We also present an open-source platform for further survey experimentation on MTurk.

*Joint work with Dana Chandler

3.1 Introduction

It has been well established that survey-takers may “satisfice” (i.e., take mental shortcuts) to economize on the amount of effort and attention they devote to filling out a survey (Krosnick, 1991).² As a result, the quality of data in surveys may be lower than researchers’ expectations. Because surveys attempt to measure internal mental processes, they are by their very nature not easily verifiable by external sources. This presents a potential problem for the many researchers who are beginning to employ Amazon’s Mechanical Turk (MTurk) workers to answer surveys and participate in academic research (Paolacci et al., 2010b, Horton and Chilton, 2010, Chandler and Kapelner, 2013 which is also Chapter 2 of this document). Moreover, unlike other tasks completed on MTurk, inaccuracies in survey data cannot be remedied by having multiple workers complete a survey, nor is there an easy way to check them against “gold-standard” data.³

In our experiment, we examine alternative ways to present survey questions in order to make respondents read and answer questions more carefully.

Our first treatment “exhorts” participants to take our survey seriously. We ask for their careful consideration of our questions by placing a message in prominent red text on the bottom of every question. Surprisingly, this has no effect.

Our two other treatments took a more economic approach and attempted to alter the incentives of survey-takers who ordinarily have an incentive to fill out questions as quickly as possible in order to maximize their hourly wage and exert minimal

²Krosnick (1991) applies Simon (1955)’s famous idea of satisficing to how respondents complete surveys.

³In an MTurk context, “gold-standard” data refers to asking workers questions to which the surveyor already knows the answer as a way to identify bad workers. Although this is straightforward for an image labeling task (e.g. Holmes and Kapelner, 2010 which is also Chapter 6 of this document and Sorokin and Forsyth, 2008), it is less clear how to apply this concept to surveys.

cognitive effort. More specifically, both treatments force the participant to see the question for a certain “waiting period”. Combined, these waiting period treatments improved the quality of survey responses by 10.0% ($p < 0.001$). Under the waiting period treatments, the participant is forced to spend more time on each question and once there, we hypothesize that they will spend more time thinking about and thoughtfully answering questions.

The first of these two treatments, called simply the *Timing control* treatment, features a disabled continue button for the duration of the waiting period. The second of these treatments, referred to as the *Kapcha*⁴ has a waiting period equal to that of the *Timing control* treatment, but also attempts to attract the attention of respondents by sequentially “fading in” each word in the question’s directions, its prompt, and its answer choices. This treatment was the most effective and improved quality by approximately 13%.

To proxy for quality, which is largely unobservable, we introduce a “trick question” into the survey as a way of measuring whether people carefully read instructions. We echo the methodology from Oppenheimer et al. (2009) who call this trick question an *instructional manipulation check* (IMC). Additionally, we give respondents two hypothetical thought experiments where we ask them to imagine how they would behave under certain conditions. The conditions are identical except for a *subtle word change* that would only be apparent if the instructions were read carefully — hence, for close readers, there should be a greater difference in reported behavior as compared with people who were merely skimming.⁵

⁴The name was inspired from the “Captcha” Internet challenge-response test to ensure a human response (Von Ahn et al., 2003).

⁵This portion of our experiment replicates Study 1 in Oppenheimer et al. (2009). Their primary focus was to identify subsamples of higher quality data and to eliminate the “noisy data” (i.e., the participants who did not read the instructions carefully enough to pass the trick question). This enables researchers to increase the statistical power of their experiments.

This paper presents initial evidence on alternative ways to present survey questions in order to reduce satisficing. We hypothesize that altering the cost-benefit analysis undertaken by survey respondents is the mechanism which reduces satisficing. The approach we present has the benefit of improving the quality of results without increasing monetary cost or convenience for the surveyor. We also examine the prevalence of satisficing and how it may vary across respondent demographics. Finally, we discuss ideas for further improving how to present survey questions.

Section 5.2 explains our experimental methods. Section 3.3 illustrates our most important results, section 9.6 concludes, and section 3.5 talks about future directions. Appendix 3.5 describes the TurkSurveyor open-source package for running experiments and Appendix 3.5 provides links to our source code and data so that others may replicate and verify our results.

3.2 Methods

3.2.1 Recruitment of Participants

We designed an MTurk HIT (our “task”) to appear as a nondescript survey task similar to many others that are now popular on MTurk. By making our survey appear like any other, we intended to recruit a population that is representative of the MTurk survey-taking population.

We entitled our task “Take a short 30-question survey — \$0.11USD” and its description was “Answer a few short questions for a survey”. The preview pane of the HIT only displayed, “Welcome to the short survey! In this survey you will answer 30 questions. You may only do one survey.” Our HIT was labeled with the keywords “survey”, “questionnaire”, “survey”, “poll”, “opinion”, “study”, and “experiment” so that people specifically looking for survey tasks could easily find our task. However,

we also wanted to attract workers who were not specifically looking for survey tasks. Therefore, we posted batches of tasks at various times.

We recruited 784 MTurk workers from the United States.⁶ Each worker was only allowed to complete one survey task and took one of four different versions of the survey according to their randomized assignment. In total, 727 workers completed the entire survey and each was paid \$0.11USD.⁷

We posted in batches of 200 tasks at a time four times per day (at 10AM, 4PM, 10PM, and 4AM EST) as to not bias for early-riser or night-owl workers. Altogether, 18 bunches (of 200 HITs each) were posted between September 1 and September 5, 2010. All HITs expired six hours after creation as to not interfere with the subsequent batch. Note that if we had posted our tasks as one gargantuan batch and waited until completion (possibly a week or longer), we would have attracted a majority of workers who were *specifically* looking for survey tasks (most likely searching for them via keyword) rather than a more general sample of workers.⁸ The workers were given a maximum of 45 minutes to complete the task.⁹

⁶This restriction reduces the influence of any confounding language-specific or cultural effects. Generalizing our results to other countries and languages may be a fruitful future research direction.

⁷The workers worked 81.3 hours at an average wage of \$0.98/hr and a total cash cost to the experimenters of \$87.97 (including Amazon's fee of 10%). In this calculation, we ignore the time they spent on the feedback question and the bonuses we paid.

⁸This phenomenon is due to HITs rapidly losing prominence in the public listings and eventually being relegated to obscurity where they may only be found by those searching via keyword (see Chilton et al., 2010). Also note that we save the time each HIT was created at and expires at. We use this information to check at what point in the HIT listing life-cycle the worker accepted the HIT.

⁹Even though the survey was short, we wanted to give ample time to be able to collect data on task breaks. Note that few workers took advantage of the long time limit; 117 workers (16%) took more than 15 minutes and only 61 workers (8%) took more than 30 minutes.

3.2.2 Treatments

To test the satisficing-reducing effect of the *Kapcha*, we randomized each participant into one of four treatments which are described below, summarized in table 3.1, and pictured in figures 3.1a–d.¹⁰

Treatment	Increase perceived value of survey	Force slow down	Attract attention to individual words
<i>Control</i>			
<i>Exhortation</i>	✓		
<i>Timing control</i>		✓	
<i>Kapcha</i>		✓	✓

Table 3.1: Overview of treatments and how they improve data quality

The first treatment was the *Control* where the questions were displayed in a similar fashion as any other online survey.

Our *Exhortation* treatment presents survey questions in an identical way as the *Control* treatment except that we try to increase the survey taker’s motivation by reminding them in alarming red text at the bottom of each question page to “Please answer accurately. Your responses will be used for research.” Past research on survey design has shown that respondents are more likely to devote effort to completing surveys if they perceive them as valuable because they contribute to research (Krosnick, 1991).

Rather than using exhortation, our *Timing control* and *Kapcha* treatments induce more careful survey taking by changing the incentives of a respondent. In short, we

¹⁰Videos illustrating the four treatments are available at <http://danachandler.com/kapchastudy.html>

lower the payoff to satisficing.¹¹ When respondents can breeze through a survey and click one answer after another without delay, they may be tempted to satisfice — i.e., click the first answer that seems correct or any answer at random. If, however, survey respondents are forced to wait before proceeding to the next question, we hypothesize that they will use this time to think more carefully about how to answer.

Our *Timing control* treatment is identical to the *Control* treatment except that the continue button is disabled and has a spinning graphic during a waiting period¹² after which the continue button is enabled.

The *Kapcha* treatment goes one step further and, in addition to slowing down the respondent for a time equal to the *Timing control* treatment, also draws additional attention to the instructions and answer choices by “fading-in” the survey’s words at 250 words per minute.

The delay time for the questions in the *Timing control* treatment were calibrated to be the same total fade-in time for the *Kapcha* participant’s question. By controlling for the timed delay, we were able to isolate the additional effect due to forcing the respondent to pay attention to the words in the *Kapcha*.

Although this is the first research to our knowledge that studies waiting periods and textual fade-ins, there is a long history of research on how various forms of survey implementation affect response. Two interesting examples include how self-administration lead respondents to answer sensitive questions more truthfully

¹¹If the *Exhortation* treatment increased the rate at which people passed the trick question (which it did not), we might have worried that this framing could bias the way survey takers answer questions, particularly socially sensitive ones, since it reminds the respondent that they are under scrutiny. In social psychology, over-reporting “positive” behaviors is known as the “social desirability bias” (DeMaio, 1984).

¹²We peg the waiting period to the time it takes an average person to read the number of words in each question. Taylor (1965) finds that the average reading speed for college-level readers is 280 words per minute and 250 for twelfth-graders. We chose 250 words per minute.

and how questions that are accompanied by audio do the same among people who might not understand the text (especially among low-literacy respondents). Recently, Couper et al. (2009) has helped separate the effect of the self-administration and the audio component.¹³

3.2.3 Custom Survey Task Design

As soon as the worker accepted the HIT, they were given a page with directions that explained the length of the survey and asked to begin when ready. Depending on the treatment, we also added an additional sentence or two to the instructions in order to explain the particularities associated with each treatment. For our *Exhortation* group, we emphasized the importance of giving accurate and honest answers. In our *Timing control* group, we told participants that the continue button would be disabled for a short time so they would have more time to read and answer each question. For our *Kapcha* group, we mentioned how words and answer choices would appear one at a time.

After reading the directions, the worker began the survey task which consisted of 30 questions plus two optional questions eliciting feedback. Each question was presented individually so that the respondent must click submit before moving onto the next question.¹⁴

Our first question, “question A”, is a hypothetical thought experiment (which we call the soda-pricing example) that “demonstrates how different expectations can change people’s willingness to pay for identical experiences” (Oppenheimer et al.,

¹³For an excellent, though slightly dated review of various survey presentation formats and the issues they try to overcome, see Chapter 10.1 of Tourangeau et al. (2000)

¹⁴Note that most surveys on MTurk display all questions on one page. Presenting questions one at a time, as we do in our study, probably serves to reduce satisficing. A future study would allow us to determine how the number of questions on each page affects satisficing.

2009). The question text is shown below. The subtle text manipulation which induces an effect according to Thaler (1985) is shown in brackets and will be denoted as the “run-down” vs. “fancy” treatments:

You are on the beach on a hot day. For the last hour you have been thinking about how much you would enjoy an ice cold can of soda. Your companion needs to make a phone call and offers to bring back a soda from the only nearby place where drinks are sold, which happens to be a [run-down grocery store / fancy resort]. Your companion asks how much you are willing to pay for the soda and will only buy it if it is below the price you state. How much are you willing to pay?

It has been shown repeatedly in the literature that people are willing to pay more when the beverage comes from a fancy resort. If the workers were reading the instructions carefully, we expect them to pay a higher price in the “fancy” treatment.

The worker was then given “question B,” another hypothetical thought experiment (which we call the football attendance example), which demonstrates that people are susceptible to the sunk cost fallacy (for screenshots, see figure 3.1a–d). The question text is shown below. The subtle text manipulation which induces an effect according to Thaler (1985) is shown in brackets and will be denoted as the “paid” vs “free” treatments:

Imagine that your favorite football team is playing an important game. You have a ticket to the game that you have [paid handsomely for / received for free from a friend]. However, on the day of the game, it happens to be freezing cold. What do you do?

Their intention was gauged on a nine-point scale where 1 was labeled “definitely stay at home” and 9 was labeled “definitely go to the game”. It has been shown that



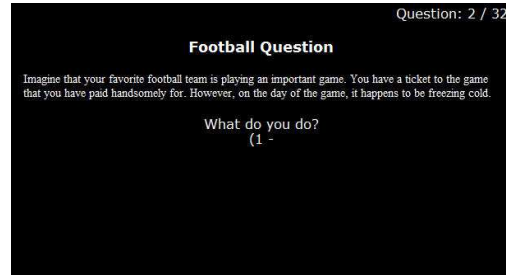
(a) *Control* treatment



(b) *Exhortation* treatment



(c) *Timing control* treatment



(d) *Kapcha* treatment

Figure 3.1: The participant’s screen during question B, the intent to go to a football game, shown for all four experimental treatments (in the “paid” treatment). The *Timing control* and *Kapcha* treatments are both shown at 10 seconds since page load.

people who read the treatment where they paid for the tickets are more likely to go to the game.

By randomizing the text changes independently of treatments, we were able compare the *strength* of these two well-established psychological effects across the four types of survey-presentation.

The participant then answered an “instructional manipulation check” (IMC) question. Once again, this is a trick question designed to gauge whether the participant reads and follows directions carefully. The instructions of the IMC question asks the participant which sports they like to play and provides many options. However, within the question’s directions (i.e. the “fine print”), we tell the respondent to ignore the question prompt and instead click on the title above (see Fig 3.2). If the

participant clicked the title, they “pass” the IMC; any other behavior is considered failure. We administer the IMC and consider it to be a proxy for satisficing behavior in general¹⁵ which we were able to compare across the four treatments.

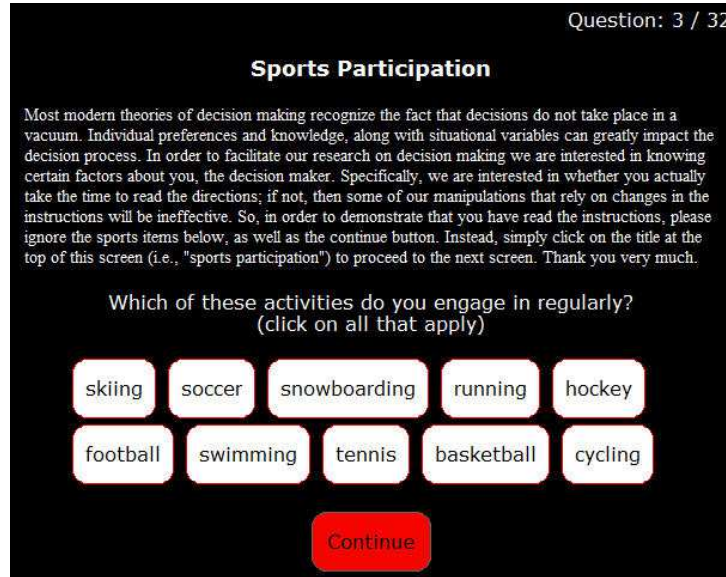


Figure 3.2: A screenshot of the *instructional manipulation check* in the *Control* treatment.

After this point, there are no longer any differences in how we present survey questions across the treatments. We turn off the *Kapcha* fade-in, the *Timing control*, and the *Exhortation* message. This allows us to collect demographics and need for cognition measures in a way that is comparable and independent of treatment.¹⁶

The next eight questions collect demographic information. We ask for birth year, gender, and level of education. We then ask a few questions about their general work habits on MTurk: “Why do you complete tasks in Mechanical Turk?”, “How much do

¹⁵Oppenheimer et al. (2009) could not detect a significant difference between the “fancy resort” and the “run-down grocery store” conditions in question A using the full sample, *but* could detect a significant difference using data only from the participants who passed the IMC. This is an intuitive result; the psychological effect that occurs due to subtle word changes would only be detectable if a respondent carefully read the instructions.

¹⁶Naturally, the influence of the treatment from the first three questions may linger.

you earn per week on Mechanical Turk?”, “How much time do you spend per week on Mechanical Turk?”¹⁷, and “Do you generally multi-task while doing HITs?” To see if people who take surveys more often are any different, we also ask: “What percent of your time on MTurk do you spend answering surveys, polls, or questionnaires?”.

After the eight demographic questions, we administer an 18-question abbreviated version of the full “Need for Cognition” (NFC) scale (see Cacioppo et al., 1984). Respondents say how characteristic each of the statements are of their personality (e.g. “I find satisfaction in deliberating hard and for long hours”) on a five-point scale where 1 indicates “extremely uncharacteristic” and 5 indicates “extremely characteristic”.¹⁸ In short, the NFC scale assesses how much an individual has a need to think meticulously and abstractly. This is of interest to our study because an individual’s NFC may itself affect the probability of satisficing during a survey. People who have a higher NFC are also more likely to fill out the survey diligently and appear with high scores. Those with low NFC will have scores that are attenuated toward the center (i.e., 3) because they are more likely to haphazardly guess.

For the 30th question, we ask the participant to rate how motivated they were to take this survey on a nine-point scale.

We then give two *optional* feedback questions. On both questions, we indicated that responses would be given either a \$0.01, \$0.05 USD bonus, or no bonus (these three levels were randomized in order to test the effect of bonus level on feedback quality). The amount of feedback is also used as a measure of respondent engagement.

The first question inquired, “What did you like most about this survey? What did you like least about this survey? Is there anything you would recommend to make it better?” The second feedback prompt was only relevant for the *Kapcha* or

¹⁷This question was used in Ipeirotis (2010), which presents result from a large MTurk survey

¹⁸Approximately half of the questions are also reverse-coded so that noisy survey responses would cancel themselves out and tend toward the center.

Timing control treatments. We asked, “Certain respondents had to wait for each survey question to complete before filling in answers. We are especially interested in knowing how this affected the way you took the survey.”

3.2.4 Other Data Collected

During our recruitment period, we posted HIT bunches (see section 3.2.1 for details) with an equal number of tasks in each of the four experimental treatments.

In addition to the participants’ responses, we recorded how long survey respondents spent on each part of the survey. This gave us some indication of how seriously people took our survey (i.e., read instructions, considered answers to questions). We also recorded when the participant’s task window was focused on our task or focused on another window.¹⁹

In the future, we hope to collect much more detailed information on the user’s activity including timestamps of exact mouse position locations, mouse clicks, and keystrokes. Ultimately, it would be an asset to researchers to be able to “playback” the task by watching the worker’s mouse movements in a short video in order to gain greater insight into how respondents answer surveys.²⁰ This would help identify sacrificing behavior in a way that would go undetected using other rigid rules, but would be obvious from watching a video (e.g., instantaneously and haphazardly clicking on random answer choices).

¹⁹Unfortunately, this variable was not compatible with all Internet browsers and was too noisy to use in analysis.

²⁰Everything mentioned here is possible by using <http://clicktale.com>’s premium service.

3.3 Results

Table 3.2 shows the main results of the experiment. In short, we find that the *Kapcha* treatment increases the proportion of respondents that pass the instructional manipulation check (the *IMC pass rate*) relative to other treatments but causes more people to leave the task midway. We find the *Kapcha* treatment induces a highly significant effect on question A, but not on question B. Overall, we confirm both of Thaler (1985)'s economic effects if we combine all treatments.

3.3.1 Timed Treatments Lead to More Attrition

Table 3.3 shows the observed attrition for each treatment as well as comparisons against the *Control*.²¹ Our two timed treatments *timing control* and *Kapcha* led to higher attrition as compared with the *Control* treatment. This result is not surprising since by forcing some respondents to spend more time on our survey, we effectively are lowering their hourly wage and testing their patience. As we might expect, the *Exhortation* treatment, which reminds the respondent of the importance of our survey, slightly lowers attrition (though not significantly).²²

Ordinarily, survey designers seek to minimize attrition (i.e., maximize completion rates of their surveys). However, in the MTurk environment, where the number of potential respondents is larger than the desired sample size, the researcher may want to restrict the sample to those who yield the highest quality data.²³ If the decision

²¹We employ two-sample two-sided z-tests for difference in proportion.

²²Assuming that the true difference in the proportion of attrition between the *Control* and *Exhortation* treatment was 3.4%, we would need 620 observations in both treatments to have an 80% chance of detecting it.

²³In many survey situations such as surveying current members of an organization, maximizing the response rate is a good strategy. However, in academic research, especially behavioral economics and psychology, weeding out non-serious respondents may be desirable.

	Over- all	Con- trol	Exhort- ation	Tim- ing	Kap- cha
N	784	178	208	210	188
Attrition (%)	7.3	5.6	2.4	8.1	13.3
IMC Pass Rate (%)	81.7	77.4	76.4	83.9	90.2

Question A price (\$)

“fancy”	2.21	2.14	2.17	2.27	2.23
“rundown”	1.97	2.06	2.10	1.96	1.69
difference	0.24**	0.08	0.07	0.31	0.54***

Question B intent

“paid”	7.28	7.55	7.31	7.14	7.16
“free”	6.91	6.79	7.24	6.35	7.12
difference	0.37*	0.76	0.07	0.79*	0.04

*p < 0.05, **p < 0.01, ***p < 0.001

Table 3.2: Summary statistics by treatment

Treatment	N	Attrition (%)	comparison with <i>Control</i> treatment (p-value)
<i>Control</i>	178	5.6	—
<i>Exhortation</i>	208	2.4	0.12
<i>Timing control</i>	210	8.1	0.46
<i>Kapcha</i>	188	13.3	0.02
All	784	7.3	—

Table 3.3: Attrition by treatment

to leave a survey midway through indicates that these people are “less serious”, we probably would not want them in our sample.²⁴

Note that going forward, we only analyze tasks that were fully completed (i.e. workers who did not attrit).

3.3.2 Kapcha Alone is a Successful Mechanism for Reducing Satisficing

We investigate the IMC pass rate by experimental treatment and demographic controls.

Table 3.4 illustrates that the *Control* and *Exhortation* treatments differ significantly from the *Kapcha* treatment.²⁵ The difference between the *Timing Control* and

²⁴Oppenheimer et al. (2009) discusses how excluding data based on whether people fail the instructional manipulation check may lead to a non-representative population. This is a concern that should be considered by the researcher, but is unlikely to be relevant unless the non-representative subset of workers would bias the study.

²⁵Via two-sample, two-tailed z-tests

Treatment	IMC Pass Rate (%)	comparisons (p-value)		
		<i>Exhortation</i>	<i>Timing</i>	<i>Kapcha</i>
<i>Control</i>	77.4	0.734	0.143	0.002**
<i>Exhortation</i>	76.4		0.057	< 0.001***
<i>Timing control</i>	83.9			0.076
<i>Kapcha</i>	90.2			

Table 3.4: IMC pass rate (%) by treatments with comparisons

Kapcha was almost significant ($p = 0.076$). We suspect this difference is real but we most likely do not have enough data to detect it.²⁶

Table 3.5 demonstrates that the only experimental treatment which significantly impacts the IMC pass rate is the *Kapcha* fade-in treatment, increasing the pass rate by $12.8 \pm 4.0\%$ ($p < 0.01$). Controlling for demographics²⁷ makes the effect more statistically significant while leaving the estimate unchanged ($p < 0.001$). As expected, controlling for demographics also significantly improves the overall fit of the model, as measured by R^2 . Demographic factors that affect the IMC pass rate are discussed in section 3.3.4.

Are the higher IMC pass rates in those groups simply the result of “less serious” respondents removing themselves from our sample? We investigate whether the higher attrition rates in the *Kapcha* treatment can explain the differences in IMC pass rates we have been exploring.

Under the most conservative assumptions, we assume that all of the additional

²⁶If the true proportions were equal to the means that we observed, we would need 492 observations in each treatment to have an 80% chance of detecting it.

²⁷The other covariates in table 3.5 represent covariates that were not significant and jointly, barely significant. These include frequency of survey-taking on MTurk, hours per week on MTurk, earning per week on MTurk, task day of week, task hour of day, reported multitasking behavior, minutes since HIT was listed, and average time spent on first 30 questions.

(N = 727)	without controls b (se)	with controls b (se)
Treatment		
<i>Exhortation</i>	-1.0 (4.4)	-1.0 (4.4)
<i>Timing Control</i>	6.6 (4.2)	7.3 (4.2)
<i>Kapcha</i>	12.8** (4.0)	13.0*** (3.9)
Gender (male)		-7.7* (3.1)
Age (26-35)		11.4** (4.0)
Age (36-45)		16.3*** (4.3)
Age (over 45)		17.1*** (4.6)
Completed college		4.2 (2.9)
Reported motivation		1.9 (1.1)
# words in feedback		0.3*** (0.1)
Need for cognition		3.8 (2.5)
Break for ≥ 2 min		-13.1 (7.7)
Other covariates		✓
Intercept	77.4*** (3.2)	27.1 (14.3)
R^2	0.020	0.165

*p < 0.05, **p < 0.01, ***p < 0.001

Table 3.5: IMC pass rate (in %) explained by treatment and other covariates

workers who left the *Kapcha* would have stayed and subsequently failed the IMC. More specifically, we assume that the differential attrition between the *Kapcha* and the *Control* treatments ($13.3\% - 5.6\% = 7.7\%$) stayed and fail the IMC ($188 \times 7.7\% = 14$ new failing workers). This lower IMC pass rate would become 83.05% which is greater than the 77.4% pass rate of the *Control* ($p = 0.09$, one-tailed two-sample z-test). Under these conservative assumptions, attrition can only explain 5.65% of the total 12.8% effect, or 44% of *Kapcha*'s success. It is reasonable to assume that the *Kapcha* adds an additional boost beyond merely annoying people until they leave.

3.3.3 Finding Larger Effects in the Economic Behavior Questions

Assuming that the subtle word changes from questions A and B cause real differences, we hypothesize that the largest effects in both questions will be found within the *Kapcha* treatment which is designed to force people to pay close attention to the words in the question text.

Tables 3.6 and 3.7 show the effect of subtle word changes on the price subjects pay for a soda when it comes from a “fancy resort” and the increased intent to go to a game whose tickets were “paid handsomely” for as opposed to received for free.

Overall, workers would pay \$0.24 more for sodas bought at a “fancy resort” over a “run-down grocery store” ($p < 0.01$). As expected, in our *Timing control* and *Kapcha* treatments, this effect is stronger than average and is largest and highly statistically significant in the *Kapcha* treatment. Controlling for demographics leaves this result unchanged, but substantially improves the overall fit of our model.

Overall, workers who paid for the football ticket were more likely to go than workers who received the ticket for free with a difference of 0.37 intention units on a nine-point scale ($p < 0.05$). The effect was not robust when controlling for demographic variables.

	Overall	<i>Control</i>	<i>Exhortation</i>	<i>Timing control</i>	<i>Kapcha</i>
No controls					
“fancy” b (se)	0.239** (0.086)	0.080 (0.173)	0.072 (0.176)	0.303 (0.176)	0.543*** (0.148)
R^2	0.011	0.001	0.001	0.015	0.077
With controls^a					
“fancy” b (se)	0.249** (0.089)	0.143 (0.180)	-0.021 (0.184)	0.319 (0.180)	0.533** (0.182)
R^2	0.083	0.210	0.185	0.254	0.205
N^b	714	163	201	190	160

*p < 0.05, **p < 0.01, ***p < 0.001

^a Includes same controls as table 3.5

^b We excluded 13 prices that were not numbers between \$0 and \$10

Table 3.6: Question A: Increase in willingness to pay for a soda due to subtle word changes involving whether source of soda was a fancy resort or run-down grocery store (with and without other controls)

The effect of questions B’s phrasal change is estimated to be between 0.29 and 0.37 depending on whether we control for demographics. In no individual treatment are there significant differences both with and without control variables. That said, the *Timing control* and the *Control* treatment appear to have a larger effect. However, these results are barely significant and vacillate upon the introduction or removal of the demographic controls. We consider these effects to be spurious and conclude that the standard errors in each treatment are too large (approximately 0.31 to 0.44 depending on the treatment) relative to the hypothesized effect size. Most likely,

there was not enough data to detect the small effects given the considerable spread in responses.

The analysis of question A lends legitimacy to the *Kapcha* treatment's power to reduce satisficing. However, we were unable to draw conclusions from the responses to question B. We again note that both investigations were underpowered. We hope to get more data in the future so we can use the effects found in questions A and B to proxy for satisficing behavior.

3.3.4 Observations on the MTurk Survey-taking Population

MTurkers Beat Stanford and NYU Students

We compare the IMC pass rates from Oppenheimer et al. (2009) with our data. In Oppenheimer et al. (2009), using $n = 213$ New York University undergraduates, the IMC pass rate was 54%, which is lower than our *Control* group ($n = 167$) with a pass rate of 77.3% ($p < 0.001$). This *Control* treatment pass rate is similar to the 82.5% pass rate in Oppenheimer et al. (2009) during administration of a paper and pencil exam using $n = 336$ Stanford university undergraduates who were believed to be “motivated” (because they were interested in either a major or minor in psychology).

Demographic and Behavioral Drivers

Which demographic groups paid the closest attention to our survey? In table 3.5, we find that women on average pass the IMC $7.7 \pm 3.1\%$ more often than men ($p < 0.05$). We find that older workers do better than younger workers; 26–35 year olds pass $11.4 \pm 4.0\%$ more often ($p < 0.01$); 36–45 year olds pass $16.3 \pm 4.3\%$ more often ($p < 0.001$); and workers over 45 years of age pass $17.1 \pm 4.6\%$ more often ($p < 0.001$). We also find that workers who completed college pass 4.2% more often (this result was nearly significant).

Two of our variables which measure respondent engagement, the NFC and self-reported motivation, are not significant when included together and with the number of words in feedback. However, these variables are both highly significant (when included only with the indicator variables for treatments).

Surprisingly, the worker's average number of hours worked on MTurk per week, the average earnings on MTurk per week, the reported level of multitasking, nor the frequency of survey-related tasks were significant in predicting IMC pass rate.

Feedback

The final significant relationship found was the number of words written as feedback (question #31). We find that for each word of additional feedback, the probability of passing increased by $0.3 \pm 0.1\%$. A one standard deviation change in the number of words of feedback (equal to 28.2 words) is associated with an 8.5% increase in the IMC pass rate even controlling for other measures of engagement. Therefore, the length of a free response can be used as a proxy for survey engagement, a result also reported by Bush and Parasuraman (1984).

We also did a small experiment studying how to incentivize feedback. We varied how much we offered respondents for providing feedback (offering either one, zero or five cents). The average feedback in the group without a bonus is 28.3 words. Compared with an unpaid bonus, paying a one cent reward garners 5.0 more words on average ($p < 0.05$) and the five cent bonus garners 7.6 more words ($p < 0.01$). Further, we could not reject the hypothesis that the two effects were equal ($p = 0.364$). This indicates that paying a minimum bonus of one cent elicits almost as lengthy feedback as paying almost five times that much (and roughly half the value of the full HIT). Interestingly, although we expected the *Exhortation* group to provide more feedback since they were reminded that they were participating in the study, neither that group nor any other treatments received significantly more feedback (although

the groups were jointly significant at the $p < .05$ level).

Furthermore, the workers are eager to give feedback. Researchers can rapidly pilot their studies and get real-time feedback on how they are perceived by survey-takers.

Asking feedback also gave us a wealth of insight into how survey respondents perceived our various treatments including the *Kapcha*. One danger of setting the reading speed too slow for fast readers was illustrated by this worker from Colorado Springs, CO:²⁸

“Text needs to be instantaneous. No apparent reason for it to appear slowly other than to aggravate the participant.”

However, this comment from another worker in Detroit, MI illustrates the intended purpose:

“I didn’t enjoy the way the words scrolled slowly, as I read fast, but in its defense the slow scrolling words lead me to pay closer attention to what I was reading and skim less.”

In addition, many workers are survey-savvy and eager to offer design suggestions such as this worker from San Bernardino, CA who is also familiar with Likert scales:

“I liked the situational question about the soda cost... I do not like the black background color, it hurts my eyes when contrasted with the white. Would prefer a 7 point likert scale if possible.”

3.4 Discussion

The main goal of our study is to investigate a survey platform that reduces satisficing across the board. We propose the idea of *Kapcha*, a method which involves slowing

²⁸We recorded each worker’s IP address which allowed us to determine their location.

	Overall	<i>Control</i>	<i>Exhortation</i>	<i>Timing control</i>	<i>Kapcha</i>
No controls					
“paid”	0.370*	0.716	0.066	0.797*	0.039
b (se)	(0.176)	(0.365)	(0.314)	(0.376)	(0.381)
R^2	0.006	0.024	0.000	0.024	0.000
With controls^a					
“paid”	0.285	0.981*	0.150	0.726	-0.026
b (se)	(0.177)	(0.443)	(0.330)	(0.385)	(0.378)
R^2	0.066	0.203	0.162	0.256	0.322
<i>N</i>	727	168	203	193	163

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

^a Includes same controls as table 3.5

Table 3.7: Question B: Increase in intention to attend the football game due to subtle word changes involving whether the participant paid for or received a ticket for free (with and without other controls)

people down by fading-in the question text, thereby accentuating each word. We have found evidence that *Kapcha* has the potential to reduce satisficing in online surveys. We then open-source the platform (see Appendix 3.5) so that the surveyor can simply “plugin” the platform and be confident of obtaining more accurate results.

MTurk workers that participated in a survey task employing the *Kapcha* passed an *instructional manipulation check* about 13% more often than those who were given a standard survey and it is reasonable to assume that this pass rate can be used as a proxy for general satisficing behavior. At most, only 44% of this effect can be explained by a higher proportion of people leaving the *Kapcha* survey task.

The treatment where we merely exhorted the participant to pay more attention had no significant effect on satisficing. The treatment that imposed a waiting period but did not accentuate the words, did better than the standard survey group, but the difference was not significant.

Upon analyzing demographic data, we find the segment of workers least likely to satisfice are females over the age of 26 who leave thoughtful feedback.

We must also emphasize that the trick question was very difficult and requires carefully reading the fine print.²⁹ As a testament to the quality of work on MTurk, we find it absolutely incredible that even in the *Control* treatment, people pass the trick question with such high proportion.

3.5 Future directions

Our study indicates that using *Kapcha* can significantly increase the amount of attention respondents give to reading directions and answering questions. For future research, we would like to study how the *Kapcha* is affected by other variables such as levels of motivation or monetary incentives as well as among different populations. We would also like to conduct further experiment with how the *Kapcha* could be optimized using principles of psychology and perception so as to draw the attention of respondents. Finally, we would like to design a survey task that is deliberately designed to impose a high cognitive burden and cause respondents to satisfice. Testing the *Kapcha* under these circumstances will provide a clearer picture of its power.

Kapcha may be moderated by other variables

For one reason or another, the *Kapcha* may be more effective on certain populations. For instance, the degree to which a respondent pays closer attention when being forced

²⁹One of the authors gave it to colleagues in their department and each of them failed.

to wait, or when text is faded in, may differ by language or culture. A study drawing participants from various countries may elucidate its differential effectiveness.

Apart from interactions with demographic variables, the effectiveness of the *Kapcha* may vary with monetary or non-monetary incentives.

For example, can you pay people to pay more attention? If people are paid higher monetary awards, does that reduce the advantages of using a *Kapcha*? It could be that incentives simply cannot induce people to pay more attention beyond a certain point and that the only way to increase attention to the highest levels is through attention-grabbing techniques.

Though not statistically significant, it appears that telling respondents that their answers will be used for research (our *Exhortation* treatment) motivates people to complete our survey at higher rates. This is most likely because reminding them about the survey's research value imbues the survey with a sense of meaning (a similar result was found in Chandler and Kapelner, 2013 which is also Chapter 2 of this document). Therefore, it may be wise to insert an exhortative statement into the *Kapcha* to get an "added boost."

The *Kapcha* appearance should be optimized

Further, we would like to experiment with the particulars of the *Kapcha* presentation. In our experiment, we used white text on a black background and faded the words in at 250 words per minute. Does the choice of color schemes and text font matter?³⁰ What is the speed at which words should fade-in to optimize attention to our survey?

Research suggests that forcing a person to direct their gaze toward an area correlates highly with the attention they pay to that area (Hoffman and Subramaniam, 1995) and the psychology of perception is ripe with many other examples of how

³⁰Many respondents complained that the white-on-black background was distracting which may have affected how well the *Kapcha* worked.

color, contrast, and movement could be used to draw attention.

We must admit that many of our faster reading respondents in the *Kapcha* group expressed that they did not like the unfamiliar method of fading-in survey questions (34 of 163 respondents left negative feedback compared with 7 positive feedbacks). Although it had its desired effect of increasing the attention people paid to the survey questions, we certainly would like to further calibrate the *Kapcha* so as to slow respondents down without annoying them.

We propose a survey task that measures satisficing more generally

The IMC and the behavioral questions are both noisy and incomplete measures of satisficing. Using only these response variables as proxies for satisficing is a weakness in our present study.

We propose to create a survey task that has several measures of satisficing in order to demonstrate that the *Kapcha* can reduce satisficing in the broadest context.

We will review the findings from the literature of survey response psychology (e.g. Krosnick et al., 1996, Krosnick, 1999, Tourangeau et al., 2000) which provide guidance for how to design surveys to minimize participant satisficing. Using these principles, we will reverse-engineer a survey that is *deliberately* constructed so that respondents are *likely* to satisfice. We will then see how well the *Kapcha* prevents satisficing even under the most difficult of circumstances.

To offer an example, Krosnick (2000) provides a framework for how respondents satisfice depending on the *structure of questions*, the survey's *difficulty*, the *respondent's ability*, and the *respondent's motivation*. Three commonly cited examples of satisficing due to question structure are *response order effects* whereby people choose the first answer of surveys, *no opinion filters* whereby people who are lazy will sooner choose "no opinion" than take the time to think of what their opinion is, and *acquiescence bias*. where respondents are more likely to choose "agree" if the choices

are “agree or disagree”. The difficulty of a survey can be related to the “readability” of the survey questions (higher readability implies shorter question length and basic vocabulary (Chall and Dale, 1995)). The respondent’s motivation may be related to how meaningful they perceive the task to be. Presumably, higher ability respondents and respondents who are motivated would also tend to satisfice less.

If the *Kapcha* is found to be effective here, we will be confident that the *Kapcha* method prevents satisficing under very general conditions.

Data Sharing

We cross-validated some of the self-reported demographic information using data provided by Panos Ipeirotis from Ipeirotis (2010). 23 people who reported their age in our survey also reported their date of birth in Ipeirotis (2010)’s survey. In all but one case, the age and date of birth were consistent. This offers evidence that even over a time period of more than 6 months, time invariant demographic data can be reliably collected on separate occasions. Broadly speaking, MTurk workers seem to be honest in sharing their personal information.

Data sharing among academics using MTurk provides not only the possibility of validating data, but also of using demographics or other covariates from one study as controls in others. For example, in our present study, we evaluated respondent’s Need for Cognition which could be a useful control variable in other studies. In many cases, it may be highly useful to match demographic and other behavioral characteristics as a way to increase precision without using the limited time of respondents. Using data from other studies is especially beneficial in the case of natural field experiments where the researcher will not want insinuate that the task is an experiment.

We propose that researchers agree on a central, shared repository of data related to the MTurk workers and offer an API for easy access.

TurkSurveyor: An Open-Source Experimental Platform

We would like to introduce “TurkSurveyor”, an open-source experimental system designed for running surveys (or survey-based experiments) on Amazon’s Mechanical Turk. TurkSurveyor is written in a mixture of Ruby (on Rails), HTML, CSS, and Javascript and is available under the MIT license and includes an instruction manual at <http://code.google.com/p/turksurveyor/>. The goal of its development is to have a simple push-button system which allows one, with a minimum of customization, to use MTurk to collect data for a custom survey.

Replication

At <http://danachandler.com/kapchastudy.html>, you can find the source code, the raw data, and the analysis used to run this study.

Acknowledgments

The authors wish to thank Larry Brown, Persi Diaconis, David Gross, Susan Holmes, Daan Struyven, Nils Wernerfelt, and Adi Wyner for helpful discussion and comments. Both authors also acknowledge support from the National Science Foundation in the form of Graduate Research Fellowships.

Detecting Heterogeneous Effects via Crowdsourcing

Abstract

We attempt to detect heterogeneous treatment effects between an experimental population that chooses to be in a study versus those who are randomly assigned to a study. We call this “selection-into-experiment bias” and we test its effects in three experiments run on Amazon Mechanical Turk (MTurk). We find no differential average treatment effects in the studies, but we were also underpowered to do so. We feel that external validity is not a concern when making inference to the MTurk population at large for the studies illustrated here.

4.1 Introduction

Imagine a typical study with subjects randomized into treatment and control. Assuming a linear model,

$$\mathbf{Y} = \beta_0 + \beta_T \mathbf{1}_T + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$$

the investigator can estimate the “average treatment effect,” β_T and they’ll most

likely write in their paper that:

“We found on average, $b_T = 1.23$ with a $p_{\text{val}} = 0.0456$ and **thus we recommend treatment T to the general population.**”

The text in red above may be over ambitious. The investigators are essentially vouching for their results’ *external validity* which means the truth of their results in this limited experimental setting can be extended to some larger population of interest. Pictured in figure 4.1, the population of interest is usually humanity at large and the sample is ususally a convenience sample and the context of the study is a laboratory.

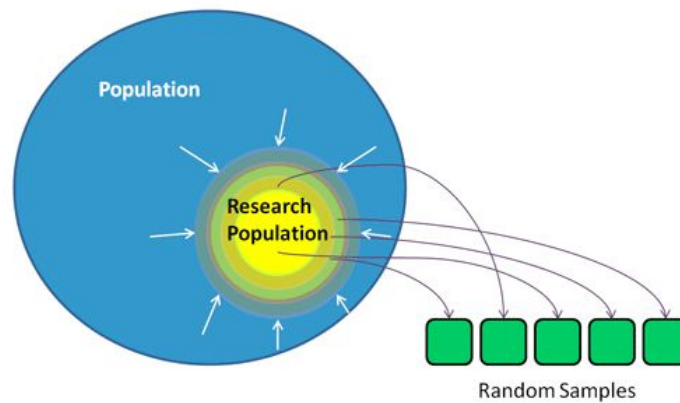


Figure 4.1: Experimental samples are drawn from the “research population” which may be different than the whole population of interest.

External validity is a well-known problem; it’s a fundamental problem in RCT inference and has been noted since the times of Fisher. It is well discussed in many fields: statistics, philosophy, political science, economics, and medicine.

An article in the Lancet (Rothwell, 2005) discusses how RCT’s were designed to have great internal validity, but the pioneers knew about the problem of applicability / generalizability, a problem which has now seemingly been forgotten. He bemoans

the ignorance and lack of guidelines for researchers and sites interesting studies where the treatment effect can vary even by small changes in latitude of the patient.

A recent article (Druckman and Kam, 2011) discusses how using undergraduates as subjects may not be a problem for generalizing experimental results; they argue that undergraduates in fact *do* reflect the American population at-large. Of course they don't speak about the *additional* selection problem that undergraduates are generally paid or given course credit, and those that participate are not even representative of their peers. The problem is open to speculation.



Figure 4.2: External validity can be had at the expense of internal validity and vice versa.

Philosophically, Cartwright (2007) argues that wringing truth out of nature is sort of like an arcade game of whack-a-mole illustrated in figure 4.2. “Clinching” methods, such as RCTs and theory, prove the conclusion but are narrow in scope; “vouching” methods, such as qualitative comparative analysis hint at the conclusion without unequivocal proof, but are more broader in application.

4.1.1 Heterogeneous Effects

External validity is suspect when it can be proven that there exists both:

- (a) *selection bias* into the study itself, meaning that the “research population” is different than the “total population” on one or many demographic variables,

$$\mathbf{X} := X_1, \dots, X_p.$$

(b) *an interaction* between treatment and demographic variables.

If the true model were as follows, selection bias into the study itself would be unproblematic.

$$\mathbf{Y} = \beta_T \mathbf{1}_T + f(\mathbf{X}) + \mathcal{E}$$

As we move from the research population to the true population, the f function will change, but the ATE, $\mathbb{E}[Y | \mathbf{1}_T = 1] - \mathbb{E}[Y | \mathbf{1}_T = 0] = \beta_T$, would remain unchanged.

However, if there are interactions between the treatment and the demographic variables,

$$\mathbf{Y} = \beta_T \mathbf{1}_T + f(\mathbf{1}_T, \mathbf{X}) + \mathcal{E}$$

there is no longer an “average treatment effect” unless we either slice the population on fixed levels of \mathbf{X} or expectorate over the population covariate distribution, since:

$$\mathbb{E}[Y | \mathbf{1}_T = 1] - \mathbb{E}[Y | \mathbf{1}_T = 0] = \beta_T + f(1, \mathbf{X}) - f(0, \mathbf{X})$$

whose behavior as subjects move around \mathbf{X} space may vary considerably due to the unknown properties of the f function.

4.1.2 Goals of the Study and Outline of this Report

This study attempts to detect if external validity is compromised in *behavioral* studies which involve text manipulation run on Amazon’s Mechanical Turk (MTurk). We pick

three studies that are designed to elicit cognitive biases. The first measures the price of beer between two groups that experience differential framing from a luxury hotel or a run-down grocery store. The second measures whether people decide to cooperate with one another after being primed with religious text. The third measures the sunk cost fallacy, we ask whether or not people are likely to go to a theatre event after paying dearly for tickets or being given them for free.

We measure if there is a difference if people are administered a study at random or if they are allowed to select which study they want to participate in when they are told the general idea of the study. This “selection” wing is designed to create different research populations for each study.

For the reader unfamiliar with MTurk and/or natural field experiments, we suggest reading Chandler and Kapelner (2013) where it is described how one major external validity concern that is manifest in most studies, laboratory bias, is *not* characteristic of carefully designed MTurk experiments. We explain the experiment in detail in section 5.2, analyze the data in 4.3, and conclude in section 4.4.

4.2 Experimental Methods and Design

We create a HIT on MTurk paying \$0.05 posing as a major marketing company asking for consumer opinions (see figure 4.3). Subjects were not told they were part of a study in order to create a real-effort task in a *natural field experiment*. We limited experimental subjects to be from America because the experimental manipulations relied on fluency in English.

Once the HIT was accepted, we move them to a demographic survey page pictured in figure 4.4 (for a list of questions asked, refer to appendix A.2.1). To avoid any systematic bias, we randomized the order of these questions as well as their answer choices (within reason).

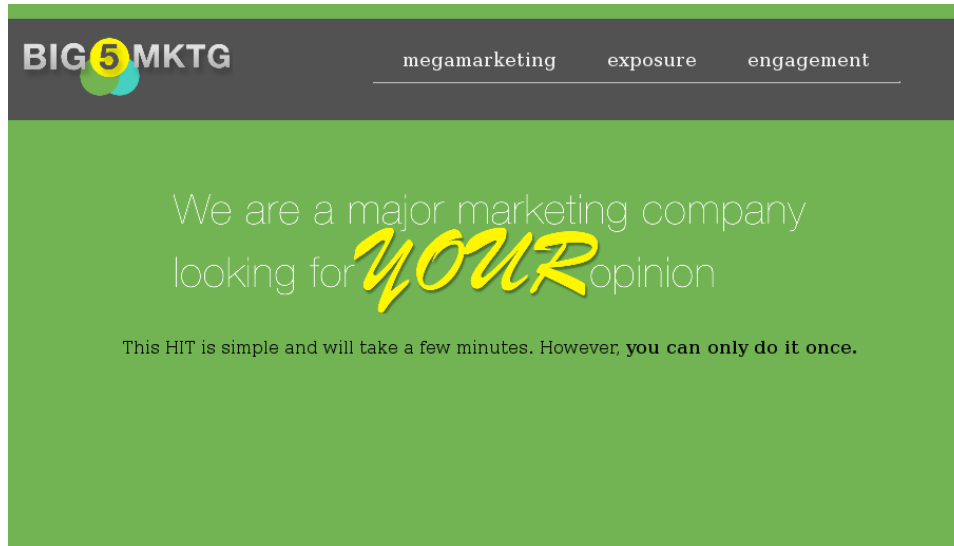


Figure 4.3: The splash page shown to a Turker viewing our HIT for the first time. The Turker will make their decision whether or not to participate in the task based on this screen and the wage offered.

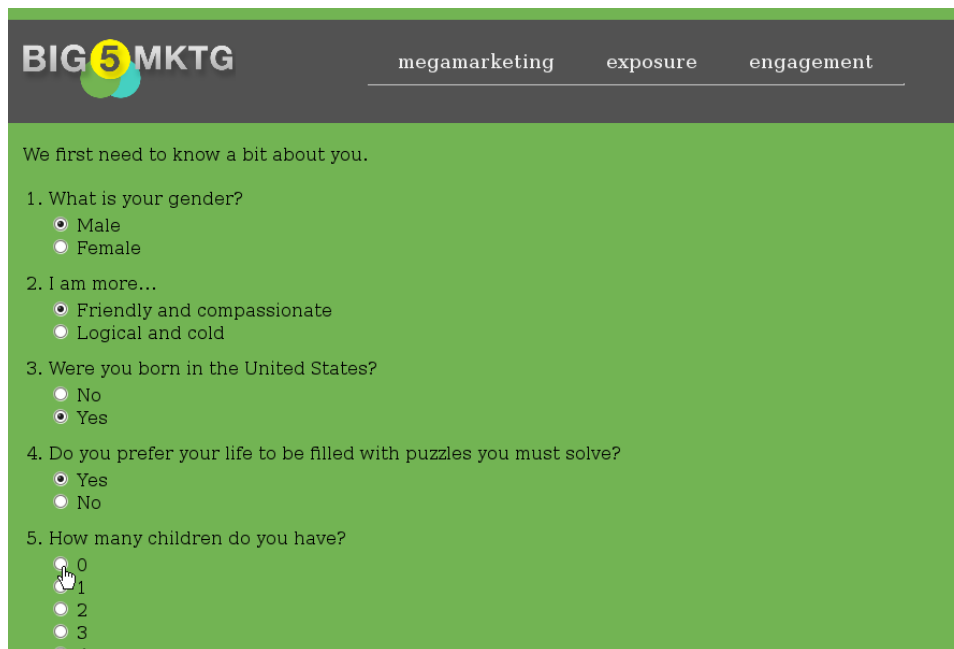


Figure 4.4: Part of the survey page which collects demographic covariates about the subjects.

Upon completion of this survey, they were randomized into the “random study” wing, which we’ll call “R,” or the “select study” wing which we’ll denote “S.”¹ If they were placed into the “R” wing, they were then randomized into one of the three studies described in the next subsection. If they were placed into the “S” wing, the subjects were allowed to choose which study they wanted to participate in. This selection is described more fully in a following subsection. Once they were assigned one of the three studies, they were further randomized evenly between treatment and control particular to the study.² The experimental flowchart illustrating this description is pictured in figure 4.5.

4.2.1 The Three Studies

We are trying to detect sampling bias into experiments on MTurk for behavioral experiments whose experimental manipulations involve text. To make any general claim, studies would ideally be sampled from the space of all possible studies that fit this criteria.

Our initial thought would be to sample from cognitive bias studies, of which there are many, and they are currently of great interest to behavioral scientists. We selected three studies investigating three different cognitive biases: framing, priming, and sunk cost. Each of the studies’ experimental manipulation consisted of *subtle* text changes between the control and the treatment. We briefly describe the studies below. Full text of the experimental screenst can be found in appendix A.2.3.

¹For this randomization, we used complete randomization with a standard bernoulli. This would prevent any systematic bias in the order in which the participants showed up but would most likely result in inbalance. Since we planned on collecting a large number of subjects, an imbalance here was not a concern.

²Once again, we employed complete randomization. We admittedly should’ve used a block design here to obtain higher power because any bias from time admitted is most likely taken care of by the R/S randomization.

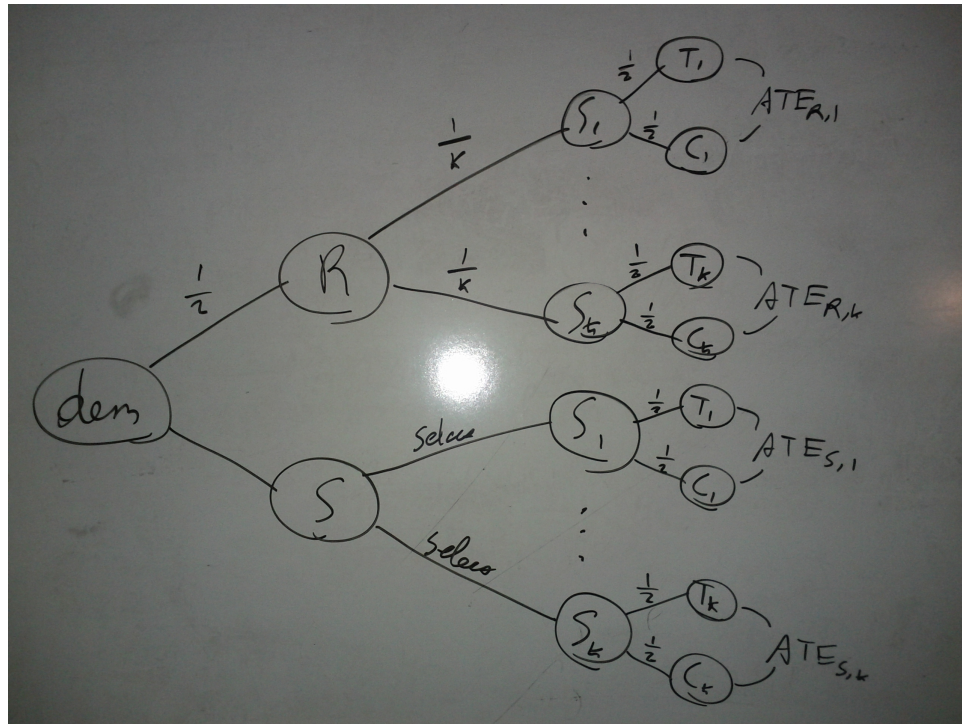


Figure 4.5: The experimental flow: wings and randomizations. The above illustration is for any number of studies. Here $k = 3$.

People are known to differentially pay higher or lower prices for identical items based on how the item is “framed.” We take a verbatim study from Thaler (1985) where participants are asked how much they would pay for a beer which comes from either a fancy resort hotel (treatment) or a run-down grocery store (control). The outcome measure would be a real number greater than zero.

It is known that people differentially change their behavior if they were “primed” before with religious ideas. We test whether subjects differentially choose to cooperate in a prisoner’s dilemma if they were given an excerpt from the bible about charity (treatment) or an excerpt of an encyclopedia page about fish (control). This experiment is a replication of one found in Horton and Chilton (2010) with minor modifications. The outcome measure is cooperate or not cooperate.

People are also known to differentially make decisions to consume if based on a cost

that has previously been sunk. According to economic theory, the cost already sunk should not affect the amount of utility, or enjoyment, one should get from a product. We design a study where the subject is hypothetically going to a theatre production on a night with extreme weather. We test whether the subjects are more likely to indicate a strong desire to go when they paid dearly for the tickets (treatment) or received tickets for free (control). The response was measured on a Likert scale from zero to nine.

Since the studies rely on subtly text manipulations, we employ the anti-satisficing technique of fading in each word of the experiment text at a high school reading page to ensure respondents pay more attention to the text (Kapelner and Chandler, 2010).

4.2.2 The Selection Wing

How are subjects to choose which study to participate in when they are randomized into the “S” wing? The purpose of the experiment is to detect selection bias into the experiments; thus, we ideally want to bias the subjects *as much as possible* into their study of preference to create the most conspicuous heterogeneous effects. We attempt to do this *without* giving away what the study is trying to measure (which would raise concerns about the studies’ internal validity).

We chose to give subjects the screen shown in figure 4.6. They know the marketing question-of-interest has something to do with beer, religion, or theatre, but they don’t know what the treatment manipulations are. This can create heterogeneous effects in two ways (1) having knowledge about the question at-hand may bias their responses and induce a different ATE vis-a-vis subjects who did not have prior knowledge (2) subjects who are more interested in “beer” may represent a different subpopulation than those who are interested in “religion” and likewise those who are interested in “theatre.”

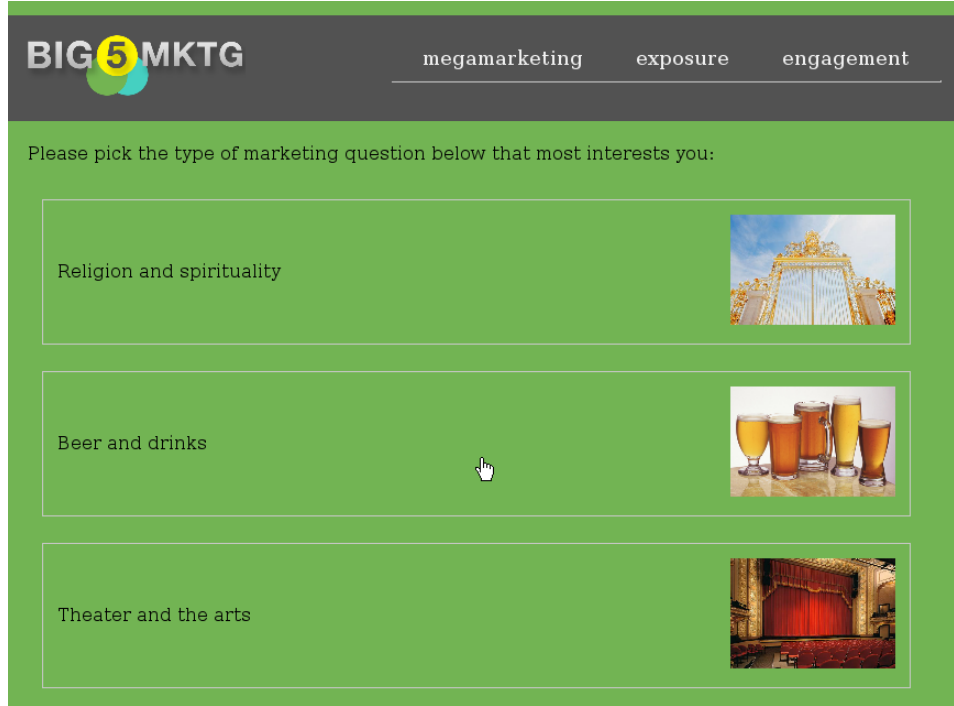


Figure 4.6: The screen shown to subjects in the “S” wing immediately after completing the demographic survey.

4.2.3 Formal Experimental Design

We are trying to demonstrate heterogeneous effects which amount to a differential ATE between the “R” and “S” wings. This can be tested via the following model (its design matrix is illustrated in figure 4.7). For study k ,

$$\mathbf{Y}_k = \mathbf{X}\boldsymbol{\beta} + \beta_{T,k}\mathbf{1}_{T_k} + \beta_{S,k}\mathbf{1}_S + \beta_{ST,k}\mathbf{1}_{T_k}\mathbf{1}_S + \boldsymbol{\mathcal{E}}, \quad k \in \{1, 2, 3\}$$

The $\boldsymbol{\beta}$ represents the nuisance parameters for the slopes of the demographic variables. The $\beta_{T,k}$'s are the ATE's for each study in the “R” wing which is of ancillary interest — we only wish to check these to ensure the study worked as designed to arrive at an estimate to the causal effect of the manipulations. The $\beta_{S,k}$'s are offsets for belonging to the “S” wing. Being allowed to select the study may shift your level

$$\mathbf{X} = \begin{bmatrix} X_{\cdot 1} & \dots & X_{\cdot p} & \mathbf{1}_S & \mathbf{1}_T & \mathbf{1}_S \mathbf{1}_T \\ \vdots & \dots & \vdots & \mathbf{1}_{n_S} & \mathbf{1}_{n_{ST}} & \mathbf{1}_{n_{ST}} \\ \vdots & \dots & \vdots & \vdots & \mathbf{0}_{n_S - n_{ST}} & \mathbf{0}_{n_S - n_{ST}} \\ \vdots & \dots & \vdots & \mathbf{0}_{N - n_S} & \mathbf{1}_{n_{RT}} & \mathbf{0}_{N - n_S} \\ \vdots & \dots & \vdots & \vdots & \mathbf{0}_{N - n_S - n_{RT}} & \vdots \end{bmatrix}$$

Figure 4.7: The design matrix for each of the k studies.

of response; this is interesting to check, but still ancillary to our question of interest. The $\beta_{ST,k}$ are the parameters of interest. They model the differential ATE between the “R” and “S” wings which is the average heterogeneous effect induced by having two different research populations participate in the same study. Based on our hypothesis, we wish to reject these being zero:

$$H_0 : \beta_{ST,1} = 0, \quad H_0 : \beta_{ST,2} = 0, \quad H_0 : \beta_{ST,3} = 0$$

For each of the studies, we have a 2×2 factorial design with n_S random (thus it cannot be balanced).

4.2.4 Pilot Study to Estimate Power

We first ran a pilot study of $N = 100$ to inform ourselves about power for the full study. For the framing study, Y is continuous, but for the sunk cost study, Y is ordinal, and for the priming study, Y is dichotomous, so the following calculation is for the framing effect study at $\alpha = 5\%$ since power calculations for non-continuous responses are more difficult.

We estimate MSE of the pilot study regression to be 11.42 and we assume this to

be our variance of the experimental error term going forward.³ We wish to calculate power for many possible values of β_{ST} which is the interaction effect in our factorial design. Letting r be the number of duplicates in each wing-treatment cell, N be the total sample size for the study, power can be calculated following Morris (2011, page 159):

$$F^* := F(95\%, 1, 4(r - 1)), \quad W \sim F\left(1, 4(r - 1), \frac{4r\beta_{ST}^2}{\sigma^2}\right), \quad \text{POW} := \mathbb{P}(W > F^*)$$

Power for a range of possible r and β_{ST} values appears below in table 4.1.

N_{study}	r	β_{ST}							
		0.01	0.05	0.1	0.15	0.2	0.25	0.5	0.75
400	100	0.05	0.06	0.09	0.14	0.22	0.31	0.84	0.99
800	200	0.05	0.07	0.13	0.24	0.39	0.55	0.99	1.00
1200	300	0.05	0.08	0.18	0.34	0.53	0.73	1.00	1.00
1600	400	0.05	0.09	0.22	0.43	0.66	0.84	1.00	1.00
2000	500	0.05	0.10	0.26	0.51	0.75	0.91	1.00	1.00
2400	600	0.05	0.11	0.30	0.58	0.83	0.95	1.00	1.00
2800	700	0.05	0.12	0.35	0.65	0.88	0.97	1.00	1.00

Table 4.1: Power by number of duplicates and interaction effect size.

³The variance found after the experiment was complete was around 14, so this was a good estimate.

4.3 Results

We hired 2112 total Turkers as subjects with total cost of about \$116. Sample sizes are reported for each wing and treatment cell for all studies in table 4.2. Note that r , the cell size, varies between 140 and 214.

Arm	Study	Treatment	Control	Total	
R	Framing	169	172	341	
	Priming	171	176	347	1053
	Sunk Cost	199	166	365	
S	Framing	167	145	312	
	Priming	203	214	417	1059
	Sunk Cost	169	161	330	

Table 4.2: Experimental sample sizes.

We analyzed data from all studies using ordinary least squares⁴ to estimate parameters noted in figure 4.7. The nuisance parameters include the demographic variables discussed in appendix A.2.1 as well as other experimental covariates we collected from subject behavior which we describe in appendix A.2.2. We summarize our main results in table 4.3, results for the framing study in figure 4.8, results for the priming study in figure 4.9, and results for the sunk cost study in figure 4.10.

In short, we did not detect significant heterogeneous effects in any of the three studies. Referencing table 4.1, we note that we only have sufficient power to detect interaction effects above \$0.50 in the framing study.

⁴Technically, this is inappropriate for the priming study which has dichotomous response and the sunk cost study which has ordinal response. Since the data turned out too noisy for effect detection, we did not try other models because we feel they would have been equally unsuccessful.

	Framing Study			Priming Study			Sunk Cost Study		
	b_T	s_{b_T}	p_{val}	b_T	s_{b_T}	p_{val}	b_T	s_{b_T}	p_{val}
ATE in “R” wing	0.69	0.43	0.107	0.06	0.03	0.078	0.19	0.29	0.507
Offset for “S” wing	0.32	0.44	0.475	-0.01	0.03	0.662	-0.20	0.31	0.525
Heterogeneous Effect	-0.19	0.61	0.752	-0.04	0.04	0.369	-0.06	0.42	0.891

Table 4.3: Experimental estimates of ATE, selection offset, and average heterogeneous effects for all studies.

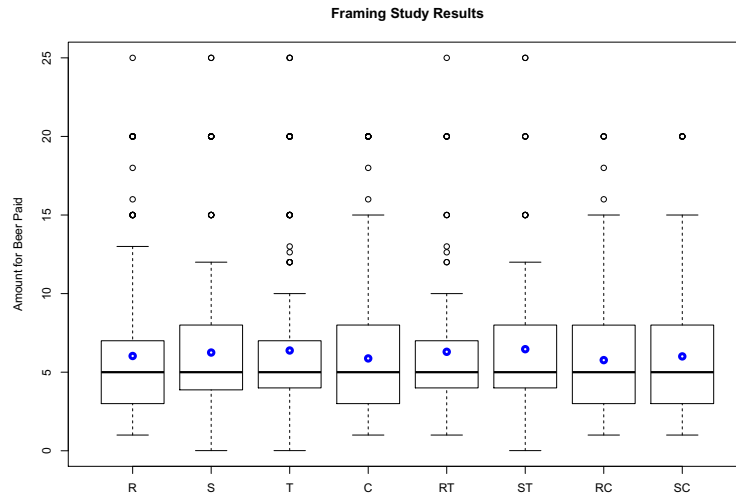


Figure 4.8: Box and whisker plots for the framing study by row, column, and cell in the factorial design. Blue dots indicate sample averages.

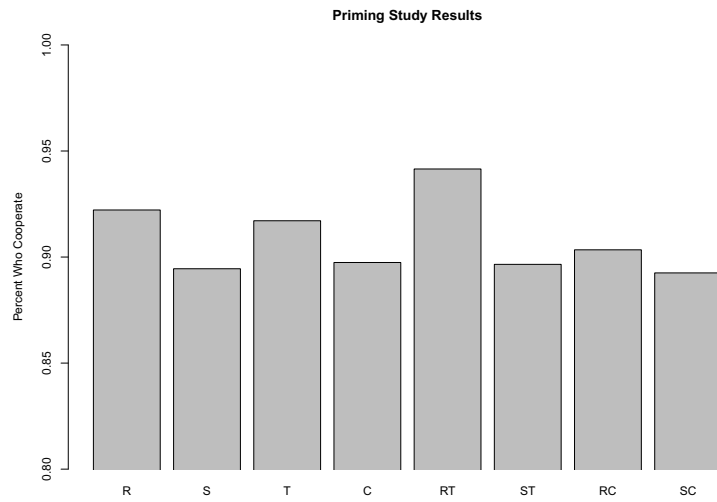


Figure 4.9: Sample proportions of cooperate in the priming study by row, column, and cell in the factorial design.

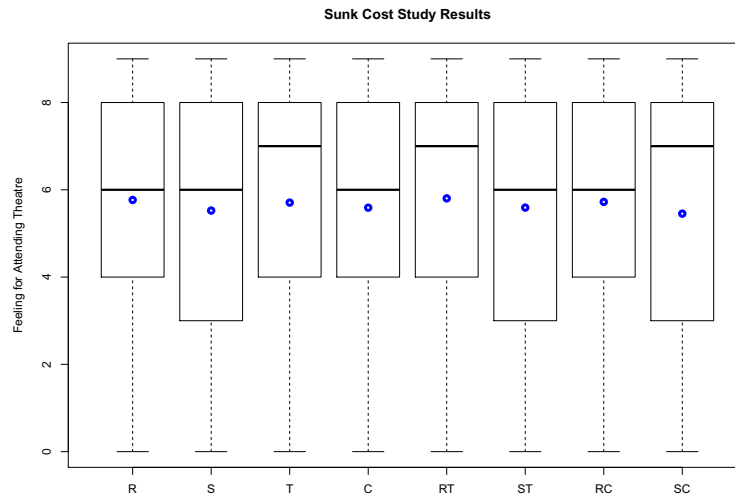


Figure 4.10: Box and whisker plots for the sunk cost study by row, column, and cell in the factorial design. Blue dots indicate sample averages.

4.4 Conclusions and Future Directions

We ran a natural field experiment on MTurk to detect selection-into-experiment bias for three behavioral studies. We did so by placing half of the participants into a random study and allowing the other half to pick the study they wished to complete. We then tried to measure selection-into-experiment bias by calculating differential ATE estimates between the random study participants and the select-your-own study participants.

We did not find significant heterogeneous treatment effects for any of the studies between the two populations either because they don't exist or because we were underpowered to find them. If we were to make a recommendation to investigators using MTurk to study behavioral phenomena using experiments that involve text manipulations, we would still recommend to not let them know anything about the study, but if they had to inform the participants it most likely would not affect the results.

We would like to run the study again with better design. First, each of the three studies should be designed to elicit a larger ATE. This would allow more room for heterogeneous effects. Second, we should employ techniques to reduce noise in the data. We are currently developing sequential matching methods for this very purpose. Third, the selection mechanism should be better designed.

Acknowledgements

Thanks to Dean Foster, Abba Krieger, Mary Putt and Paul Rosenbaum for helpful discussions.

Matching on-the-fly in Sequential Experiments*

Abstract

We propose a dynamic allocation procedure that increases power and efficiency when measuring an average treatment effect in fixed sample size sequential randomized trials. Subjects arrive iteratively and are either randomized *or* paired via a matching criterion to a previously randomized subject and administered the alternate treatment. We develop estimators for the average treatment effect that combine information from both the matched pairs and unmatched subjects as well as an exact test. Simulations illustrate the method's higher efficiency and power over competing allocation procedures in both controlled scenarios and historical clinical trial data.

5.1 Introduction

Randomization in experimentation is likely to work well with large sample size. With small sample size, the empirical distributions of relevant covariates can be different across treatments possibly masking an effect by creating bias and inflating variance.

*Joint work with Abba Krieger

Some improvements over completely randomized design are rerandomization (Morgan and Rubin, 2012), *a priori* matching (Raudenbush et al., 2007), and *adaptive design* which involves any change to the experiment or statistical procedures while the experiment is underway (Chow and Chang, 2008).

We limit our focus to *sequential experiments*, where subjects enter iteratively over time and the experimental condition is administered upon entrance, but the outcome does not necessarily have to be assessed before the next subject is enrolled. We develop a new adaptive design for sequential experiments whose goal is to elucidate an average treatment effect (ATE) between two arms, which we call treatment (T) and control (C). Sequential experiments are very popular in both clinical trials and recently, crowdsourced-Internet experimentation (Horton et al., 2011; Chandler and Kapelner, 2013).

Our design is a new type of *dynamic allocation*, a means of assigning T/C to newly-arrived subjects based on decisions about previous assignments, covariates, and/or responses (Hu and Rosenberger, 2006). Proposals of dynamic allocation began with Efron (1971)'s biased coin design. Here, a coin is biased in favor of the treatment with the fewest subjects, hence leading to better balance in treatment allocation. However, this procedure ignores subjects' covariates.

The first line of defense to balance covariates is stratification (or "blocking"), a classic strategy dating back to Fisher's agricultural experiments. Stratification becomes quickly impractical when the number of total covariate levels is large relative to sample size. Taves (1974) and others tackle these shortcomings by "minimizing" the imbalance between the treatments among all levels of covariates present. The most popular and widely implemented among these methods is found in Pocock and Simon (1975) whose procedure involves picking a few arbitrary functions to tailor the imbalances. The selection of these functions is still an ongoing area of research (for instance, see Han et al., 2009). Concerned by this arbitrariness, Atkinson (1982)

posits a method solidly rooted in linear model theory using D_A optimality.

If the end goal of experimentation is to find an effect when one is present, then the primary concern is estimator efficiency and test power. Stratification and minimization methods rely on the logic that greater balance among the covariates implies greater efficiency which is mathematically true *only* in homoskedastic linear models (Rosenberger and Sverdlov, 2008). D_A optimality iteratively maximizes efficiency assuming the linear model without explicitly focusing on balance. Thus, we see one of the fundamental problems in previous allocation procedures is the reliance on the homoskedastic linear model, an assumption that is rarely true in practice. We wish to develop a dynamic allocation which is robust when the covariates combine non-linearly and with interactions to produce the response and no worse than current methods when the linear model holds.

The seminal guidebook, Cook and Campbell (1979), states “whenever possible, it is recommended to minimize error terms.” They then recommend *matching* subjects before randomization on covariates to create better stratification. It was not a novel idea; Student (1931) commented on the $n = 20,000$ children Lanarkshire Milk Experiment proposing the experiment should be performed exclusively on 50 identical twin pairs which would be randomly assigned T/C.

We propose matching iteratively, *on-the-fly*. As subjects walk in the door or engage a survey online they should be matched with people “similar” to them who came in previously, a procedure which Whitehead (1997) believes to be “especially difficult.” Imagine the following scenario of a trial testing whether a pill decreases blood pressure. The investigators determine that age, height, weight, and race should be collected as covariates as they are known to be related to blood pressure. Bob, 28, 5’10”, 180lb and White enters and is determined to fit the requirements of the study. By the flip of a coin, he receives the pill. The next day, Grace, 45, 5’2”, 105lb and Asian enters. Based on inspection of this demographic data, she is clearly different

from Bob; thus she is also randomized. Soon after, Joe, 29, 5'11", 185lb and White enters. We determine that he is similar to Bob, pair them, and deterministically administer to him the placebo. The trial continues and Grace would then await someone to be matched with, which may or may not occur.

The algorithm is simple: incoming subjects are either randomized and placed in a holding pool, called the "reservoir," or if they're found to match a subject already in the reservoir, they're matched and given their match's alternate treatment. The matches and the reservoir form two independent samples yielding two different estimators which are combined to estimate the ATE.

The closest idea we find in the literature is in Raghavarao (1980) who computes the Mahalanobis distances between a new subject's covariates and the average covariates in the different treatment groups. The allocation to the treatment is then made via a biased coin with probabilities proportional to these distances. We use the idea of Mahalanobis distance which creates robustness to collinearity in the covariates, but we use it to match individual subjects together in pairs.

We layout our scenario assumptions, explain our algorithm and develop testing procedures in Section 5.2. We demonstrate our procedure's improvements over previous procedures via simulations in Section 5.3. Our method performs particularly well in the case where the model is non-linear, performs respectably with linear models, and also performs respectively when the covariates do not inform the response. We then demonstrate higher efficiency using historical data from randomized controlled trials (RCT's) in Section 5.4, where the covariate-response model was unknown but most likely non-linear. We discuss and describe future directions in Section 9.6.

5.2 The Algorithm, Estimation, and Testing

5.2.1 Problem Formulation

Subjects arrive sequentially and their covariates, denoted by $\mathbf{x}_i := [x_{i1}, \dots, x_{ip}]$ which are either continuous or binary, are immediately recorded. The subjects must then be assigned to a treatment on-the-spot. We develop our method for allocating one of two treatments, T or C, to subject i denoted by the treatment indicator $\mathbb{1}_{T,i}$. The response y_i can be collected at any time after allocation. We assume the following model with independent observations, an additive treatment effect, a possibly non-linear covariate effect, normal and homoskedastic errors, fixed covariate design, and sample size n fixed in advance:

$$\begin{aligned} Y_i &= \beta_T \mathbb{1}_{T,i} + z_i + \mathcal{E}_i, & z_i &:= f(\mathbf{x}_i), \\ \mathcal{E}_i &\stackrel{iid}{\sim} \mathcal{N}(0, \sigma_e^2), & i &\in \{1, \dots, n\}. \end{aligned} \tag{5.1}$$

We wish to develop a dynamic allocation method followed by an unbiased estimator for β_T with higher efficiency and thereby more powerful when testing classic null hypotheses compared to popular approaches.

5.2.2 The Algorithm

The first n_0 subjects enter the experiment and are randomized to T or C with the flip of a coin. These subjects comprise the initial “reservoir.” After a certain point, we would like to potentially match an incoming subject with subjects in the reservoir. We would like to match them on $f(\mathbf{x})$, which is latent, so we match on what we consider is the next best thing, the \mathbf{x} ’s themselves. We hope that $\mathbf{x}_1 \approx \mathbf{x}_2$ implies

$f(\mathbf{x}_1) \approx f(\mathbf{x}_2)$ which is true if the function is sufficiently smooth.

We match using squared Mahalanobis distance which gives a convenient scalar distance between points in \mathbb{R}^p adjusting for collinearities. This metric has a long implementation history in matching applications dating back to Rubin (1979). Matching using Mahalanobis distance and then randomizing the pairs to T/C has been demonstrated to result in better balance and higher power (Greevy et al., 2004). Further, the assumption of normal covariates seems to work well with real data even when the covariates are non-normal (see Section 5.4). After matches are produced, we do not make use of the normality assumption of the \mathbf{x} 's further in the development of the estimator. Improvements to the matching machinery that may be more robust to real-world covariate distributions are also discussed in Section 9.6.

Thus, the new subject enters and the squared Mahalanobis distance between its covariate vector, \mathbf{x}_{new} , and each of the previous subject covariate vectors, the \mathbf{x}_{old} 's, are calculated. Denote \mathbf{S} as the covariates' sample variance-covariance matrix calculated with all subjects including the new subject. Assuming normal covariates, the squared Mahalanobis distance then has a scaled F distribution given below:

$$D^2 := (\mathbf{X}_{\text{new}} - \mathbf{X}_{\text{old}})^\top \mathbf{S}^{-1} (\mathbf{X}_{\text{new}} - \mathbf{X}_{\text{old}}), \quad (5.2)$$

$$\frac{n-p}{2p(n-1)} D^2 \stackrel{\text{approx}}{\sim} F_{p, n-p}.$$

We then take the minimum of the squared Mahalanobis distances between the new observation and each observation in the reservoir and calculate its probability. Let the minimum distance squared come from the previous subject, $\mathbf{x}_{\text{old}}^*$. If the probability is less than λ , a pre-specified hyperparameter, then \mathbf{x}_{new} and $\mathbf{x}_{\text{old}}^*$ are matched together; if it's not, \mathbf{x}_{new} is randomized and added to the reservoir. If \mathbf{x}_{new} is matched, it is not added to the reservoir and $\mathbf{x}_{\text{old}}^*$ is removed from the reservoir. $\mathbb{1}_{T, \text{new}}$ is then

assigned to be $1 - \mathbb{1}_{T, \text{old}^*}$, i.e. $\mathbf{x}_{\text{old}^*}^*$'s opposite treatment. The process is repeated until the n^{th} entrant. We left out other implementation details in this discussion but make them explicit in algorithm 1. Note that our proposed procedure is considered a form of *covariate-adaptive randomization* (Rosenberger and Sverdlov, 2008, Section 2) because we are using the covariates to determine the dynamic allocation.

Note that upon matching, the treatment indicator is assigned *deterministically* to be the opposite of its match. This can cause selection bias if the investigator is not properly blinded. In defense of our decision to make allocation deterministic for a little less than half the subjects, note that our algorithm is sufficiently complicated that a duplicitous investigator would not be able to guess whether the entering subject will be assigned T or C based on previous information during a clinical trial. McEntegart (2003) discusses how this type of machination is unrealistic even in multi-center block permuted designs, which is much simpler than the allocation strategy proposed here. Further, if the procedure is implemented in an Internet-based experiment, the algorithm would be hard-coded and would not be subject to human tampering.

5.2.3 Estimation and Hypothesis Testing

We assume the covariate design matrix is fixed, the classic regression assumption. If so, the subjects ultimately matched and the subjects ultimately found in the reservoir are fixed as well. Thus conditioning on the design is equivalent to conditioning on the sigma field given below:

$$\mathcal{F} = \sigma\left(\underbrace{\langle \mathbf{x}_{T,1}, \mathbf{x}_{C,1} \rangle, \langle \mathbf{x}_{T,2}, \mathbf{x}_{C,2} \rangle, \dots, \langle \mathbf{x}_{T,m}, \mathbf{x}_{C,m} \rangle}_{\text{matched pairs}}, \underbrace{\mathbf{x}_{R,1}, \mathbf{x}_{R,2}, \dots, \mathbf{x}_{R,n_R}}_{\text{reservoir}}\right).$$

Algorithm 1 The sequential matching algorithm for subjects entering the experiment. The algorithm requires λ and n_0 to be prespecified, which controls the ease of creating matches.

```

1: for  $t \leftarrow \{1, \dots, n\}$  do ▷  $n$  is the total sample size, fixed a priori
2:   if  $t \leq n_0$  or reservoir empty then
3:      $\mathbb{1}_{T,t} \leftarrow \text{Bern}(\frac{1}{2})$  and subject  $t$  is added to the reservoir ▷ randomize
4:   else
5:      $S_t^{-1}$  is calculated using  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t$  ▷ Estimate the true var-cov matrix
6:     Lookup  $F_{\lambda,p,t-p}^* \triangleright F^*$  is the critical cutoff from the  $F$  distribution quantile
7:     for all  $\mathbf{x}_r$  in the reservoir do
8:        $F_r \leftarrow \frac{t-p}{2p(n-1)}(\mathbf{x}_t - \mathbf{x}_r)^\top S_t^{-1}(\mathbf{x}_t - \mathbf{x}_r)$ 
9:     end for
10:     $F_{r^*} \leftarrow \min\{r\}\{F_r\}$ ,  $r^* \leftarrow \arg \min_r \{F_r\}$  ▷ among ties, randomly select
11:    if  $F_{r^*} \leq F_{\lambda,p,t-p}^*$  then ▷ a match is found
12:       $\mathbb{1}_{T,t} \leftarrow 1 - \mathbb{1}_{T,r^*}$  ▷ assign subject  $t$  the opposite of  $r^*$ 's assignment
13:       $[\mathbf{x}_{r^*}, \mathbb{1}_{T,r^*}]$  is removed from the reservoir
14:      record  $\langle \mathbf{x}_t, \mathbf{x}_{r^*} \rangle$  as a new match
15:    else ▷ a match is not found
16:       $\mathbb{1}_{T,t} \leftarrow \text{Bern}(\frac{1}{2})$  and subject  $t$  is added to the reservoir ▷ randomize
17:    end if
18:  end if
19: end for

```

Upon completion of the experiment, there are m matched pairs and n_R subjects in the reservoir, both quantities fixed since the sample size and the design are fixed ($n = 2m + n_R$). In our development of estimators and testing procedures, we always assume conditioning on \mathcal{F} , thus this notation is withheld going forward.

We focus on testing the classic hypotheses $H_0 : \beta_T = \beta_0$ versus $H_a : \beta_T \neq \beta_0$. We consider testing under three model assumptions (a) the response has normal noise and may possibly depend on covariates but we do not wish to model their effect (b) the response has normal noise and depends linearly on covariates (c) the response has mean-centered noise and depends on covariates through an unknown model. We develop testing procedures for each of these models: (a) a modification to the classic $\bar{Y}_T - \bar{Y}_C$ in Section 5.2.3.1, (b) a modification to ordinary least squares regression in Section 5.2.3.2 and (c) an exact permutation test in Section 5.2.3.3.

5.2.3.1 The Classic Test

We define \bar{D} as the estimator for the average of the differences of the m matched pairs (treatment response minus control response) and $\bar{R} := \bar{Y}_{R,T} - \bar{Y}_{R,C}$ as the difference in averages of the treatments and controls in the reservoir. We combine these two estimators of the treatment effect using a weight parameter, $w\bar{D} + (1 - w)\bar{R}$. We can find a w^* to minimize variance thereby obtaining the estimator

$$B_T = \frac{\sigma_R^2 \bar{D} + \sigma_D^2 \bar{R}}{\sigma_R^2 + \sigma_D^2} \quad \text{with} \quad \text{Var}[B_T] = \frac{\sigma_R^2 \sigma_D^2}{\sigma_R^2 + \sigma_D^2}. \quad (5.3)$$

B_T is unbiased because under the assumption of fixed covariates, taking the expectation over treatment allocation and noise yields $\mathbb{E}[\mathbb{E}[\bar{D}]] = \mathbb{E}[\mathbb{E}[\bar{R}]] = \beta_T$. Standardizing B_T gives a standard normal due to the assumption of normal noise. To create a usable test statistic, note that the true variances are unknown, so we

plugin S_D^2 , the matched pairs sample variance estimator, and S_R^2 , the pooled two-sample reservoir difference variance estimator. Note that the standard calculation of S_R^2 assumes the number of treatments and controls in the reservoir is fixed in advance but in our algorithm, these quantities are random. A more careful calculation could include this randomness.

Equation 5.4 shows the resulting statistic which has an asymptotically standard normal distribution.

$$\frac{B_T - \beta_0}{\text{SE}[B_T]} \approx \frac{\frac{S_R^2 \bar{D} + S_D^2 \bar{R}}{S_R^2 + S_D^2} - \beta_0}{\sqrt{\frac{S_R^2 S_D^2}{S_R^2 + S_D^2}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \quad (5.4)$$

This asymptotic result holds if both $m \rightarrow \infty$ and $n_R \rightarrow \infty$ since $S_D^2 \xrightarrow{p} \sigma_D^2$ and $S_R^2 \xrightarrow{p} \sigma_R^2$. The results also holds if $m \rightarrow \infty$ and n_R is bounded which would occur in practice as the covariates are typically categorical or have compact support. In this case, $S_R^2/(S_R^2 + S_D^2) \xrightarrow{a.s.} 1$ which reduces Equation 5.4 above to $(\bar{D} - \beta_0)/S_{\bar{D}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$. Also, by the assumption of additive and normal noise, the estimator is unbiased for finite n .

Note that in the case where there are no matched pairs, we default to the classic estimator and in the case where there are fewer than two treatments or controls in the reservoir, we default to the matched pairs estimator.

When is this estimator more efficient than the standard classic estimator, $\Delta \bar{Y} := \bar{Y}_T - \bar{Y}_C$? In other words, when is $\sigma_{\Delta \bar{Y}}^2 / \sigma_D^2 > 1$? Assuming perfect balance in its treatment allocation ($n_T = n_C = \frac{n}{2}$) for the classic estimator and taking the expectation over both noise and treatment allocation, it can be shown that the variances are:

$$\begin{aligned}\sigma_D^2 &= \frac{1}{m^2} \sum_{k=1}^m (z_{T,k} - z_{C,k})^2 + \frac{2}{m} \sigma_e^2, \\ \sigma_{\Delta Y}^2 &\approx \frac{4}{n^2} \sum_{i=1}^n z_i^2 + \frac{4}{n} \sigma_e^2.\end{aligned}\tag{5.5}$$

This means that the better the matching, the smaller $\sum_{k=1}^m (z_{T,k} - z_{C,k})^2$ will be, the smaller the variance becomes, and the higher the power. If we further allow $n_R = 0$ (*all* the subjects matched), then it's clear that $\sigma_{\Delta Y}^2 / \sigma_D^2 > 1$ if and only if $\sum_{k=1}^m z_{T,k} z_{C,k} > 0$. Note that the approximation in the last expression is due to ignoring covariance terms which do not exist when conditioning on n_T and n_C .

5.2.3.2 The Least Squares Test

To construct a test when the response is linear in the covariates or when we wish to make linear adjustments, we extend the idea in the previous section where we combined an effect estimate from the matched pairs data and an effect estimate from the reservoir data to regression models. Consider the model for the response differences among the matched pairs: $D_k = \beta_{0,D} + \beta_{1,D} \Delta x_{1,k} + \dots + \beta_{p,D} \Delta x_{p,k} + \mathcal{E}_{k,D}$ where $\mathcal{E}_{k,D}$ is $\overset{iid}{\sim}$ normal noise and k ranges from 1 to m . The parameter of interest is the intercept, $\beta_{0,D}$, with OLS estimator $B_{0,D}$, the analogue of \bar{D} in the previous section. $\Delta x_{1,k}, \dots, \Delta x_{p,k}$ are the differences between treatment and control within match k for each of the p covariates respectively. $\beta_{1,D}, \dots, \beta_{p,D}$ are nuisance parameters that adjust for linear imbalances in the covariate differences not accounted for in the matching procedure.

For the responses in the reservoir, consider the classic model: $Y_i = \beta_{T,R} \mathbb{1}_{T,R,i} + \beta_{0,R} + \beta_{1,R} x_{1,i} + \dots + \beta_{p,R} x_{p,i} + \mathcal{E}_{i,R}$ where $\mathcal{E}_{i,R}$ is $\overset{iid}{\sim}$ normal noise and i ranges from 1 to n_R . The parameter of interest is the additive effect of the treatment, $\beta_{T,R}$, with

OLS estimator $B_{T,R}$, the analogue of \bar{R} in the previous section. $\beta_{0,R}, \beta_{1,R}, \dots, \beta_{p,R}$ are nuisance parameters that adjust for a mean offset and linear imbalances in the covariates.

Using the parallel construction in Equations 5.3 and 5.4, our modified OLS estimator has the form

$$\frac{\frac{S_{B_{T,R}}^2 B_{0,D} + S_{B_{0,D}}^2 B_{T,R}}{S_{B_{T,R}}^2 + S_{B_{0,D}}^2} - \beta_0}{\sqrt{\frac{S_{B_{T,R}}^2 S_{B_{0,D}}^2}{S_{B_{T,R}}^2 + S_{B_{0,D}}^2}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \quad (5.6)$$

with the same asymptotics described in the last section below Equation 5.4. $S_{B_{T,R}}^2$ is the sample variance of $B_{T,R}$ and $S_{B_{0,D}}^2$ is the sample variance of $B_{0,D}$.

5.2.3.3 The Permutation Test

An application of Fisher's exact test is straightforward. For the matched pairs component of the data, we examine the 2^m configurations (each match can have T-C or C-T) to compute all \bar{d} 's. For the reservoir portion of the estimator, we condition on $n_{R,T}$ and examine every possible arrangement of the treatment vector to compute every $\bar{y}_{R,T} - \bar{y}_{R,C}$. For each arrangement, we also compute s_D^2 and s_R^2 to create values of the test statistic in Equation 5.3. In practice, the 2-sided p -value is approximated by comparing the observed $|b_T|$ from the true sample data to absolute values of Monte-Carlo samples from the space of all possible test statistics. A similar exact test is available using the modified OLS estimates but we do not explore it in this study.

5.2.4 Properties of the Reservoir Size

We wish to gain insight about how λ , n_0 and n affect n_R . Assume for now that we only have one covariate, x (which may also be the largest principal component of a collection of covariates) and we allow immediate matching ($n_0 = 1$). Mahalanobis distance matches on standardized distance. For this illustration, assume we match if the two x 's sample quantiles are within λ of each other. For example, the latest subject in the experiment had a sample quantile of 0.96, they would be matched to the closest subject in the reservoir with quantile between 0.91 and 1 at $\lambda = 0.10$.

Consider dividing the unit interval into $K := 1/\lambda$ intervals of equal length. Two items in one interval qualify to be matched. Assume that K is even (similar results follow for K odd). Consider the Markov process that transitions after each pair of subjects. Let s be the state that $2m$ of the K cells are occupied for $s \in \{0, 1, \dots, K/2\}$. It is straightforward that $P_{i,j}$, the transition probability of pairs from state i to j , satisfies:

$$P_{s,j} = \begin{cases} \frac{2s(2s-1)}{K^2}, & j = s - 1, \quad s \neq 0 \\ \frac{K(4s+1)-8s^2}{K^2}, & j = s \\ \frac{K^2 - K(4s+1) + 2s(2s+1)}{K^2}, & j = s + 1, \quad s \neq \frac{K}{2}. \end{cases}$$

Note the inherent symmetry: $P_{s,j} = P_{K/2-s, K/2-j}$. Hence, the steady-state probabilities are symmetric about $s = K/4$. Therefore, the mean number of items in the reservoir goes to $K/2 = (2\lambda)^{-1}$ as n grows arbitrarily large. For example, $\lim_{n \rightarrow \infty} \mathbb{E}[N_R \mid \lambda = 0.10] = 5$.

5.3 Simulation Studies

We demonstrate our method’s performance by simulating in three scenarios: covariates affect the response non-linearly (the “NL” scenario), covariates affect the response linearly (the “LI” scenario) and covariates do not affect the response (the “ZE” scenario). These scenarios were simulated via the settings found in Table 5.1. In practice, we simulated many settings for the NL and LI scenarios with similar results.

Scenario	Y_i
NL	$\beta_T \mathbf{1}_{T,i} + x_{1,i} + x_{2,i} + x_{1,i}^2 + x_{2,i}^2 + x_{1,i}x_{2,i} + \mathcal{E}_i$
LI	$\beta_T \mathbf{1}_{T,i} + 2x_{1,i} + 2x_{2,i} + \mathcal{E}_i$
ZE	$\beta_T \mathbf{1}_{T,i} + \mathcal{E}_i$

Table 5.1: The response models for the three scenarios proposed. The covariates were $X_{1,i} \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ and $X_{2,i} \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ and the errors were $\mathcal{E}_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$.

We set the treatment effect to be $\beta_T = 1$. n was varied over $\{50, 100, 200\}$, n_0 was set to the minimum to admit calculation of \mathbf{S}^{-1} (at $n_0 = p = 2$) and λ was varied over nine settings between 0.01 and 0.75. We then used σ_e^2 to modulate the resolution in our comparisons. We chose $\sigma_e^2 = 3$ to be a good balance because even at $n = 200$ comparisons were clear.

In choosing which competitor dynamic allocation methods to simulate against, we wanted to pick methods that are in use in sequential trials. According to Scott et al. (2002), stratification is very popular and Efron’s biased coin has been used in a few studies. Most popular is minimization which has been used in over 1,000 trials (McEntegart, 2003).

Thus, we choose to compare our method against (1) complete randomization (CR).

(2) stratification: both x_1 and x_2 were stratified into three levels based on the tertiles of the standard normal distributions, creating $3 \times 3 = 9$ blocks. Within blocks we alternate T / C in order to coerce $n_T \approx n_C$ so no power is lost on allocation imbalance. (3) Efron's biased coin design (BCD): we use the bias parameter of $\alpha = 2/3$ which is Efron's "personal favorite." (4) Minimization: we used the same blocks as stratification, the "D" function (Begg and Iglewicz, 1980 found the variance method performed slightly better than alternatives), the "G" function, and we set $p = 1$ for deterministic assignments in order to force $n_T \approx n_C$ so no power is lost on allocation imbalance. (5) stratification as above followed by post-matching which we detail in Web Appendix A in the Supplementary Materials.

Note that power for our stratification and minimization simulations may be higher than what is expected in practice due to two reasons. First, we use deterministic assignments. Second and more importantly, real world covariates are collinear which reduces the effectiveness of stratification and minimization. Here, we generate covariates independently. Note that collinearity is not an issue for our sequential matching proposal due to the use of Mahalanobis distance.

There are three scenarios (NL, LI, and ZE) and four competitors. Naturally, we want to gauge performance if we assumed the correct underlying model, but we also want to ensure we are robust if the model is misspecified. Therefore, we simulate each of these under the three model assumptions discussed in Section 5.2.3. For the classic estimator, all competitors employed $\bar{Y}_T - \bar{Y}_C$; for the linear estimator, all competitors employed OLS; and for the exact test, all competitors employed the standard conditional permutation test.

We hypothesize that in the case of no effects (the ZE scenario), we will slightly underperform against competitors under all three testing procedures because of the loss of power due to the lower effective sample size when analyzing paired differences. If the effects are linear, we hypothesize to do slightly worse against the OLS procedures

due to lower effective sample size. Under all other scenarios and models, we expect to do better.

We simulated each scenario 2,000 times and for exact testing, we Monte-Carlo sampled 1,000 times within each simulation.

Our main metric for comparison is power, the proportion of the times the null was rejected under the Type I error rate of $\alpha = 5\%$. We also record standard error of the estimate (when the estimator was parametric) as well as balance (the maximum standardized difference in the averages of covariates between treatment and control samples: $\max_{j \in \{1,2\}} \{(\bar{x}_{j,T} - \bar{x}_{j,C})/s_j\}$). Results for power under $\lambda = 10\%$ against the null of no treatment effect are illustrated in Figure A.17c and results for balance and relative efficiency vis-a-vis other methods are found in Table 5.2. Power results for other values of λ appear as Web Figures 1 to 8 in the Supplementary Materials. The levels of λ have little effect the comparisons against the four competitors. A table of empirical bias results appears as Web Table 1 in the Supplementary Materials.

In the NL scenario, our sequential matching procedure dominates competitors in power and efficiency, sometimes doubling power and nearly tripling efficiency. Even at $n = 200$, the gains are large. Regression adjustment helps the competitors, but it cannot adjust for the non-linear portion of the quadratic terms and interaction term; they will appear as higher noise.

In the LI scenario, sequential matching dominates competitors in the classic and the exact test because the competitors do not use the covariate information. In the linear assumption, sequential matching performs similarly in power but has a lower efficiency across all n 's. This loss is due to a lower effective sample size when using matched pairs. The gap decreases as n increases because there are benefits to matching even when employing regression adjustment. As Greevy et al. (2004) explain, better balance reduces collinearity resulting in a smaller standard error for the estimate. Balance is improved over competitors that do not allocate based on the

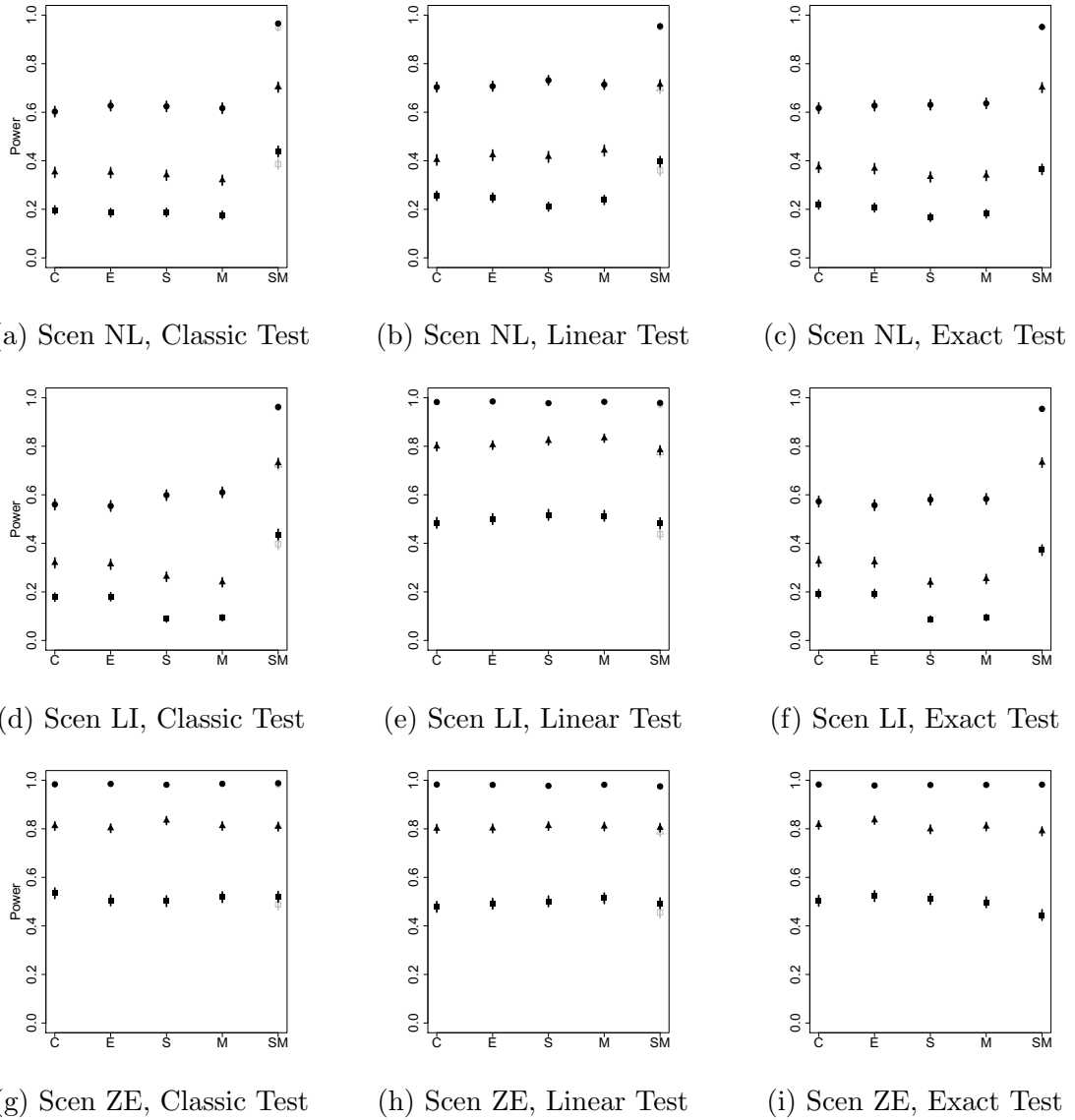


Figure 5.1: Power at $\alpha = 0.05$ illustrated for matching parameter $\lambda = 10\%$ for the three scenarios by the three testing procedures, all five allocation methods (C: Complete Randomization, E: Efron's BCD, S: Stratification, M: Minimization and SM: our sequential matching algorithm) and sample sizes (squares illustrates results for $n = 50$, triangles for $n = 100$ and circles for $n = 200$). Plotted points represent the sample proportion of null hypothesis rejections in 2,000 simulations and segments represent 95% confidence intervals. The grey points and segments plot the T approximation to the finite distribution of Equations 5.4 and 5.6.

Sample Relative Efficiency Over Competitors								
Allocation			Scenario NL		Scenario LI		Scenario ZE	
n	Method	Balance	Classic	Linear	Classic	Linear	Classic	Linear
50	CR	0.812	2.233	1.645	2.543	0.814	0.896	0.779
	Efron's BCD	0.810	2.134	1.744	2.358	0.805	0.854	0.825
	Stratification	0.422	1.664	1.282	1.125	0.780	0.798	0.717
	Minimization	0.397	1.750	1.478	1.163	0.778	0.842	0.718
	Seq. Matching	0.506	—	—	—	—	—	—
100	CR	0.804	2.610	1.762	2.637	0.898	0.944	0.885
	Efron's BCD	0.802	2.500	1.688	2.709	0.855	0.967	0.885
	Stratification	0.392	1.785	1.401	1.224	0.820	0.928	0.879
	Minimization	0.376	2.975	1.532	1.132	0.842	0.932	0.879
	Seq. Matching	0.430	—	—	—	—	—	—
200	CR	0.801	3.139	2.011	3.183	0.969	0.991	0.851
	Efron's BCD	0.796	2.695	2.034	3.187	0.979	1.005	0.866
	Stratification	0.378	2.084	1.586	1.400	0.977	1.002	0.917
	Minimization	0.370	2.241	1.757	1.289	0.938	0.964	0.883
	Seq. Matching	0.361	—	—	—	—	—	—

Table 5.2: Balance results and relative sample efficiency results (the ratio of the two unbiased sample variances) of sequential matching ($\lambda = 0.10$, Z approximation) versus competitors by scenario and testing procedure. Balance results are averages over all scenarios and model assumptions (60,000 simulations for SM and 36,000 for other allocations). Efficiencies in **gray** indicate our algorithm performed worse than a competitor and efficiencies in **bold** indicate our algorithm performed better (as measured by an F-test with 1% significance level unadjusted for multiple comparisons). Exact tests are not shown because they do not admit a standard error calculation.

covariates and this better balance implies higher power and efficiency. Parenthetically, we note that as n increases, it appears as if balance is approaching levels observed in both stratification and minimization. This is expected and is an added bonus of our procedure.

In the ZE scenario, our approach is most severely impacted by lower effective sample size. However, power is not as low as expected. Efficiency seems to be lost for all simulated n 's but most significantly when $n = 50$ as expected.

All in all, sequential matching shines in the case of non-linear covariate models which is the most realistic case in practice. If the covariate model is truly linear, sequential matching does worse when OLS is employed but our relative inefficiency is only observed for small sample sizes. In the case when covariates do not matter at all, we begin to perform about equally with competitors when $n \geq 100$. This is an important result in practice because investigators sometimes choose useless covariates which do not affect the outcome measure.

A possible criticism of the high power achieved is we assume our n was large enough for the estimators in Equations 5.4 and 5.6 to converge. Via unshown simulations, we have reason to believe the true density to be similar to a T with degrees of freedom $\max(m - 1, \min(n_{R,T}, n_{R,C}))$ where $n_{R,T}$ and $n_{R,C}$ are the number of treatments and controls left in the reservoir respectively. We illustrate power results for this putative T distribution in gray in Figure A.17c. To further investigate the convergence, we also simulated the size of the tests in Table 5.3. For the classic estimator, at $n = 50$, the size of the Z approximation test is generally larger than the Type I error rate of 5% even with the T procedure (in most cases). At $n = 100$ and 200, the Z approximation is appropriately sized for most scenario-model situations, and the T procedure is appropriately sized for almost all cases. Other anomalies observed in this table are discussed in Section 9.6.

n	Allocation	Scenario NL			Scenario LI			Scenario ZE		
	Method	Classic	Linear	Exact	Classic	Linear	Exact	Classic	Linear	Exact
50	CR	0.047	0.047	0.052	0.046	0.046	0.056	0.053	0.050	0.048
	Efron's BCD	0.050	0.046	0.040*	0.054	0.055	0.046	0.053	0.051	0.046
	Stratification	0.016**	0.031**	0.018**	0.006**	0.051	0.006**	0.050	0.049	0.046
	Minimization	0.021**	0.041	0.024	0.004**	0.046	0.005**	0.046	0.055	0.050
	Seq. Match (Z)	0.078**	0.072**		0.077**	0.089**		0.087**	0.093**	
	Seq. Match (T)	0.056	0.052	0.052	0.061*	0.077*	0.046	0.076**	0.073**	0.044
100	CR	0.050	0.047	0.050	0.042	0.053	0.054	0.054	0.049	0.048
	Efron's BCD	0.053	0.050	0.048	0.054	0.046	0.046	0.052	0.054	0.050
	Stratification	0.034*	0.034*	0.023**	0.004**	0.046	0.004**	0.048	0.050	0.046
	Minimization	0.032**	0.040*	0.043	0.004**	0.048	0.061*	0.046	0.046	0.050
	Seq. Match (Z)	0.054	0.054		0.054	0.062*		0.063*	0.061**	
	Seq. Match (T)	0.047	0.052	0.051	0.050	0.063*	0.050	0.060*	0.070**	0.048
200	CR	0.052	0.043	0.051	0.051	0.053	0.050	0.043	0.053	0.060*
	Efron's BCD	0.052	0.050	0.050	0.044	0.052	0.048	0.052	0.051	0.050
	Stratification	0.018**	0.027**	0.014**	0.004**	0.054	0.003**	0.051	0.050	0.048
	Minimization	0.018**	0.032*	0.024**	0.004**	0.053	0.001**	0.054	0.046	0.052
	Seq. Match (Z)	0.053	0.056		0.052	0.049		0.054	0.062*	
	Seq. Match (T)	0.049	0.060	0.049	0.054	0.058	0.048	0.050	0.058	0.054

Table 5.3: Simulated size of tests for all scenarios, competitors, and all tests at $\lambda = 10\%$. Numbers followed by a ** indicate they are different from the purported 5% size at a Bonferroni-corrected significance level (162 comparisons). Numbers followed by a * indicate they are different from the purported 5% size without Bonferroni correction.

5.4 Demonstration Using Real Data

5.4.1 Behavioral Experiment

Kapelner and Chandler (2010) (Chapter 3 of this document) ran experiments using the Amazon Mechanical Turk platform, a global outsourcing website for small one-off tasks that can be completed anonymously on the Internet. They focused on measuring subjects' stated preference for a beer price when the beer came from different

purchasing locations (an online replication of Thaler, 1985’s demonstration of the “framing effect,” a cognitive bias). The treatment involved subtle text manipulations: the same beer came from either a *fancy resort* or a *run-down grocery store*. In their control wing ($n = 168$), no tricks were employed to ensure the subjects were paying attention to the text. Thus, in this wing, the subtle text manipulation did not seem to affect the subjects’ stated beer prices. The effect may have been real, but the data was either too noisy or there was insufficient sample size to find it. We demonstrate here that if our sequential matching procedure was employed, the effect estimator would have been more efficient.

For matching, we first used most of the covariates found in the original dataset: age, gender, level of earnings, number of weekly hours spent doing one-off tasks, level of multitasking when performing tasks, stated motivation level, passing the “instructional manipulation check” (Oppenheimer et al., 2009) and a survey gauging the subject’s “need for cognition” (Cacioppo and Petty, 1982).

We note that R^2 under OLS was about 18.7%. We then run two off-the-shelf machine learning algorithms that are designed to find interactions and non-linearities in the response function. The in-sample pseudo- R^2 using Chipman et al. (2010)’s Bayesian Additive Regressive Trees (BART) was 42.4% and Breiman (2001b)’s Random Forests (RF) was 70.4%. Although this is not a formal test, it is pretty compelling evidence that the covariates do not combine strictly linearly to inform beer price. Thus, as demonstrated in figure A.17a and column 4 of table 5.2, our method should be more powerful and more efficient than using previous dynamic allocation strategies with a classic estimator. The results for 200 simulations at $\lambda = 0.10$ are shown in table 5.4a. Many of the covariates are binary. Thus, the variance-covariance matrix was *not* invertible in line 5 of algorithm 1 for many of the early iterations, so we used the Moore-Penrose generalized inverse instead.

We now match on four selected covariates that come out most significant in an

	purported n	average actual n	average efficiency	approx. sample size reduction
(a)	50	37.8	1.84	45.7%
	100	71.9	1.23	16.9%
	168 (all)	116.1	1.06	5.4%
(b)	50	34.9	2.01	50.1%
	100	67.8	1.60	37.3%
	168 (all)	112.1	1.57	36.3%

Table 5.4: Results for 200 simulations of the sequential matching procedure over many values of n and $\lambda = 0.10$. (a) all covariates matched on (b) four cherry-picked covariates are matched on (OLS has an $R^2 = 20.8\%$, BART, 32.1% and RF, 26.5%). Note that many of the covariates were binary, and thus the variance-covariance matrix was *not* invertible in line 5 of algorithm 1 for many of the early iterations of these simulations, so we used the Moore-Penrose generalized inverse.

OLS regression on the full dataset: age, level of earnings, level of multitasking when performing tasks and one question from the survey gauging the subject’s “need for cognition.” The results are found in table 5.4b. Note that the efficiencies are higher and do not drop off as quickly when n increases. Thus, matching on *relevant* covariates yields a performance enhancement in our procedure.

5.4.2 Clinical Trial

We now examine sequential clinical trial data from Foster et al. (2010), a twelve-week, multicenter, double-blind, placebo-controlled clinical trial studying whether amitriptyline, an anti-depressant drug, can effectively treat painful bladder syndrome. The study measured many outcomes, including change in pain after 12 weeks (difference in Likert scale scores). The confidence interval for the ATE between pill and

placebo for this outcome measure was $[-1.00, 0.30]$ with ap value of 0.29 (Table 2, row 1, page 1856). Such results indicate the effect may have been real but there wasn't enough power to detect it due to low sample size or a high error variance.

For matching, we first use most of the covariates found in the original dataset: age, gender, race (White / Hispanic), level of education, level of employment, living with a partner, presence of sexually transmitted diseases and urinary tract infection, as well as baseline measures of pain, urination frequency and urgency, quality of life, anxiety and depression as well as syndrome symptom levels.

R^2 was about 25.2% in an OLS regression. We then run two off-the-shelf machine learning algorithms that are designed to find interactions and non-linearities in the response function. The in-sample pseudo- R^2 using Chipman et al. (2010)'s Bayesian Additive Regressive Trees (BART) was 42.6% and Breiman (2001b)'s Random Forests (RF) was 82.4%. Although this is not a formal test, it is pretty compelling evidence that the covariates do not combine strictly linearly to inform change in pain after 12 weeks. Thus, as demonstrated in Figure A.17a and column 4 of Table 5.2, our method should be more powerful and more efficient than using previous dynamic allocation strategies with a classic estimator.

We simulate the subjects being dynamically allocated using the sequential matching procedure by first assuming the entering subjects do not exhibit any time trend; this will allow us to permute their order. During the iterative procedure, all subjects assigned to the reservoir keep whichever assignment they had during the experiment. During matching, if the subject happened to have been assigned the treatment which the sequential matching procedure allocated, they are kept in the subject pool; if not, they are discarded (this is illustrated in Figure 5.2). Thus, during our simulations, we result in a *subset* of the data we began with. Note that we only show results for the classic estimator versus the modified estimator in Equation 5.4, not the OLS modified estimator whose results we suspect to be similar.

```

nsim: 1 .....o.xx....o....xxx.oxxx.o.xxooxooxx.xo. (37)
nsim: 2 .....xxx..x.xxoo.xxx.o.o.oo..ox.xxxxx.o (35)
nsim: 3 .....x.x..oo.xxxo.x.xxxxxxxxxo.x.ox.oox.o (34)
nsim: 4 .....o...o...ox.o.o.xo.ox.o....xxooox.oox.xo (42)
nsim: 5 .....x.xoo.o.o..xxoxx.o.x.x.oo.o.xoxx.. (39)
nsim: 6 .....x..xx.x..xx.oo.xoxoxxx.ooxo.o.xxxo (35)
nsim: 7 .....x.o..oo..oxx.x..o.oxo.xxox.o.xoxxxoxo (37)
nsim: 8 .....xx.oxxxo..x.xxx.x..oxoo.o.xx.x..xox.o (34)
nsim: 9 .....o...xxoo.xo.ooo.....o.oxxox.oxx..o.x (42)

```

Figure 5.2: Running an $n = 50$ subset of historical data through the sequential procedure. The dots represent a subject being placed into the reservoir. The “o” signifies that the subject was matched and that their treatment allocation was *opposite* of their matching partner. The “x” signifies that the subject was matched but their treatment allocation was the *same* as their matching partner, resulting in the subject being discarded. The number in parenthesis at the end of the line is the sample size retained of the purported 50.

The results for 200 simulations are shown in Table 5.5a. We also match on the top four covariates which are the most significant in an OLS of the full dataset: living with a partner, baseline pain, frequency of pain and syndrome symptom levels. The results are found in Table 5.5b. Note that these efficiencies are higher when compared to matching on all covariates and they do not suffer the steep drop off as n increases.

5.5 Discussion

Estimation in sequential experiments, of which many are clinical trials, can have higher power and efficiency if the covariate information is leveraged. We present a dynamic allocation of treatment and control that matches subjects on-the-fly via a novel algorithm and present modified estimators of classic approaches: average difference, linear regression, and permutation testing. We simulate under different scenarios and illustrate higher power in scenarios where competing methods cannot

	purported n	average actual n	average efficiency	approx. sample size reduction
(a)	50	38.9	1.30	23.0%
	100	75.2	1.10	9.2%
	150	111.3	1.05	4.9%
	224 (all)	165.5	1.07	6.7%
(b)	50	38.0	1.27	21.2%
	100	72.9	1.23	18.8%
	150	108.6	1.15	13.2%
	224 (all)	160.3	1.13	11.3%

Table 5.5: Results for 200 simulations of the sequential matching procedure over many values of n and $\lambda = 0.10$. (a) all covariates matched on (b) four cherry-picked covariates are matched on (OLS has an $R^2 = 18.9\%$, BART, 35.1% and RF, 45.0%). Note that many of the covariates were binary, and thus the variance-covariance matrix was *not* invertible in line 5 of algorithm 1 for many of the early iterations of these simulations, so we used the Moore-Penrose generalized inverse.

make proper use of covariate information. We underperform only in the case of low sample size when the covariate model is linear or non-existent. In simulations with real clinical data, we find the efficiency of our method over complete randomization increases as the covariates become more important. This is most likely due to the fact that real-world response functions, such as this one in the clinical setting, combine covariates non-linearly, and this is when our procedure is most advantageous.

We note that “analysis assumptions may be compromised due to the ‘pseudo’-random allocation” (Scott et al., 2002) and would like to address this criticism which can be made about our procedure. Note that in Table 5.3, the size of the tests under stratification and minimization are less than 5%. One should not use the classic

estimator in these cases because one implicitly tried to balance on the covariates, but then did not include the covariates in the model (as seen by the poor sizes using the classic model in scenario NL using the classic model in scenario LI in Table 5.3). Simon (1979) and Senn (2000) have very good discussions about this issue and recommend using regression adjustment (as seen in the competitors in scenario LI, linear response model in the Table 5.3 for $n > 50$). As for exact testing under dynamic allocation, Kalish and Begg (1987) and many others warn that permutation distributions in stratification and minimization are incorrect unless the investigator permuted the treatment allocation according to how the allocation was determined by the covariates. This is not straightforward in practice for stratification and even less straightforward for minimization.

In contrast, the sizes of the test for our approach seem to be correct in Table 5.3, especially when making use of the low-sample T approximation. Usage of our classic estimator seems ill-advised for the same reasons that the classic estimator is not recommended under stratification and minimization. However, using linear regression on the covariates is also ill-advised when the model is non-linear or otherwise does not satisfy the OLS model assumptions (Freedman, 2008) although Rubin (1979) finds that covariance adjustment of matched pair differences is robust to model misspecification. Thus, since our permutation test performs well in the non-linear case, properly sized and can avoid many of the above issues, we recommend the permutation test in practice. We permuted according to the structure of our dynamic matching allocation, thus our permutation tests are valid.

5.5.1 Further Developments

We view this contribution as a step forward in covariate-adaptive randomization in sequential experiments but it is far from complete. We list extensions below which

we believe are in difficulty order.

Sequential Analysis Although we assume fixed n in our construction, it is relatively straightforward to adapt to a fully or group sequential design whose methods can be found in Jennison and Turnbull (2000). It would be hard to tabulate values when our estimator has unknown convergence properties, thus it would probably have to be done by waiting until the estimator most likely converged, and then using standard sequential analysis software.

Multi-armed designs This paper is devoted to studies where there are only two arms. If the study is multi-armed, the approach can be modified. Upon matching, the newly matched subject could be randomly assigned to a random treatment (an arm other than the treatment of the reservoir subject) or perhaps, alternatively, assigned to one of the other arms to best achieve balance. Analogous estimators can then be created for contrasts of interest.

Non-continuous outcomes We believe with adjustment of the estimators, our method can apply beyond continuous responses to binary, ordinal, or count responses.

Imbalance protection Complete randomization when allocating subjects in the reservoir may result in treatment imbalances in the reservoir after the experiment is completed. In clinical trials, block permutations are popularly employed to avoid imbalance. In our case, blocks for reservoir allocation would not work since, upon matching, the block sequences would be broken. A solution would be to use a biased coin allocation scheme (Efron, 1971) or an urn model. This would involve modification of the exact test of Section 5.2.3.3 to reflect the restricted randomization scheme.

Multi-centered designs Clinical trial recruitment is frequently spread across multiple centers and center-center variation is expected. Our procedure can be modified

in two ways to be of use in multiple centers. First, if large enrollment is expected in each center, the procedure can run independently in each center and resulting estimates from each center can be combined upon study completion. Second, if small enrollment is expected in each center, the subjects can be matched across centers and the least squares estimator of section 5.2.3.2 can be used to adjust for center-center variation.

Better matching distance We feel that the most significant improvement would be better matching. Mahalanobis distance is logical, but prone to strange behavior with departures from the normality assumption. Another natural extension is to bootstrap the distribution of the nominal metric in Equation 5.2 as to not rely on probabilities from the scaled F distribution. Also, practitioners may want to weight the variables in the matching as well as force some variables to always match (see Rosenbaum, 2010, chapter 8).

On-the-fly variable selection Of course, our procedure suffers from the central issue of all matching: selecting the variables to match on. A poor choice makes a big difference as evidenced by the simulations on clinical trial data (Table 5.5a vs. 5.5b). There may be a way to *iteratively* match and differentially weight by covariates that are found to be important, so the set of perceived important covariates is updated during the sequential experiment.

Reservoir optimization We have begun to consider how large n_R grows asymptotically as a function of n_0 and λ (Section 5.2.4). To understand the optimal tradeoff of n_0 and λ to maximize estimator efficiency as functions of n , p , the variance-covariance matrix of the covariates, and how strong the signal of f is to the noise will involve a lot of theory. Also, perhaps a variable rule for λ would be effective: if the sample size is large, the algorithm can afford to be conservative about the matches during

the beginning of the experiment, but then become less conservative as time passes. Waiting until n_0 is almost half n can be another strategy.

Supplementary Materials

Original data and source code for reproduction can be found at github.com/kapelner/sequential_matching_simulations_2.

Acknowledgements

We wish to thank Larry Brown, Dean Foster, John Horton, Stephen Kapelner, Katherine Propert, Paul Rosenbaum, Andrea Troxel, and the reviewers for helpful suggestions and Hugh MacMullan for help with grid computing. Adam Kapelner also acknowledges the National Science Foundation for his graduate research fellowship.

Estimating the Number of Objects in Images*

Abstract

We develop a method to estimate the count of objects or features in an image such as tallying the number of birds in a photograph or the the number of cells in a microscopic image. Our approach has two novel steps. We first develop software that records the labelings of many naive workers via Amazon’s Mechanical Turk. We then view each worker’s training as a “capture” in a set of capture-recapture experiments. We use statistical learning to eliminate falsely-trained objects and then take a Bayesian approach and use a Gibbs sampler to estimate the true number of objects.

6.1 Introduction

Counting and locating specific features in an image has many applications such as cell counting, harvest estimation, and crowd counting. Standard methods in counting (how many words did Shakespeare known? (Efron and Thisted, 1976) and animal

*Joint work with Susan Holmes

abundance) use models where the frequencies of counts form sufficient statistics. However, with large amounts of noisy data, such models have to be refined. In this paper, we develop a two-stage solution to the counting and localization of objects in images. Some images feature objects that are easy to locate, thus the count is obvious; other images are difficult featuring ambiguous and obfuscated objects, this is where our solution can be applied.

In section 6.2 we give an introduction to object identification in the field of computer vision and our previous research that served as the inspiration for this project. In section 6.3 we talk about the engineering effort of our new tool, `DistributeEyes`, that uses crowd labor to label images. The use of this new tool has many applications, but we focus here on improving the estimate of the number of objects in an image. In section 6.4 we talk about statistical methods that use data from `DistributeEyes` to estimate the number of objects in images using Bayesian capture-recapture methods. In section 6.5 we run an experiment with a wide variety of images, and test our count estimation method on the experimental data. We are also able to present a quality control analysis and present observations about spatial and temporal patterns in workers' training points. We conclude and talk about future directions in the final section 6.6.

6.2 Crowdsourcing Object Identification in Images

Many automated computer vision algorithms are available and have been tailored for specific types of images and object identification questions. Cell recognition techniques applied to microscopy images are transforming the field of pathology (Kapeller et al., 2007b), thermal/aerial photographs can be analyzed for car detection and counts used in traffic analysis.

Projects like LabelMe (Russell et al., 2008) provide annotation for objects in an

image but have concentrated on images where there are less than 15 objects, most are of different types. In contrast, we are interested in counting and locating many objects of similar types; in general, there will be more than 50 objects of each type which need to be counted. In our implementation, we ask the participants to identify the object by clicking on one central pixel, instead of providing a full outline of the object, or textual annotation.

Our previous research (Holmes et al., 2009; Kapelner et al., 2007b) yielded an interactive statistical learning software named “GemIdent” (Kapelner et al., 2007a) which was designed to locate cells in histological samples. The program can be adapted to any object recognition task on images with few colors and relative conformity in the shape and size of the sought objects. GemIdent was instrumental in the localization and identification of T-cells and cancer cells in immunohistologically stained microscope images enabling an architectural comparative analysis in tumor-draining lymph nodes from breast cancer patients and healthy lymph nodes (Setiadi et al., 2010). The software package has been open-sourced and available for download since May, 2007. We have received much feedback about its shortcomings. Overwhelmingly, users feel that the training step for the different object instances is too time-consuming. We were inspired to address this weakness by introducing a new possibility for incorporating training data — from a *crowd* of inexperienced trainers, located around the world, who assist in identifying objects of interest. “Crowd-sourcing’ is the act of outsourcing tasks, traditionally performed by an employee or contractor, to an undefined, large group of people or community (a crowd), through an open call”.²

Why charge anonymous laborers with such as task? General automatic object identification using statistical learning is difficult in cases where the contexts are heterogeneous but the objects are not. Manually marking the objects would yield the

²<http://en.wikipedia.org/wiki/Crowdsourcing>

most accurate solution, but would be too time-consuming in real applications such as cell counting where there could be millions of cells in a single histological sample.

Our solution is to provide *training data* through a set of points collected via Internet-based crowdsourcing. We facilitate the distribution of the task of marking the objects. We thus extended our image segmentation software and names it “DistributeEyes”³. It is an HTML application that plugs into a crowdsourcing platform, currently Amazon’s Mechanical Turk⁴ and communicates with a Gemident-enhanced client. We give the details of the integration of these platforms in the following section.

6.3 Engineering

The data acquisition solution has several different components which communicate with each other for an integrated solution (see figure 6.1).

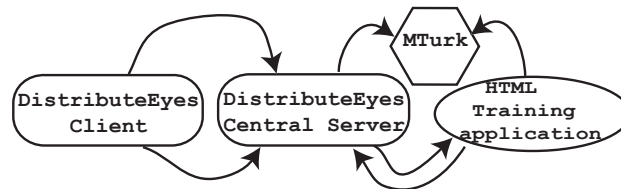


Figure 6.1: The DistributeEyes Components and their interactions

³<http://www.distributeeyes.com>

⁴<http://www.mturk.com>

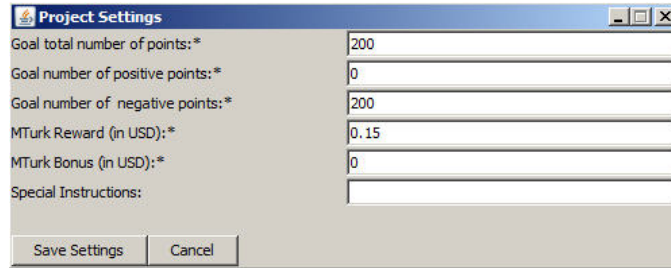


Figure 6.2: DistributeEyes Project Settings Dialog Window

6.3.1 The DistributeEyes client

A user of the client (the “user”) creates a new project and identifies the different objects of interest types (for historical reasons, we call these “phenotypes”) and then provides a few examples of each. The user then initializes the project on the central `DistributeEyes` server (the “server”). Then, images that we would like to extract training data from (“images”) are added to the project.

The “Distribute Images” panel (not shown) shows each image as a row in a table and allows users to select which images they want to distribute for online marking. The marking of one image is done as a “Human Intelligence Task” (“HIT”) on Amazon’s MTurk by a worker and is dubbed a “submission.”

A project settings panel allows customization of the following parameters: minimum number of total points, minimum number of positive points (objects) sought, minimum number of negative points (non-objects) sought, monetary reward for training, bonus reward for training, time-to-expire for the HIT, and special instructions.

After images are distributed, the client polls the server regularly to see if the HITs have been completed. Once a submission is detected, the client provides two means of assessing the quality of the work. The first is a window that shows the local area around each marked point which allows for quick deletions of simple mistakes (see figure 6.3). The second is the phenotype training window which displays the original

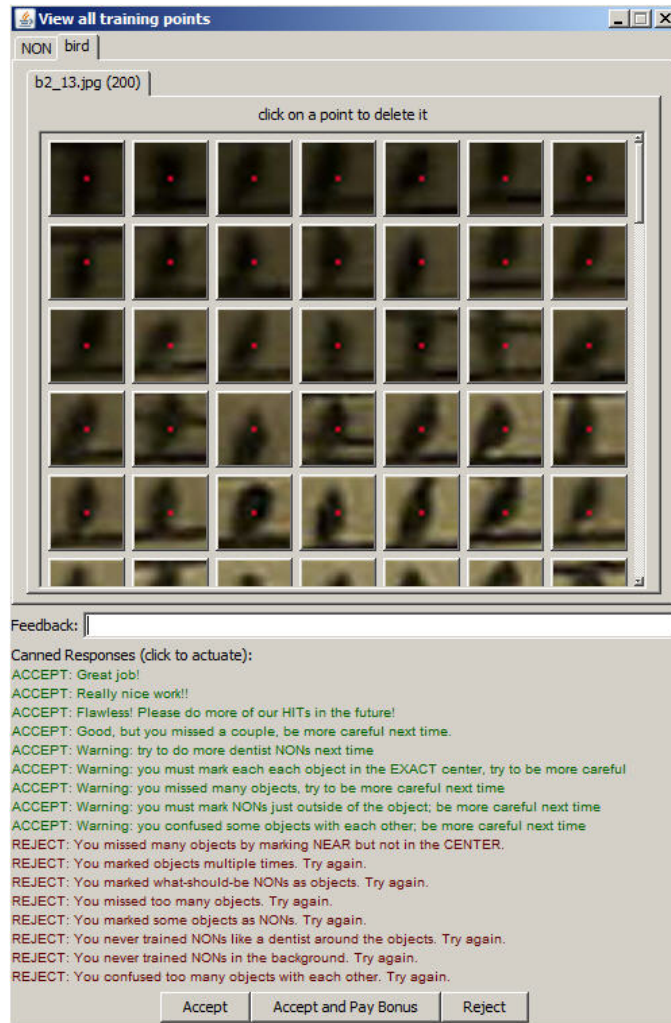


Figure 6.3: Worker submission check window

image with an overlay of the training points (not shown, see Holmes et al., 2009 Figure 8, p9).

Usually within 5-10 seconds the user can recognize if the worker did a satisfactory job and accept or reject accordingly. A rejection means zero compensation for the worker. The client also provides common canned feedback responses such as “Great job!” or “You missed too many objects. Try again.” If the worker went above and beyond, a bonus can be rewarded. As an additional tool, the worker’s historical record (acceptances, rejections, bonuses, and feedback) are displayed.

It is this speed-up of the training process which makes *DistributeEyes* a powerful

research tool with enormous potential.

6.3.2 The Central DistributeEyes Server

The function of the server is to be an intermediary for passing data between the client and MTurk's workers, to render the HTML training application, and to serve the public homepage URL.⁵ User accounts are created via the public portal and after email verification, the user can download the DistributeEyes client.

6.3.3 The HTML Training Application and the Raw Data

Each image can be rendered inside of a training application for marking by a worker. The training application comes complete with each example image of each phenotype, a magnified view of the image where the user marks / deletes points, magnification adjustments, a thumbnail, a scaled-to-fit view, and a help link.

Before the worker is allowed to use the application, he must “qualify” by watching a five-minute tutorial video which teaches training concepts and techniques, pass a six-question quiz (see figure 6.4), and agree that DistributeEyes can monitor his usage. Workers must pass the quiz in order to be able to work on a HIT. The worker is forced to watch the entire video before having the opportunity to answer the questions. Workers that pass the qualification quiz are recorded on the central server and no longer need to re-qualify to work on future HITs.

Then, the worker moves to the training application (see figure 6.5). The worker marks points via point-and-click until all goals set by the user are met and all special instructions are fulfilled. The “submit training” button becomes enabled, and the worker can upload his points to the central server. When the worker submits, he is led through a series of dialogs inquiring if he has made the common mistakes that

⁵The server was written in Ruby on Rails using MySQL for the database.

Qualification Test

- 1 - When you are training for positive points, how many times do you click on the phenotype example?
 - Once
 - Twice
 - As many dots as you can fit inside the object
- 2 - When you are required to train NONs, your hit will be rejected if you do not...
 - train around the borders of positive objects like a dentist
 - use the right mouse button
 - use the thumbnail viewer to look at the whole image
- 3 - When you make a mistake by clicking something incorrectly, you should...

Figure 6.4: The training video and example quiz questions

lead to rejections, and only then is he allowed to submit.

While the worker trains, we collect information about his actions e.g. when and where he clicked, when he switched phenotypes, when he used the magnification tools, and we log these actions every ten seconds.

To be clear, we refer equivalently to a “training” or a “labeling” as a worker’s submitted points for one image which are coordinates (x_{i_p}, y_{i_p}) where $i_p \in \{1, \dots, n_p\}$, $p \in \{1, \dots, P\}$ where P denotes the total number of phenotypes (usually one) and n_p denotes the total number of points per phenotype that the worker labeled.

In our setup using `DistributeEyes`, acquiring many workers’ labelings is quick

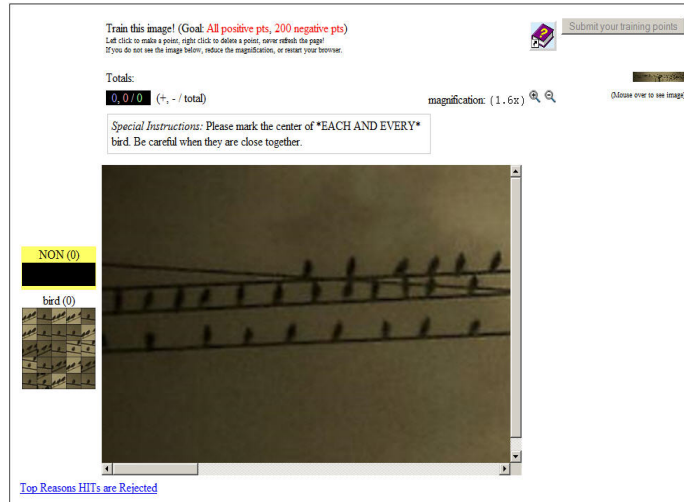


Figure 6.5: The HTML training application embedded into MTurk. The worker is training the project “birds” as part of our experiment (see section 6.5).

and inexpensive.

6.3.4 Interfacing with MTurk

Amazon Mechanical Turk (MTurk) is a marketplace that coordinates the use of human intelligence to perform tasks which computers are unable to do. This labor paradigm has been coined “crowdsourcing” which Amazon whimsically refers to as “artificial, artificial intelligence.”

Since it is the largest such crowdsourcing marketplace on the Internet, and it provides a convenient Applications Programming Interface, it was chosen as the venue to host `DistributeEyes` tasks. To create an MTurk HIT, the HTML Training Application is rendered inside an `IFrame` which is wrapped inside of MTurk’s worker interface. From the perspective of the worker, everything occurs within MTurk; he is completely unaware of `DistributeEyes`.

After submission, the client will detect the completion within 10 minutes. Acceptances or rejections are relayed from the client to MTurk via the central server using

the MTurk API. The server records each acceptance or rejection. The workers are warned upon rejections. If a worker receives many rejections, he may be banned from working on DistributeEyes HITs.

6.4 Statistical Model and Implementation

In many instances pooling together work done by many naive workers can have greater quality than an expert. Suppose we have one type of object (one phenotype) we would like to count (*e.g.* faces in a photograph of a crowd or cell nuclei in a microscopic field-of-view) and suppose we have W workers who submit training coordinates for where the objects are located. Our goal is to combine all W trainings together to create \hat{N} , an estimate for N , the total number of objects.⁶

To first approximation, we can view each of the W trainings as “captures” of different objects and all trainings together as a *capture-recapture* experiment in a closed population: that of the actual objects.

6.4.1 Capture-Recapture Models

Capture-Recapture experiments trace their history back to estimating the number of animals in a certain locality. The number of animals caught in a particular sample n may give little information about N , the total population size, considered to be static. Hence, many captures are done where all animals in a capture are marked via a unique identifier and released. In subsequent capture trials, the number of animals previously marked is noted. It is for this reason that this procedure is sometimes known as a discrete-time *capture-mark-recapture* experiment. For a more thorough

⁶We also use their most likely locations in practical applications, but this is not addressed in this investigation.

introduction we refer the reader to the beginning of Otis et al. (1978) and Robert (2007, Chapter 5.1 and 5.2).

Methods for combining the data from the different captures to provide an estimate for N can vary based on the assumptions one makes about the heterogeneity of the animals and the way they are captured. Otis et al. (1978) notes three types of heterogeneity:

- t The probability of capturing an animal varies with the time of the capture, that is it varies trial-trial within the capture-recapture experiment.
- b The probability of capture of an animal is dependent on its behavioral response to a previous capture.
- h There is heterogeneity in each of the animal's capture probability.

Each type of heterogeneity can be combined together for a total of eight possible models.

6.4.2 Object Identification as a Capture-Recapture Model

The objects in the image of interest can be thought of analogously in the Capture-Recapture paradigm as animals. The number of objects are now the closed population size. Each of the W worker labelings are now viewed as capture trials. If two different workers mark the same object, it is counted as a "recapture".

In an arbitrary image, some objects are more difficult to detect than others, therefore heterogeneity of type h must be accounted for. Since the objects are pixels unchanged in an image, there cannot be any heterogeneity due to previous captures, hence heterogeneity of type b does not need to be considered. Lastly, we assume that all workers have the same ability. This may not be fully justified, but we do this to

avoid extra parameterization and to keep our model simple. There is also substantial evidence that all workers can be considered the same in ability for a variety of crowdsourced tasks (for example, see Kapelner et al., 2012 which is also Chapter 7 of this document). Therefore, capture probabilities are assumed not vary with the trial number, hence heterogeneity of type t does not apply. In conclusion, we believe that the M_h model (notation being standard in the literature) with W trials is a reasonable model that we can apply to our data in order to estimate the number of objects in images.

There is one conceptual issue that does not perfectly adhere to the M_h model. Workers are charged with the task of clicking on the objects. Sometimes, they make mistakes by clicking on a point that is not actually an object, *i.e.* they label a “false positive” (these can be likened to traps capturing “ghost” animals), an intractable problem. We first present the model without this complication below and then we address this problem in section 6.5.5 using statistical learning.

6.4.3 Estimating N in an M_h model

Let X_i denote the number of times object i has been labeled when all W labelings of one image of interest are aggregated. Let p_i denote the probability that object i is detected by any worker (we assumed worker ability is homogenous). It is clear that the number of objects found is binomial. We further assume that the p_i 's are drawn from a beta distribution, an assumption that is justified in section 6.5.3.

$$X_1 | p_1 \sim \text{Binomial}(W, p_1), \dots, X_N | p_N \sim \text{Binomial}(W, p_N)$$

$$p_1, \dots, p_N \stackrel{iid}{\sim} \text{Beta}(\alpha, S\alpha)$$

Following the work of George and Robert (1992) and Gilks et al. (1996), we choose to estimate N via Bayesian estimation. The Bayesian paradigm fits our purposes because

we have built up sufficient prior information that can be incorporated into a realistic coherent model. We now outline the prior setup and Gibbs sampling procedure used in the hierarchical model found in Yip et al. (2005).

Let r denote the number of unique objects detected when all W labelings are pooled. Denote $M = N - r$ as the number of objects left undetected. Let f_1 denote the number of objects detected only by one worker, let f_2 denote the number of objects detected by only two workers, etc, and let f_W denote the number of objects detected by all workers. We now use the following flexible conjugate priors:

$$\alpha \sim \text{Gamma}(k, \lambda), \quad S \sim \text{Gamma}(\ell, \tau), \quad M \sim \text{Geometric}(\rho)$$

Integrating out p_1, \dots, p_N , we arrive at the following Gibbs conditional sampling scheme:^{7,8}

$$\begin{aligned} M \mid \alpha, S, r, f_1, \dots, f_W, W, \rho &\sim \text{NegBin} \left(r + 1, 1 - \rho \frac{\Gamma(\alpha + S\alpha) \Gamma(S\alpha + W)}{\Gamma(S\alpha) \Gamma(\alpha + S\alpha + W)} \right) \\ \alpha \mid M, S, r, f_1, \dots, f_W, W, k, \lambda &\propto \left(\frac{\Gamma(\alpha + S\alpha)}{\Gamma(S\alpha) \Gamma(\alpha + S\alpha + W)} \right)^{r+M} \frac{\Gamma(S\alpha + W)^M}{\Gamma(\alpha)^r} \\ &\quad \times \alpha^{k-1} e^{-\lambda\alpha} \prod_{j=1}^W (\Gamma(\alpha + j) \Gamma(S\alpha + W - j))^{f_j} \\ S \mid \alpha, M, r, f_1, \dots, f_W, W, \ell, \lambda &\propto \left(\frac{\Gamma(\alpha + S\alpha)}{\Gamma(S\alpha + \alpha) \Gamma(\alpha + S\alpha + W)} \right)^{r+M} \Gamma(S\alpha + W)^M \\ &\quad \times S^{\ell-1} e^{-\lambda S} \prod_{j=1}^W \Gamma(S\alpha + W - j)^{f_j} \end{aligned}$$

N can now be sampled by using the posterior samples of M , the number of undetected

⁷Note that the distributions for α and S are non-standard, and further, can only be obtained up to a normalizing constant, therefore they are sampled via a Metropolis-Hastings acceptance or rejection.

⁸Note that the parameterization of the beta distribution in the model definition as $[\alpha, S\alpha]$ is done for computational reasons to reduce the correlation of the beta's two parameters during Gibbs sampling which is a common trick.

objects, and adding back the number of detected objects, r .

6.4.4 The Gibbs Sampler Implementation

For the prior distributions of α and S we pick hyperparameters to have a target mean probability of detection of 66% with large spread which is reasonable for both easy and difficult images. We aim for a target mean $\alpha = 4$ and mean $S\alpha = 2$ but with considerable variance in order to be relatively uninformative. Therefore we choose hyperparameters:

$$\alpha \sim \text{Gamma}(k = 2, \lambda = 0.5) \quad \text{and} \quad S \sim \text{Gamma}(\ell = 1, \tau = 0.25)$$

As for the prior distribution on M , the number of undetected objects, note that the geometric distribution has one parameter and thereby is difficult to control both the mean and variance. We choose rather arbitrarily a mean $\mu_M = 5$ ($\sigma_M^2 = 20$) which coerces ρ to be 0.8. Note that the results in this study were not sensitive to these choices.

For the Gibbs sampling, we do 100,000 sample iterations and we discard the first 10,000 as a burn-in and we thin by using only 10% of the remaining samples evenly spaced (after which autocorrelation is negligible for all three parameters), leaving us with empirical posterior distributions of 9,000 samples.

We use logs of the densities and proportional densities to avoid computational overflow when large values of the gamma function are required. For the Metropolis-Hastings steps, we follow Yip et al. (2005) and generate α_1 from the current α_0 using the lognormal distribution with scale parameter 1 (which we found moves us around the parameter space sufficiently quickly).

$$\alpha_1 \sim \log\mathcal{N}(\ln(\alpha_0), 1) \quad \text{and} \quad S_1 \sim \log\mathcal{N}(\ln(S_0), 1)$$

We then accept α_1 if:

$$\ln(U(0, 1)) < \ln\left(\frac{\mathbb{P}(\alpha_1 | M, S, r, f_1, \dots, f_W, W, k, \lambda)}{\mathbb{P}(\alpha_0 | M, S, r, f_1, \dots, f_W, W, k, \lambda)} \cdot \frac{\alpha_0}{\alpha_1}\right)$$

where $U(0, 1)$ denotes a draw from a standard uniform distribution. S_1 is sampled analogously.

6.5 Experiment and Results

To test the `DistributeEyes` system and our count estimation techniques, we allowed up to 50 workers on MTurk to label the locations of objects in twelve diverse images. We then used the labelings to estimate the total number of objects, N . Since we also have the baseline truth, we were able to build a reasonable model to account for the workers' false positive mistakes. Finally, we were able to assess trends of worker trainings over time, bias, difficulty, spatial effects, and others.

6.5.1 Experimental Setup

We designed three types of object-recognition tasks increasing in complexity: (a) find all the objects of one phenotype (b) find all the objects of one phenotype and mark points that do not serve as examples of the object, the "NONs" (c) find all the objects of multiple phenotypes and mark points that do not serve as examples of any of the objects, the "NONs." Complexity class (c) features projects with multiple phenotypes. For the analyses and estimation found in the following sections, these multiple phenotypes were collapsed into one. In each complexity class we tested four images for a total of 12 "projects". We tried to represent a diverse range of phenotype types (see figure 6.6) across the projects.

For each of the projects, we created 50 identical HITs which was enough to inves-

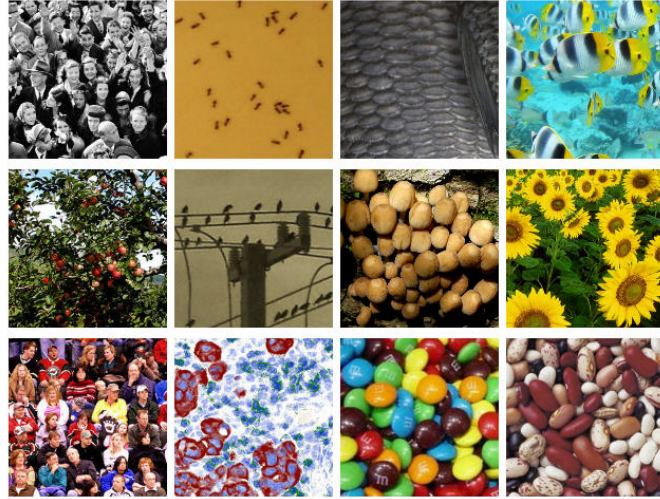


Figure 6.6: Cropped selections from the project images. From left to right then top to bottom: Faces, Ants, Scales, Fish, Tree, Birds, Mushrooms, Sunflowers, Fans, Cells, Candies, Beans

tigate the distribution of quality in worker submissions. Workers were barred from marking the same image more than once. Our reward for a completed HIT was constant at \$0.15 USD. HITs were given a 60-minute time limit. The HITs were opened to the MTurk community at large without geographical restrictions. Worker submissions were checked within 10-20 seconds using the tools described in Section X.

If the submission was especially poor, or if a server error caused an incorrect submission, it was rejected and the data was not analyzed. This is the reason that W varies between 23 and 47 in the next section.

6.5.2 Data and Basic Results

Worker training data is output from the server into csv files that are then analyzed using R (R Development Core Team, 2005). A truth set was manually created for each of the twelve images. The data for each worker's training was then compared to

the truth set. If the worker’s object coordinate was within an average object radius from the truth point, it was marked as correct i.e. a “true positive”, otherwise, it was tallied as a “false positive.” Therefore, for each project we have a true positive binary matrix with workers on the rows and the true positives as the columns where a 1 represents worker j found object i . We also have a false positive binary matrix structured analogously. Overall, rejected trainings aside, the workers performed well. We measured “accuracy” for training a certain phenotype by the $F1$ -score — a harmonic average between *precision* (the percentage of worker’s training points that successfully corresponded to an object) and *recall* (the percentage of total objects found by the worker). The results for the positive phenotypes are summed up in table 6.1. The NONs (the negative phenotype) results were not analyzed, but anecdotally appear to be accurate as well. Broadly speaking, MTurk laborers seem to take this image-labeling task seriously.

6.5.3 Object Detection Probabilities are Distributed as a Mixture of Betas

Before we begin estimating N , we justify the assumption that the probability of detecting an object is beta-distributed.

We did not find compelling evidence that the probability of detecting the objects are beta-distributed for all 12 projects. However, we found strong evidence that the percent of objects detected by each worker (the true-positive rates) are beta distributed.

Because we know the truth, we know if each worker detected each true positive. We compute the proportion of true positives detected for each worker by taking the number found over the total number of true points. We then estimate the α, β parameters of a putative beta distribution using maximum likelihood, and then test the appropriateness of the sample being drawn from the Beta(α, β) using an Anderson-

Project	$F1$ -score ($\bar{x} \pm SD$)	W
Faces	0.96 ± 0.03	27
Ants	0.96 ± 0.02	47
Scales	0.89 ± 0.05	28
Fish	0.90 ± 0.04	42
Tree	0.73 ± 0.16	27
Birds	0.97 ± 0.02	39
Mushrooms	0.92 ± 0.06	36
Sunflowers	0.88 ± 0.07	35
Fans	0.96 ± 0.04	33
Cells	0.79 ± 0.06	23
Candies	0.92 ± 0.04	28
Beans	0.90 ± 0.06	29

Table 6.1: Averages and standard deviations of $F1$ scores for worker labelings. For the four images that had multiple phenotypes, we report the average $F1$ score considering all points over all phenotypes. 50 image tasks were released for each project. The W is the number of trainings that were accepted and therefore the number of trainings that were analyzed.

Darling test (table 6.2). For each project, the test passed at $\alpha = 0.05$ implying that the true positives are beta-distributed (for an example, see figure 6.7).

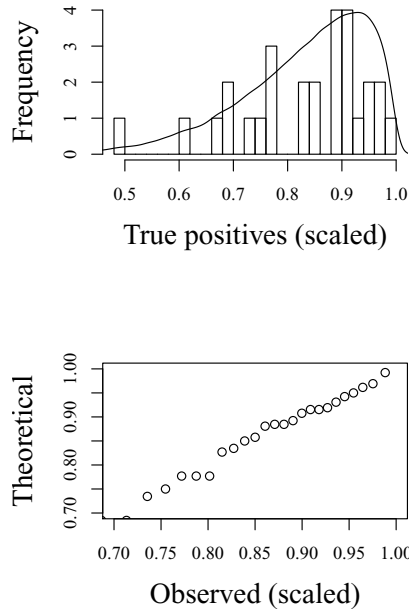


Figure 6.7: Top: Putative beta fit (found via maximum likelihood estimation) overlaid atop the true positive rate for workers in the “Fish” project. Bottom: a Q-Q plot for the putative beta distribution.

Since the workers and the true positives are exchangeable, we have a jointly exchangeable binary array. We already know the probability model among workers is plausibly beta in all projects. Hence, by the Aldous-Hoover theorem, we know that the probability model across the true positives (the detection probabilities p_1, \dots, p_N) must be a *mixture* of betas.

6.5.4 Estimating N using the Experimental Data when the Truth is Known

We now apply the methods from section 6.4.3 to the data collected from the 12 images.

Project	Anderson-Darling p -value
Faces	0.94
Ants	0.74
Scales	0.99
Fish	0.50
Tree	0.76
Birds	0.31
Mushrooms	0.13
Sunflowers	0.53
Fans	0.89
Cells	0.90
Candies	0.87
Beans	0.97
Apples	0.95

Table 6.2: The results of the Anderson-Darling test for the number of true positives of being Beta-distributed for all 12 projects.

In the situation where the “truth is known”, we know the locations of the TP’s and can generate the statistics r and f_1, \dots, f_W using the binary true positive matrix (see section 6.5.2 for information about this construction). Note that we *ignore* all false positive points. We then proceed to use the Gibbs sampler explained in section 6.4.4.

The results for all projects are presented in figure 6.8. All of the 12 projects feature 95% posterior predictive intervals that capture the true number of TP’s. These results inform us that the model works very well when we have oracular knowledge.

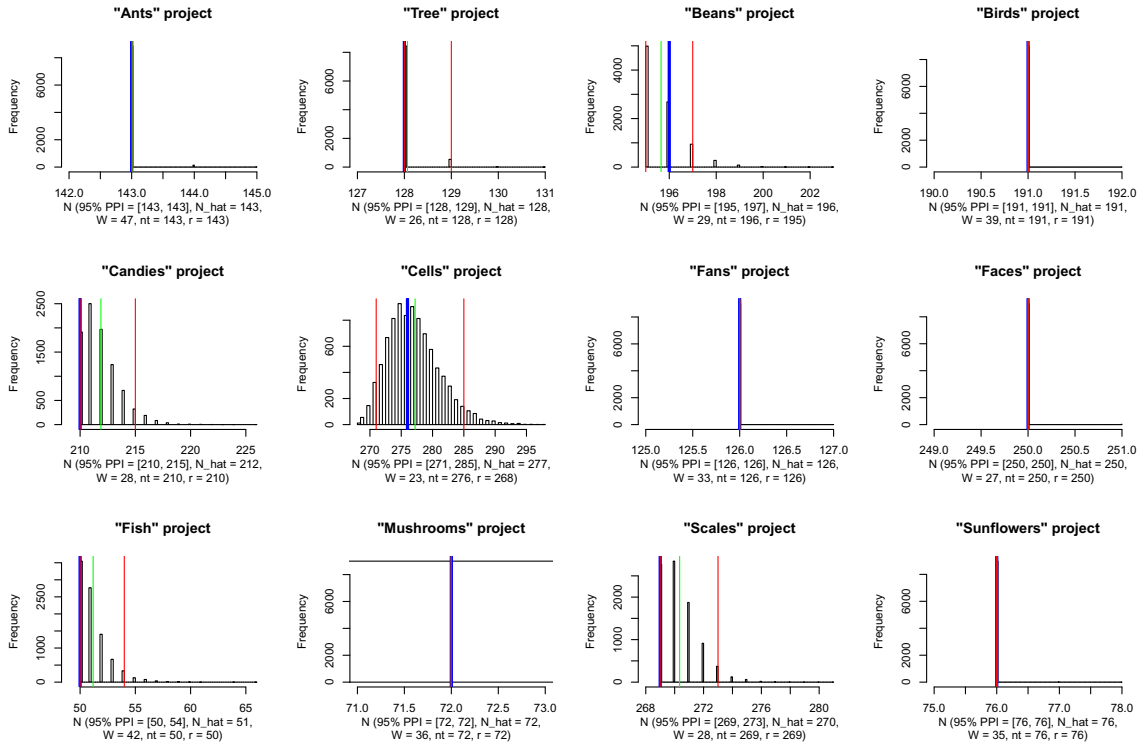


Figure 6.8: Histograms of the empirical posterior of N for each project considering the truth is known. The blue line is the number of true positives (labeled as “ nt ” below each histogram); the red lines are the bounds of a 95% posterior predictive interval (labeled as “95% PPI = $[a, b]$ ”); the green line is the posterior mean (labeled as “ $N.hat$ ”). Note that 6 of the 12 projects were extremely easy, featuring large and unambiguous objects and hence there was no variation in the estimates.

6.5.5 Estimating N when the Truth is Unknown

In the realistic situation where “truth is unknown”, we only have the workers’ labelings. In order to estimate N without changing the model nor the Bayesian framework, we must estimate the sufficient statistics r and f_1, \dots, f_W . Recall in the M_h model that these sufficient statistics are computed *only* from true captures. Therefore, we proceed in three steps: (1) we group the worker training points into common objects via a clustering algorithm, (2) we estimate the likelihood that each cluster is a true object, and finally (3) we update the Gibbs sampler to account for this uncertainty.

6.5.5.1 Grouping Training Points into Common Objects

We start with an unsophisticated clustering algorithm that resolves the workers’ trainings to common objects. We construct the clusters through agglomeration. Starting with an initial point, we then agglomerate the closest point (measured by Euclidean distance) within a user-defined radius, then agglomerate the next closest point to the center of the initial two points, etc. We allow 10% of the workers to label a point twice as a duplicate.⁹

The total number of clusters represents the total number of objects found (both true objects and false objects) and the number of points in each cluster represents the number of workers that found that object.

Now we must estimate the probability that each cluster represents a false positive. We take a statistical learning approach.

6.5.5.2 Classifying Clusters as True Captures or False Captures

The goal is to build a machine that accepts a set of points (a cluster of workers’ coordinates), and outputs the probability that the cluster is a true positive (or a false

⁹Our results were not sensitive to this choice.

positive).

We build a design matrix X where each row represents a cluster and the column entries are features of the cluster. The raw features are found in table 6.3.¹⁰ Note that the radius of the objects of interest are *not* used during the computations of any of the features.

Number	Name	Description
1	n	The number of points in the cluster
2	ε	The eccentricity of the minimum volume enclosing ellipse (MVVE)
3	foc	The focus of the MVVE
4	al	The length of the semi-major axis of the MVVE
5	bl	The length of the semi-minor axis of the MVVE
6	SSE	The sum squared Euclidean distances of each point from the cluster centroid
7	A	The area of the α -hull of the points (where $\alpha \gg 0$)
8	x_c, y_c	The centroid (coordinates) of the cluster
9	d_1, d_2, d_3, d_4	The centroid - centroid Euclidean distance of the first closest cluster, second closest, etc

Table 6.3: The raw features used to discriminate a cluster between being a true positive and a false positive. We also add many derived features as A/n , a_ℓ/n , etc. The total number of features used was 14.

We use the statistical learning algorithm Classification and Regression Trees (CART,

¹⁰The minimum volume enclosing ellipse (MVVE) was computed using an R-implementation of the Khachiyan Algorithm (Gacs and Lovász, 1981). The α -hull was computed using the `alphahull` package.

Breiman et al., 1984)¹¹. We built the classifier in-sample and achieved an average error rate of 3.1%. Examining the tree split rules, we found that the features that were the most important in the discrimination were n , A , and SSE .

6.5.5.3 Updating the Capture-Recapture Gibbs Sampler

We use CART to output the probability of being a true positive.¹² We use these estimates to augment the M , α , and S Gibbs sampler (from section 6.4.4) by adding a sampling step for which simulates whether or not each cluster is a true positive based on the probability estimate from CART.¹³

$$\begin{aligned}
 M \mid \alpha, S, r, f_1, \dots, f_W, W, \rho &\sim \text{same as before} \\
 \alpha \mid M, S, r, f_1, \dots, f_W, W, k, \lambda &\propto \text{same as before} \\
 S \mid \alpha, M, r, f_1, \dots, f_W, W, \ell, \lambda &\propto \text{same as before} \\
 \text{cluster}_1 &\sim \text{Bern}(\mathbb{P}(\text{cluster 1 is a TP according to CART})) \\
 \text{cluster}_2 &\sim \text{Bern}(\mathbb{P}(\text{cluster 2 is a TP according to CART})) \\
 &\vdots \\
 f_1 &= \text{The number of clusters with 1 training point} \\
 &\vdots \\
 f_W &= \text{The number of clusters with } W \text{ training points} \\
 r &= \sum_{i=1}^W f_i
 \end{aligned}$$

¹¹The implementation used was the R package `rpart` with standard options

¹²This is estimated by examining the empirical distribution in each of the leaves of the pruned classification tree.

¹³If the outputted probability is zero or one, we draw from a $\text{Beta}(1, r_o + 1)$ or $\text{Beta}(r_o + 1, 1)$ respectively where r_o is the number of *total* clusters found across all workers for the given project.

To compute f_1 , we count the the number of clusters that have one training point; to compute f_2 , we count the number of clusters that have two training points, etc. r represents the total number of clusters found, which is equivalent to summing all the f_i 's.

Figure 6.9 shows our promising results. The scales and beans project features very ambiguous or hidden objects. We would argue that this type of object estimation is not possible with our system due to the difficulty in distinguishing FP's from TP's.

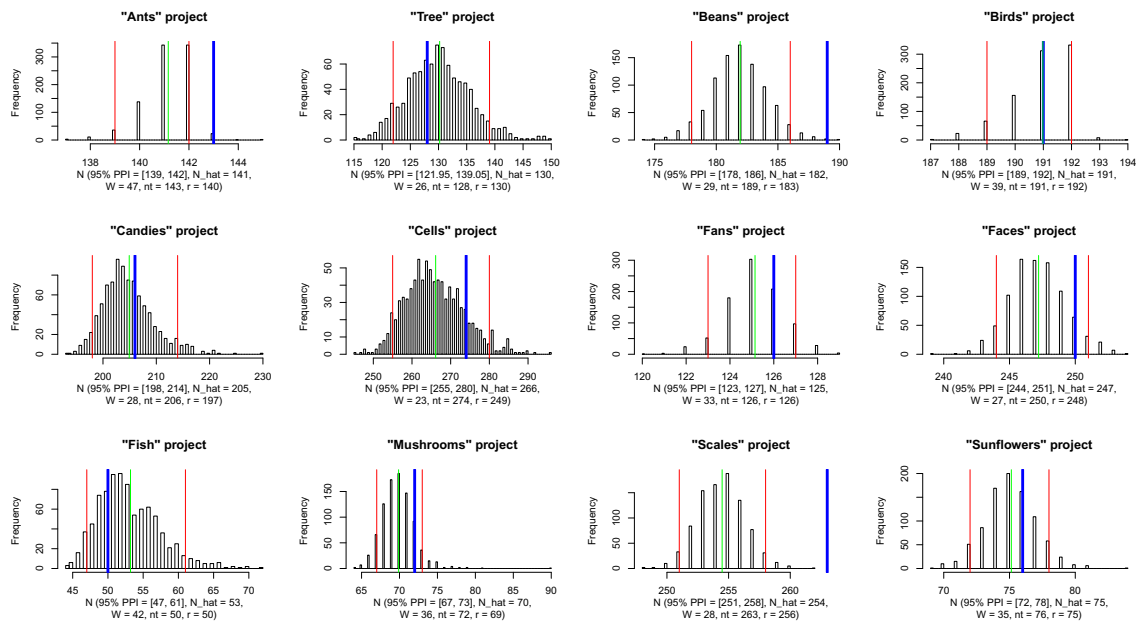


Figure 6.9: Histograms of the empirical posterior of N for each project when truth is unknown. f_1, \dots, f_w, r are sampled via a three-step procedure. Layout follows figure 6.8.

6.5.6 Spatial and Temporal Findings

In addition to estimating counts using the Bayesian capture-recapture model, we carried out other investigations on the worker labelings. Below are the most important of our findings.

6.5.6.1 Jumping Around Improves Accuracy

We find weak evidence that workers who “jump around” the image during training tend to have more accurate trainings. This is expected, since workers that go back and check their work do indeed jump around (see figure 6.10a). The effect was weak; on average over all projects, a one standard deviation increase in training distance yields a two-tenths standard deviation increase in $F1$ score ($p = 0.0001$).

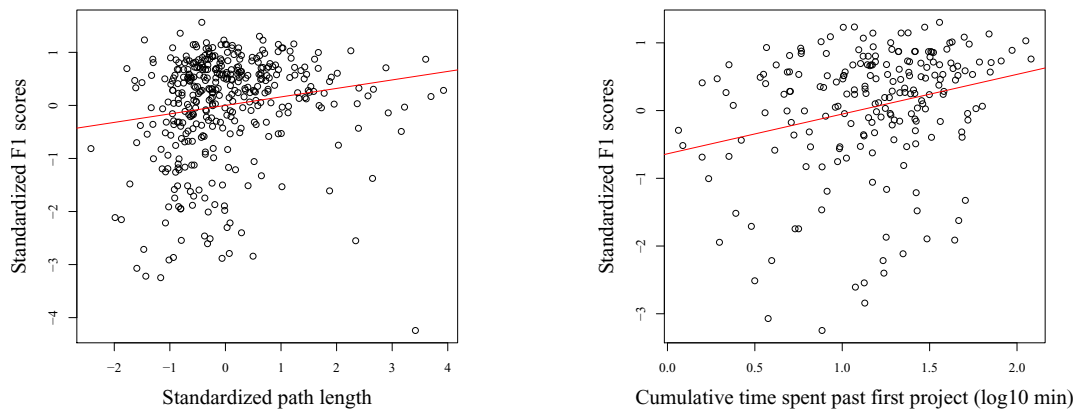


Figure 6.10: Left: Standardized accuracy regressed on standardized training distance (across all projects). Right: Standardized accuracy regressed on log 10 cumulative time training in minutes beyond first image (across all projects and workers who completed more than three images)

6.5.6.2 Workers Improve with Time

We found evidence that workers overall become more accurate in training with more experience. Investigating workers that finished 3 or more projects (see figure 6.10b), we found on average that beyond the first training, for each order of magnitude increase in cumulative time training, the $F1$ score increases by half of a standard deviation ($p < 0.001$). Also expected, we find a similar effect when regressing standardized

$F1$ scores on total number of projects instead of cumulative time training.

6.5.6.3 Longer Training does not Yield More Accurate Results

Surprisingly, spending more time training *within a single image* does not correspond to more accurate results. We regressed $F1$ scores on time training for all twelve projects individually. Not one of the twelve projects features a significant correlation at a Bonferroni-corrected level.

6.5.6.4 Some Images are More Difficult than Others

We ordered each project by average $F1$ score and then compared the $F1$ score distributions using Wilcoxon-rank tests. Anecdotally, the *easy* projects seem to have easily identifiable objects all in the spatial foreground; the *intermediate* projects either had ambiguous objects in either the foreground background; the *difficult* projects had very ambiguous objects in the foreground and/or background. The best results from DistributeEyes seem to come from images that have features that are both easily-identifiable and in a single spatial perspective.

We find that our capture-recapture expresses image difficulty via the width of the posterior distribution that it outputted. These widths more or less correspond to the results we present here.

6.5.6.5 Trainings Can be Biased

We investigated whether workers on average tend to train above, below, or exactly the number of true objects. We find in projects where objects are ambiguous (i.e. the *intermediate* difficulty level described in section 6.5.6.4), the training is *unbiased*. We hypothesize that lazy trainers and over-ambitious trainers balance each other out. We also find that on projects where objects are well-defined (i.e. the *easy* difficulty level),

the trainings are generally *biased downwards*. Since there was a clear maximum, the assiduous trainers could never balance the lazy trainers. We find that our capture-recapture approach is excellent for mitigating these effects.

6.6 Conclusion

As large amounts of data are accumulated over many images and image-contexts, we can build data-based prior distributions for recognizing situations where many small objects could be located and counted. Eventually, we plan to construct a generalized “salience” classifier. A lofty goal would be to perform *unsupervised* machine learning to find objects of interest given an arbitrary context.

Another potential application is that `DistributeEyes`-trained data can lead to more efficient object tracking in videos and 3D extensions. One of the applications of the training/segmentation/counting paradigm presented here is to the tracking of animals leading to behavioral analysis (the “Ants” project). An example of passing from the static image analysis to video tracking has been developed recently (Pinter-Wollman et al., 2011).

`DistributeEyes` provides a platform for combining human generated data of high quality with computer based segmentation and feature detection output. This bridges the gap between areas when only human based systems are available (cell counting) and those where automated systems provide adequate data (image segmentation).

We can now use this on future studies both to determine the number of workers needed to acquire a reasonable estimate of a crowd count and for the collection of training data of a necessary accuracy level to be fed into automated learning algorithms.

Acknowledgements

Thanks to Larry Brown for helpful discussions.

Collecting labels for Word Sense Disambiguation*

Abstract

We use crowdsourcing to disambiguate 1000 words from among coarse-grained senses, the most extensive investigation to date. Ten unique participants disambiguate each example, and, using regression, we find surprising features which drive differential WSD accuracy: (a) the number of rephrasings within a sense definition is associated with higher accuracy; (b) as word frequency increases, accuracy decreases even if the number of senses is kept constant; and (c) spending more time is associated with a decrease in accuracy. We also observe that all participants are about equal in ability, practice (without feedback) does not seem to lead to improvement, and that having many participants label the same example provides a partial substitute for more expensive annotation.

*Joint work with Krishna Kaliannan, Lyle Ungar and Dean Foster

7.1 Introduction

Word sense disambiguation (WSD) is the process of identifying the meaning, or “sense,” of a word in a written context (Ide and Véronis, 1998). In his comprehensive survey, Navigli (2009) considers WSD an AI-complete problem — a task which is at least as hard as the most difficult problems in artificial intelligence. Why is WSD difficult and what is driving its difficulty? This study examines human WSD performance and tries to identify drivers of accuracy. We hope that our findings can be incorporated into future WSD systems.

To examine human WSD performance, we tap pools of anonymous untrained human labor; this is known as “crowdsourcing.” A thriving pool of crowdsourced labor is Amazon’s Mechanical Turk (MTurk), an Internet-based microtask marketplace where the workers (called “Turkers”) do simple, one-off tasks (called “human intelligence tasks” or “HITs”), for small payments. See Snow et al. (2008); Callison-Burch (2010); and Akkaya et al. (2010) for MTurk’s use in NLP, and Chandler and Kapelner (2013) which is also Chapter 2 of this document and Mason and Suri (2011) for further reading on MTurk as a research platform.

We performed the first extensive look at coarse-grained WSD on MTurk. We studied a large and variegated set of words: 1,000 contextual examples of 89 distinct words annotated by 10 unique Turkers each. In the closest related literature, Snow et al. (2008) found high Turker annotation accuracy but only annotated a single word, while Passonneau et al. (2011) focused on only a few words and annotated fine-grained senses. The extensive size of our study lends itself to the discovery of new factors affecting annotator accuracy.

Our contribution is three-fold. First, we use regression to identify a variety of factors that drive accuracy such as (a) the number of rephrasings within a sense definition is associated with higher accuracy; (b) as word frequency increases, accuracy

decreases even if the number of senses is kept constant; and (c) time-spent on an annotation is associated with lower accuracy.

Second, we echo previous findings, mostly from non-WSD experiments, demonstrating that Turkers are respectably accurate (Snow et al., 2008), they’re approximately equal in ability (Parent, 2010; Passonneau et al., 2011), spam is virtually non-existent (Akkaya et al., 2010), responses from multiple Turkers can be pooled to achieve high quality results (Snow et al., 2008; Akkaya et al., 2010), and that workers do not improve with experience (Akkaya et al., 2010). Third, we present a system of crowdsourcing WSD boasting a throughput of about 5,000 disambiguations per day at \$0.011 per annotation.

7.2 Methods and data collection

We selected a subset of the OntoNotes data (Hovy et al., 2006), the SemEval-2007 coarse-grained English Lexical Sample WSD task training data (Pradhan et al., 2007). The coarse-grained senses in OntoNotes address a concern that nuanced differences in sense inventories drives disagreement among annotators (Brown et al., 2010). We picked 1,000 contextual examples at random from the full set of 22,281.² Our sample is detailed in table 1. It consisted of 590 nouns and 410 verb examples that had between 2-15 senses each (nouns: 5.7 ± 3.0 senses, verbs: 4.7 ± 3.3 senses). For each snippet, ten annotations were completed by ten *unique* Turkers.

²We later disqualified 9 of the 1,000 because they had words with only one sense.

7.2.1 The WSD HIT

We designed a simple WSD task that was rendered inside an MTurk HIT.³ The Turker read one example in context with the target word emboldened, and then picked the best choice from among a set of coarse-grained senses (see Figure 7.1). We gave a text box for soliciting optional feedback and there was a submit button below. We term a completed HIT an “annotation.”

We employed anti-spam and survey bias minimizing techniques to obtain better data. We faded in each word in the context and the sense choices one-by-one at 300 words/min.⁴ Additionally, we randomized the display order of the sense choices. This reduces “first response alternative bias” as explained in Krosnick (1991), but may decrease accuracy when compared to displaying the senses in descending frequency order as observed by Fellbaum et al. (1997). We also limited participation to US Turkers to encourage fluency in English.

Upon completion, the Turker was given an option to do another of our WSD tasks (this is the MTurk default). A Turker was not limited in the number of annotations they could do.⁵ The entire study took 51 hours and cost \$110. The code, raw data, and analysis scripts are available under GPL2 at github.com/kapelner/wordturk_pilot.

³The HIT was entitled “Tell us the best meaning of a word... do many and earn a lot! Really Easy!”, the wage was \$0.01, the time limit for each task was seven minutes, and the HITs expired after one hour. We posted batches of 750 new HITs to MTurk hourly upon expiration of the previous batch. Thus, the task was found readily on the homepage which drove the rapid completion.

⁴As Kapelner and Chandler (2010) found (also Chapter 3 of this document), this accomplishes three things: (1) Turkers who plan on cheating will be more likely to leave our task, (2) Turkers will spend more time on the task and, most importantly, (3) Turkers will more carefully read and concentrate on the meaning of the text.

⁵The actual upper limit was all 1,000 examples but in practice, not one Turker came close to completing all of them. The most productive Turker completed 405 annotations while the median completed was 4.

target word	# inst	# senses	target word	# inst	# senses	target word	# inst	# senses
affect-v	1	3	end-v	8	4	policy-n	10	3
allow-v	8	2	enjoy-v	3	2	position-n	13	7
announce-v	4	3	examine-v	3	3	power-n	12	4
approve-v	3	2	exchange-n	17	6	president-n	34	3
area-n	15	5	exist-v	1	2	produce-v	6	3
ask-v	16	6	explain-v	2	2	promise-v	3	2
attempt-v	2	2	express-v	3	3	propose-v	1	3
authority-n	3	6	feel-v	24	3	prove-v	2	6
avoid-v	2	2	find-v	7	6	raise-v	6	9
base-n	5	12	fix-v	2	6	rate-n	49	2
begin-v	7	4	future-n	16	4	recall-v	2	4
believe-v	9	2	go-v	12	14	receive-v	4	2
bill-n	18	9	hold-v	5	10	regard-v	1	3
build-v	1	4	hour-n	9	4	remember-v	8	6
buy-v	7	7	job-n	6	10	remove-v	4	2
capital-n	15	5	join-v	2	4	report-v	7	4
care-v	6	3	keep-v	11	8	rush-v	1	4
carrier-n	4	13	kill-v	5	9	say-v	104	5
chance-n	5	4	lead-v	9	7	see-v	9	10
claim-v	4	4	maintain-v	3	4	set-v	10	12
come-v	7	11	management-n	13	2	share-n	103	3
complain-v	2	2	move-n	19	4	source-n	10	6
complete-v	1	3	need-v	11	2	space-n	2	8
condition-n	7	4	network-n	9	4	start-v	8	7
defense-n	8	8	occur-v	2	4	state-n	33	4
development-n	8	3	order-n	11	9	system-n	15	7
disclose-v	5	2	part-n	14	7	turn-v	19	15
do-v	4	6	people-n	38	6	value-n	16	5
drug-n	7	3	plant-n	12	3	work-v	4	9
effect-n	7	5	point-n	27	14			

Table 7.1: The 89 words of the sample of 1000 OntoNotes snippets used in this study.

“# inst” is the number of instances in the 1,000 with the corresponding target word.

“# senses” is the number of sense choices provided by OntoNotes.

Word Meaning Task

Read the following snippet which will fade in slowly:

Apple shares fell 75 cents in over-the-counter trading to close at \$48 a share. Fiscal fourth-quarter sales grew about 18% to \$1.38 billion from \$1.17 billion a year earlier. Without the Adobe gain, Apple's full-year operating profit edged up 1.5% to \$406 million, or \$3.16 a **share**, from \$400.3 million, or \$3.08 a share. Including the Adobe gain, full-year net was \$454 million, or \$3.53 a share. Sales for the year rose nearly 30% to \$5.28 billion from \$4.07 billion a year earlier.

Please pick the meaning of the word **share** which best fits the context of the paragraph above:

- capital stock in a corporation
- a tool for tilling soil
- a portion or percentage of a whole

Submit my definition of "share" (and whatever optional feedback I left below)

My feedback:

We also welcome and give bonuses to feedback, comments, and bug reports:

Figure 7.1: An example of the WSD task that appears inside an MTurk HIT. This was displayed piecewise as each word in the example ("snippet") and senses faded-in slowly.

7.3 Results and data analysis

We were interested in investigating (1) which features in the target word, the context, and sense definition text affect Turker accuracy, (2) which characteristics in the Turker's engagement of the task affect accuracy, (3) heterogeneity in worker performance, and (4) the combination of Turker responses to boost accuracy.

We recruited 595 Turkers to work on our tasks, yielding an average accuracy of 73.4%.

We measured inter-tagger agreement (ITA) using the alpha-reliability coefficient (Krippendorff, 1970) to be 0.66 (0.70 for nouns and 0.60 for verbs) which comports with Chklovski and Mihalcea (2003)'s *Open Mind Word Expert* system.

However, OntoNotes was specially designed by Hovy et al. (2006) to have 90% ITA by experts. Our measure is significantly less. Untrained Turkers should not be expected to be experts.

7.3.1 Performance and language characteristics

What makes WSD difficult for untrained Turkers? Are there too many senses to choose from? Is the example difficult to read? With 10,000 instances from 600 workers, we can attempt to answer these questions.

We first construct the features of interest:

- target word part-of-speech (*target word is noun?*)
- target word length in characters (*# chars in target word*)
- target word frequency (*log target word frequency*)
log of frequency in the contemporary corpus of American English (Davies, 2008).
- number of senses to choose from (*# senses to disambiguate*)
- number of characters in the correct sense definition (*# chars in definition*)
- number of rephrasings in definition text (*# rephrasings in definition*)

For example, the word “allot” has a sense with definition text “let, make possible, give permission” which would be counted as three rephrasings.

- number of characters in context (*# chars in context*)

We add a fixed intercept for each Turker to account for correlation among tasks completed by the same worker. The result of an ordinary least squares (OLS) regression of correct (as binary) on the variables above is presented in table 2.⁶

⁶We also ran a variety of fixed and random effects linear and logit models, all of which gave

	estimate	t
<i>target word is noun?</i>	8.4%	7.5 ***
<i># chars in target word</i>	-1.0%	3.6 ***
<i>log target word frequency</i>	-3.7%	7.6 ***
<i># senses to disambiguate</i>	-2.9%	19.8 ***
<i># chars in definition</i>	-0.063%	2.6**
<i># rephrasings in definition</i>	3.4%	5.4 ***
<i># chars in context</i>	-0.0062%	2.6 **

Table 7.2: OLS regression of instance correctness on features of the target word, context, and senses. Fixed effects for each of the 595 Turkers are not shown. ** indicates significance at the $< .01$ level, *** indicates significance at the < 0.001 level.

We found that, controlling for all other variables, nouns have 8% higher disambiguation accuracy. This difference between noun and verb accuracy is also reflected in automatic system performance on the SemEval-2007 task Pradhan et al. (2007), and often attributed to the idea that nouns “commonly denote concrete, imagible referents” (Fellbaum et al., 1997). For each extra sense, accuracy suffers 3% which also is expected since the Turkers have more choices. We show accuracy by number of senses and part of speech in figure 7.2. We also found the longer the target word, the more difficult the task, reflecting the fact that longer words are often more complex. Similarly, the longer the context or length of definitions decreased accuracy but the effect was quite small.

Surprisingly, with each extra rephrasing of the definition of the correct sense there is a gain of 3.5%. This suggests untrained annotators benefit from receiving a variety of sense descriptions, or that more rephrasings suggests a more coarse-grained sense

the same significance results. We chose to present the OLS output because of its familiarity and interpretability.

which is easier for annotators to understand.

Finally, as the word becomes more common in the English language (controlling for all other variables, including length of word and number of senses) accuracy still suffers. Possibly the more prevalent the word in our language, the more likely it will have senses that overlap conceptually.

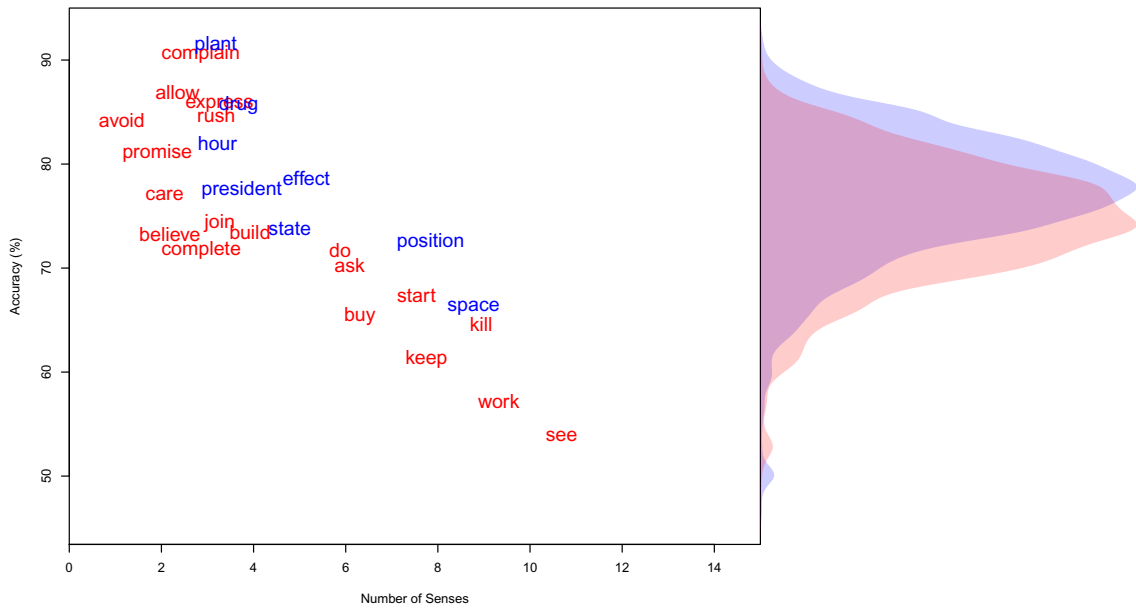


Figure 7.2: Predicted accuracy vs. number of senses for a sample of the words in our study. Nouns are blue; verbs are red. The densities are smoothed histograms of the noun and verb predicted accuracies. Note that the word display is jittered; there are at least two senses for each word.

7.3.2 Performance and Turker characteristics

Are there any characteristics about the Turker’s engagement with our task that impacts accuracy? We create the following features: time spent on task, the number of words in their optional feedback message, and the number of annotations that worker completed prior to the response being examined. To control for the difficulty of each

task, we added 1,000 fixed intercepts — one for each unique task; and to control for correlation among the workers, we added a fixed intercept for each worker. An ordinary least squares regression of the WSD task being correct (as binary) on the variables above⁶ was run. We found that for each additional second spent on the task, accuracy drops by 0.06% ($p < 0.001$). We found that, contrary to Kapelner and Chandler (2010) (also Chapter 3 of this document), leaving comments does not correspond to higher accuracy, and, in agreement with Akkaya et al. (2010), the number of tasks completed prior does not impact accuracy. This may imply that a learning effect does not exist; practice (without feedback) does *not* make perfect.

Surprisingly, spending more time on the disambiguation task associates with a significant *reduction* in accuracy ($p < 0.001$).⁷ Note that this is *after* we non-parametrically control for instance difficulty and worker ability. For every additional minute spent, a Turker is 3.6% less likely to answer correctly. We posit three theories: (1) taking breaks leads to loss of concentration (2) the “knee-jerk” response is best (rumination should be discouraged), and (3) although we control for instance difficulty, an instance may only be difficult for particular workers as evidenced by their taking longer.

7.3.3 Turker equality

In order to replicate previous work, we investigate Turker equality and the presence of spammers and superstars via plotting the number of annotations correct by the number of annotations completed in figure 7.3. To test the null hypothesis that all workers are equal (and thus, average), each worker’s *total contributions* are assumed to be drawn from independent Binomial random variables with probability of success $p = 73.4\%$ (the experimental average). Does the worker’s confidence interval (CI)

⁷We validated this linear approximation by regressing time spent as a polynomial and found the effect to be monotonically decreasing with a flat stretch in the middle.

contain p ? Figure 7.3 reveals that every worker has approximately the same capacity for performing coarse-grained WSD except for two above-average superstars and two below-average.

To test for spammers, we test against the null hypothesis of random answering, $p = 25.5\%$ (determined by simulation). Among workers who did a significant number of tasks,⁸ we find only one worker who may be a spammer. We echo Akkaya et al. (2010), Snow et al. (2008), and Singh et al. (2002) and conclude there is minimal spammer contribution. Once again, we do not observe a change in accuracy by quantity of tasks completed, an observation confirmed using regression (table 3).

7.3.4 Combining responses to optimize prediction

We can combine the 10 unique disambiguation responses for each of the 1000 examples to yield higher accuracy. Our algorithm is naive — we take the plurality vote and arbitrate ties randomly. Snow et al. (2008) found such an approach results in higher accuracy for disambiguating ‘president’. We wondered if the same is true for our more extensive dataset and annotations.

# of Annotations	2	3	4	5	6	7	8	9	10	2.4 (1st plurality)
Accuracy	.734	.795	.808	.824	.830	.837	.840	.843	.857	.811

Table 7.3: Accuracy of the WSD task using plurality voting for different numbers of Turkers. The last column is the accuracy of the variable algorithm: starting with two workers and adding an additional worker until plurality.

Table 4 illustrates our results. There is an overall accuracy of 85.7% when anno-

⁸We do not have significant power to claim a worker has accuracy of even 50% until about $n = 79$ at the Bonferroni-corrected α level.

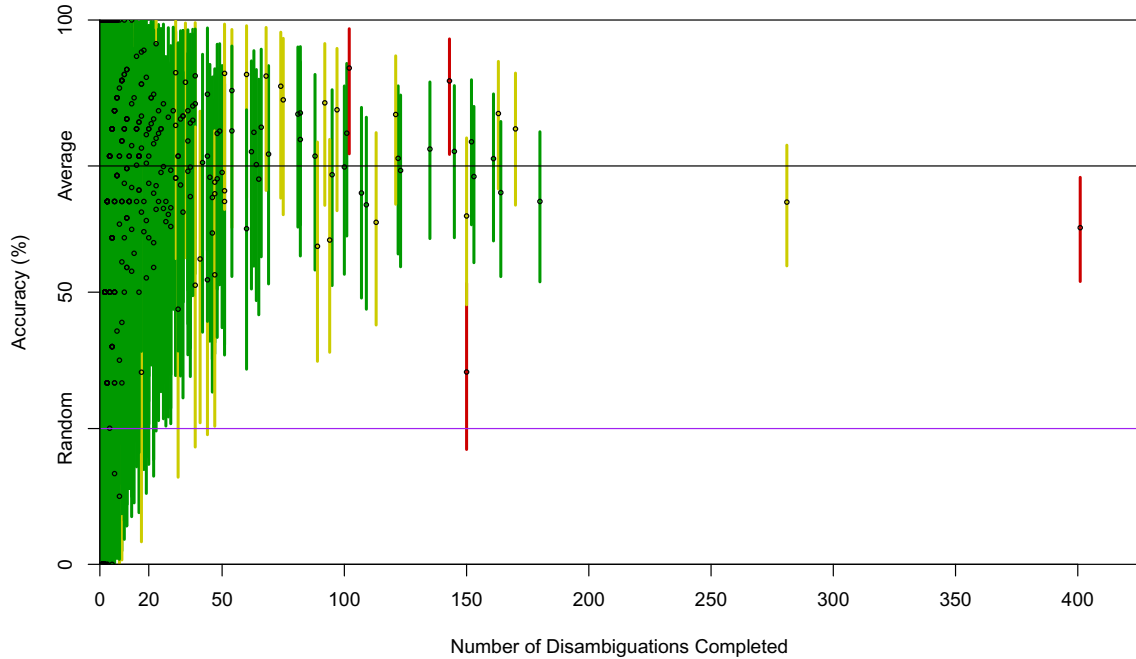


Figure 7.3: Accuracy of all 595 Turkers. The black line is the average accuracy ($p = 73.4\%$) and the purple line represents random sense choice accuracy (25.5%). We plot the Bonferroni-corrected Binomial proportion confidence intervals in green if they include p , yellow if the non-Bonferroni-corrected confidence intervals do not include p , and red if neither include p .

tations from all workers are aggregated. This is in the ballpark of the best supervised statistical learning techniques which boast almost 90% (Pradhan et al., 2007).⁹ We determined the marginal accuracy of each added Turker by simulating random subsets of two Turkers, three Turkers, etc and employed the same plurality vote.

With techniques such as discarding results from annotators who often disagree, and giving the annotators sense choices in order of sense frequency, we could likely achieve higher accuracy.

⁹Note that this is not a fair comparison. These supervised algorithms were given all the training data while Turkers were *not* given any previous examples. They also arbitrated based on the senses' frequencies while we randomized the order that the senses appeared in. Finally, they were not limited to polysemous words as we were.

Given MTurk annotation costs, we believe this system can be extended to accurately disambiguate a million words a year at 80% accuracy for about \$25,000. This demonstrates the system’s potential for mass annotation, but we reiterate that the main goal of this current work was to gain insight into drivers of WSD accuracy.

7.4 Conclusion

We performed the first extensive study of crowdsourced coarse-grained word sense disambiguation in order to gain insight into the behavioral and linguistic features that affect accuracy of the untrained annotations. As expected, we found results improved when there were less sense choices or when the target word was a noun, and that untrained workers did not improve with experience. However, we also discovered surprising insights: (1) the number of rephrasings in the correct sense definition corresponded with improved annotator accuracy, (2) frequency of target word corresponded with lower accuracy, and (3) time-spent on an individual annotation corresponded with lower accuracy. It also seems that time pressure may increase accuracy. Future experiments that prove these relationships causally may be fruitful. Lastly, we looked at Turker ability and found that they are all roughly equal in ability, and although individually not as accurate as experts, many Turkers may be pooled to improve accuracy.

Acknowledgments

We thank Mark Liberman and Lynn Selhat for helpful comments and discussions. Adam Kapelner thanks the National Science Foundation for the Graduate Research Fellowship that made this work possible.

Bayesian Additive Regression Trees Implementation*

Abstract

We present a new package in R implementing Bayesian Additive Regression Trees (BART). The package introduces many new features for data analysis using BART such as variable selection, interaction detection, model diagnostic plots, incorporation of missing data and the ability to save trees for future prediction. It is significantly faster than the current R implementation, parallelized, and capable of handling both large sample sizes and high-dimensional data.

8.1 Introduction

Ensemble-of-trees methods have become popular choices for forecasting in both regression and classification problems. Algorithms such as Random Forests (Breiman, 2001b) and stochastic gradient boosting (Friedman, 2002) are two well-established and widely employed procedures. Recent advances in ensemble methods include Dynamic Trees (Taddy et al., 2011) and Chipman et al. (2010)'s Bayesian Additive Regres-

*Joint work with Justin Bleich

sion Trees (BART), which depart from predecessors in that they rely on an underlying Bayesian probability model rather than a pure algorithm. BART has demonstrated substantial promise in a wide variety of simulations and real world applications such as predicting avalanches on mountain roads (Blattenberger and Fowles, 2014), predicting how transcription factors interact with DNA (Zhou and Liu, 2008) and predicting movie box office revenues (Eliashberg, 2010). This paper introduces `bartMachine`, a new R package that significantly expands the capabilities of using BART for data analysis.

Currently, there exists a single implementation of BART on CRAN: `BayesTree`, the package developed by the algorithm's original authors. One of the major drawbacks of this implementation is its lack of a `predict` function. Test data must be provided as an argument during the training phase of the model. Hence it is impossible to generate forecasts on future data without re-fitting with the entire model. Since the run time is not trivial, forecasting becomes an arduous exercise. A significantly faster implementation of BART that contains master-slave parallelization exists as Pratola et al. (2013), but this is only available as standalone C++ source code and not integrated with R.

The goal of `bartMachine` is to provide a fast, easy-to-use, visualization-rich machine learning package for R users. Our implementation of BART is in Java and is integrated into R via `rJava` (Urbanek, 2011). From a runtime perspective, our algorithm is significantly faster and is parallelized, allowing computation on as many cores as desired. Not only is the model construction itself parallelized, but the additional features such as prediction, variable selection, and many others can be divided across cores as well.

Additionally, we include a variety of expanded and new features. We implement the ability to save trees in memory and provide convenience functions for prediction on test data. We also include plotting functions for both credible and predictive

intervals and plots for visually inspecting convergence of BART's Gibbs sampler. We expand variable importance exploration to include permutation tests and interaction detection. We implement recently developed features for BART including a principled approach to variable selection and the ability to incorporate in prior information for covariates (Bleich et al., 2013). We also implement the strategy found in Chapter 9 of this document (Kapelner and Bleich, 2013a) to incorporate missing data during training and handle missingness during prediction.

In Section 9.2, we provide an overview of BART with a special emphasis on the features that have been extended. In Section 8.3 we provide a general introduction to the package, highlighting the novel features. Section 8.4 provides step-by-step examples of the regression capabilities and Section 8.5 introduces additional step-by-step examples of features unique to classification problems. We conclude in Section 9.6. Appendix A.4.1 discusses the details of our implementation and how it differs from BayesTree. Appendix A.4.5 offers predictive performance comparisons.

8.2 Overview of BART

BART provides a unique approach to nonparametric function estimation using regression trees. Regression trees rely on recursive binary partitioning of predictor space into a set of hyperrectangles in order to approximate some unknown function f . Such trees have received praise for their ability to flexibly fit interactions and nonlinearities. Models composed of sums of regression trees are able to capture additive effects in f better than single trees.

BART can be considered a sum-of-trees ensemble, with a novel estimation approach which relies on a fully Bayesian probability model. Specifically, the BART model can be expressed as:

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon} \approx \mathfrak{T}_1^{\text{leaf}}(\mathbf{X}) + \mathfrak{T}_2^{\text{leaf}}(\mathbf{X}) + \dots + \mathfrak{T}_m^{\text{leaf}}(\mathbf{X}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n) \quad (8.1)$$

Here we have m distinct regression trees, each composed of a tree structure, denoted by \mathfrak{T} , and the parameters at the terminal nodes (also called leaves), denoted by leaf . The two together, denoted as $\mathfrak{T}^{\text{leaf}}$ represents an entire tree with both its structure and set of leaf parameters.

The structure of a given tree \mathfrak{T}_t includes information on how any observation recurses down the tree. For each nonterminal (internal) node of the tree, there is a “splitting rule” taking the form $\mathbf{x}_j < c$ consisting of the “splitting variable” \mathbf{x}_j and the “splitting value” c . An observation moves to the left daughter node if the condition set by the splitting rule is satisfied and to the right daughter node otherwise. The process continues until a terminal node is reached. Then, the observation receives the leaf value of the terminal node as its predicted value. We denote the set of tree’s leaf parameters as $\text{leaf}_t = \{\mu_{t,1}, \mu_{t,2}, \dots, \mu_{t,b_t}\}$ where b_t is the number of terminal nodes for a given tree.

BART can be distinguished from other ensemble-of-trees models due to its underlying probability model. As a Bayesian model, BART consists of a set of priors for the structure and the leaf parameters and a likelihood for data in the terminal nodes. The aim of the priors is to provide regularization, preventing any single regression tree from dominating the total fit.

We provide an overview of the BART priors and likelihood and then discuss how draws from the posterior distribution are made. A more complete exposition can be found in Chipman et al. (2010).

8.2.1 Priors and Likelihood

There are three priors for the BART model: a prior on the tree structure itself, a prior on the leaf parameters, and a prior on the error variance σ^2 . The prior on σ^2 is independent from the other two and each tree is independent, yielding:

$$\begin{aligned} \mathbb{P}(\mathfrak{T}_1^{\text{leaf}}, \dots, \mathfrak{T}_m^{\text{leaf}}, \sigma^2) &= \left[\prod_t \mathbb{P}(\mathfrak{T}_t^{\text{leaf}}) \right] \mathbb{P}(\sigma^2) = \left[\prod_t \mathbb{P}(\mathcal{L}_t | \mathfrak{T}_t) \mathbb{P}(\mathfrak{T}_t) \right] \mathbb{P}(\sigma^2) \\ &= \left[\prod_t \prod_{\ell} \mathbb{P}(\mu_{t,\ell} | \mathfrak{T}_t) \mathbb{P}(\mathfrak{T}_t) \right] \mathbb{P}(\sigma^2) \end{aligned}$$

where the last line follows from an additional assumption of conditional independence of the leaf parameters given the tree's structure.

The first prior is on the locations of nodes within the tree. Nodes at depth d are nonterminal with probability $\alpha(1+d)^{-\beta}$ where $\alpha \in (0, 1)$ and $\beta \in [0, \infty]$. This prior keeps the trees shallow, limiting complexity of any single tree. Default values for these hyperparameters of $\alpha = 0.95$ and $\beta = 2$ are recommended by Chipman et al. (2010).

For nonterminal nodes, splitting rules have the following prior. First, a predictor is randomly selected to serve as the splitting variable. In the original formulation, each available predictor is equally likely to be chosen, but this is relaxed in our implementation to allow an arbitrary discrete distribution (see Section 8.4.11). Then, the splitting value is selected by randomly choosing a value of the selected predictor with equal probability.

The third prior is on the leaf parameters. Given a tree with a set of terminal nodes, each terminal node (or leaf) has a continuous parameter (the leaf parameter) representing the “best guess” of the response in this partition of predictor space. This parameter is the fitted value assigned to any observation that lands in that node. The

prior on each of the leaf parameters is given as: $\mu_\ell \stackrel{iid}{\sim} \mathcal{N}(\mu_\mu, \sigma_\mu^2)$. The expectation, μ_μ , is picked to be the range center, $(y_{\min} + y_{\max})/2$. The variance is empirically chosen so that the range center plus or minus $k = 2$ variances cover 95% of the provided response values in the training set (by default). The aim of this prior is to provide model regularization by shrinking the leaf parameters towards the center of the distribution of the response.

The final prior is on the error variance and is chosen to be $\text{InvGamma}(\nu/2, \nu\lambda/2)$. λ is determined from the data so that there is a $q = 90\%$ a priori chance (by default) that the BART model will improve upon the RMSE from an ordinary least squares regression. Therefore, the majority of the prior probability mass lies below the RMSE from least squares regression. Additionally, this prior limits the probability mass placed on small values of σ^2 to prevent overfitting.

Note that the adjustable hyperparameters are α , β , k , ν and q . Default values generally provide good performance, but optimal tuning can be achieved via cross-validation, an automatic feature implemented and described in Section 8.4.2.

Along with a set of priors, BART specifies the likelihood of responses in the terminal nodes. They are assumed a priori Normal with the mean being the “best guess” in the leaf at the current moment (in the Gibbs sampler) and variance being the best guess of the variance at the moment i.e. $\mathbf{y}_\ell \sim \mathcal{N}(\mu_\ell, \sigma^2/m)$. Note that σ^2 is scaled by the number of trees in order that the sum of the variances across the m trees is σ^2 .

8.2.2 Posterior Distribution and Prediction

A Gibbs sampler (Geman and Geman, 1984) is employed to generate draws from the posterior distribution of $\mathbb{P}(\mathfrak{F}_1^{\otimes m}, \dots, \mathfrak{F}_m^{\otimes m}, \sigma^2 \mid \mathbf{y})$. A key feature of the Gibbs sampler for BART is to employ a form of “Bayesian backfitting” (Hastie and Tibshirani, 2000)

where the j th tree is fit iteratively, holding all other $m - 1$ trees constant by exposing only the residual response that remains unfitted:

$$\mathbf{R}_j := \mathbf{y} - \sum_{t \neq j} \mathfrak{F}_t^{\mathcal{L}}(\mathbf{X}). \quad (8.2)$$

The Gibbs sampler,

$$\begin{aligned} 1 : & \quad \mathfrak{F}_1 \mid \mathbf{R}_{-1}, \sigma^2 \\ 2 : & \quad \mathcal{L}_1 \mid \mathfrak{F}_1, \mathbf{R}_{-1}, \sigma^2 \\ 3 : & \quad \mathfrak{F}_2 \mid \mathbf{R}_{-2}, \sigma^2 \\ 4 : & \quad \mathcal{L}_2 \mid \mathfrak{F}_2, \mathbf{R}_{-2}, \sigma^2 \\ & \quad \vdots \\ 2m - 1 : & \quad \mathfrak{F}_m \mid \mathbf{R}_{-m}, \sigma^2 \\ 2m : & \quad \mathcal{L}_m \mid \mathfrak{F}_m, \mathbf{R}_{-m}, \sigma^2 \\ 2m + 1 : & \quad \sigma^2 \mid \mathfrak{F}_1, \mathcal{L}_1, \dots, \mathfrak{F}_m, \mathcal{L}_m, \mathbf{E}, \end{aligned} \quad (8.3)$$

proceeds by first proposing a change to the first tree's structure \mathfrak{F} which are accepted or rejected via a Metropolis-Hastings step (Hastings, 1970). Note that sampling from the posterior of the tree structure does not depend on the leaf parameters, as they can be analytically margined out of the computation (see Appendix A.5.3.1). Given the tree structure, samples from the posterior of the b leaf parameters $\mathcal{L}_1 := \{\mu_1, \dots, \mu_b\}$ are then drawn. This procedure proceeds iteratively for each tree, using the updated set of partial residuals \mathbf{R}_j . Finally, conditional on the updated set of tree structures and leaf parameters, a draw from the posterior of σ^2 is made based on the full model residuals $\mathbf{E} := \mathbf{y} - \sum_{t=1}^m \mathfrak{F}_t^{\mathcal{L}}(\mathbf{X})$.

Within a given terminal node, since both the prior and likelihood are normally distributed, the posterior of each of the leaf parameters in \mathcal{L} is conjugate normal with its mean being a weighted combination of the likelihood and prior parameters (lines 2, 4, \dots , $2m$ in equation set A.7). Due to the normal-inverse-gamma conjugacy, the posterior of σ^2 is inverse gamma as well (line $2m + 1$ in equation set A.7). The complete expressions for these posteriors can be found in Gelman et al. (2004).

Lines 1, 3, \dots , $2m - 1$ in equation set A.7 rely on Metropolis-Hastings draws from the posterior of the tree distributions. These involve introducing small perturbations to the tree structure: growing a terminal node by adding two daughter nodes, pruning two daughter nodes (rendering their parent node terminal), or changing a split rule. We denote these possible alterations as: GROW, PRUNE, and CHANGE.² The mathematics associated with the Metropolis-Hastings step is simple but is tedious, and we refer the interested reader to Appendix A.4.1 for the explicit calculations. Once again, over many Gibbs samples, trees can dynamically morph their structure in an effort to capture the fit left currently unexplained.

Pratola et al. (2013) argue that a CHANGE step is unnecessary for sufficient mixing of the Gibbs sampler. While we too observed this to be true for estimates of the posterior means, we found that omitting CHANGE can negatively impact the variable inclusion proportions (the feature introduced in Section 8.4.6). As a result, we implement a modified CHANGE where we only propose new splits for nodes that are singly internal: both children nodes are terminal nodes (details are given in Appendix A.5.3.3).

All $2m + 1$ steps represent a *single* Gibbs iteration. We have observed that generally no more than 1,000 iterations are needed as “burn-in” (see Section 8.4.4 for convergence diagnostics). An additional 1,000 iterations is usually sufficient to serve

²In the original formulation, Chipman et al. (2010) include an additional alteration called SWAP. Due to the complexity of bookkeeping associated with this alteration, we do not implement it.

as draws from the posterior for $f(\mathbf{x})$. A single predicted value $\hat{f}(\mathbf{x})$ can be obtained by taking the average of the posterior values and a quantile estimate can be obtained by computing the appropriate quantile of the posterior values. Additional features of the posterior distribution will be discussed in Section 8.4.

8.2.3 BART for Classification

BART can easily be modified to handle classification problems for categorical response variables. In Chipman et al. (2010), only binary outcomes were explored but recent work has extended BART to the multiclass problem (Kindo et al., 2013). Our implementation handles binary classification and we plan to implement multiclass outcomes in a future release.

For the binary classification problem (coded with outcomes “0” and “1”), we assume a probit model,

$$\mathbb{P}(Y = 1 \mid \mathbf{X}) = \Phi \left(\mathfrak{F}_1(\mathbf{X}) + \mathfrak{F}_2(\mathbf{X}) + \dots + \mathfrak{F}_m(\mathbf{X}) \right),$$

where Φ denotes the cumulative density function of the standard normal distribution. In this formulation, the sum-of-trees model serves as an estimate of the conditional probit at \mathbf{x} which can be easily transformed into a conditional probability estimate of $Y = 1$.

In the classification setting, the prior on σ^2 is not needed as the model assumes $\sigma^2 = 1$. The prior on the tree structure remains the same as in the regression setting and a few minor modifications are required for the prior on the leaf parameters.

Sampling from the posterior distribution is again obtained via Gibbs sampling with a Metropolis-Hastings step outlined in Section 8.2.2. Following the data augmentation approach of (Albert and Chib, 1993) an additional vector of latent variables \mathbf{Z} is

introduced into the Gibbs sampler. Then, a new step is created in the Gibbs sampler where draws of $\mathbf{Z} | \mathbf{y}$ are obtained by conditioning on the sum-of-trees model:

$$Z_i | y_i = 1 \sim \max N \left(\sum_t \mathbb{1}_t^{\otimes}(\mathbf{X}), 1 \right), 0 \text{ and}$$

$$Z_i | y_i = 0 \sim \min \left\{ N \left(\sum_t \mathbb{1}_t^{\otimes}(\mathbf{X}), 1 \right), 0 \right\}.$$

Next, \mathbf{Z} is used as the response vector instead of \mathbf{y} in all steps of Equation A.7.

Upon obtaining a sufficient number of samples from the posterior, inferences can be made using the the posterior distribution of conditional probabilities and classification can be undertaken by applying a threshold to the to the means (or another summary) of these posterior probabilities. The relevant classification features of `bartMachine` are discussed in Section 8.5.

8.3 The `bartMachine` package

The package `bartMachine` provides a novel implementation of Bayesian Additive Regression Trees in R. The algorithm is substantially faster than the current R package `BayesTree` and our implementation is parallelized at the Gibbs sample level during prediction. Additionally, the interface with `rJava` allows for the entire posterior distribution of tree ensembles to persist throughout the R session, allowing for prediction and other calls to the trees without having to re-run the Gibbs sampler (a limitation in the current implementation). The model object cannot persist across sessions (using R's `save` command for instance) and we view the addition of this feature as future work. Since our implementation is different from `BayesTree`, we provide a predictive accuracy bakeoff on different datasets in Appendix A.4.5 which illustrates that the two are about equal.

8.3.1 Speed Improvements and Parallelization

We make a number of significant speed improvements over the original implementation.

First, `bartMachine` is fully parallelized (with the number of cores customizable) during model creation, prediction, and many of the other features. During model creation, we chose to parallelize by creating one independent Gibbs chain per core. Thus, if we have 500 burn-in samples and 1,000 post burn-in samples and four cores, each core would sample 750 samples: 500 for a burn-in and 250 post burn-in samples. The final model will aggregate the 250 post burn-in samples for the four cores yielding the desired 1,000 total post burn-in samples. This has the drawback of effectively running the burn-in serially, but has the added benefit of reducing auto-correlation of the sum-of-trees samples in the posterior samples since the chains are independent which may provide greater predictive performance. Parallelization at the level of likelihood calculations is left for a future release. Parallelization for prediction and other features scale linearly in the number of cores.

Additionally, we take advantage of a number of additional computational shortcuts:

1. Computing the unfitted responses for each tree (Equation 8.2) can be accomplished by keeping a running vector and making entry-wise updates as the Gibbs sampler (Equation A.7) progresses from step 1 to $2m$. Additionally, during the σ^2 step $2m + 1$, the residuals do not have to be computed by dropping the data down all the trees.
2. Each node caches its acceptable variables for split rules and the acceptable unique split values so they do not need to be calculated at each tree sampling step. This speed enhancement, which we call *memcache* comes at the expense of memory and may cause issues for large data sets. We include a toggle in our implementation

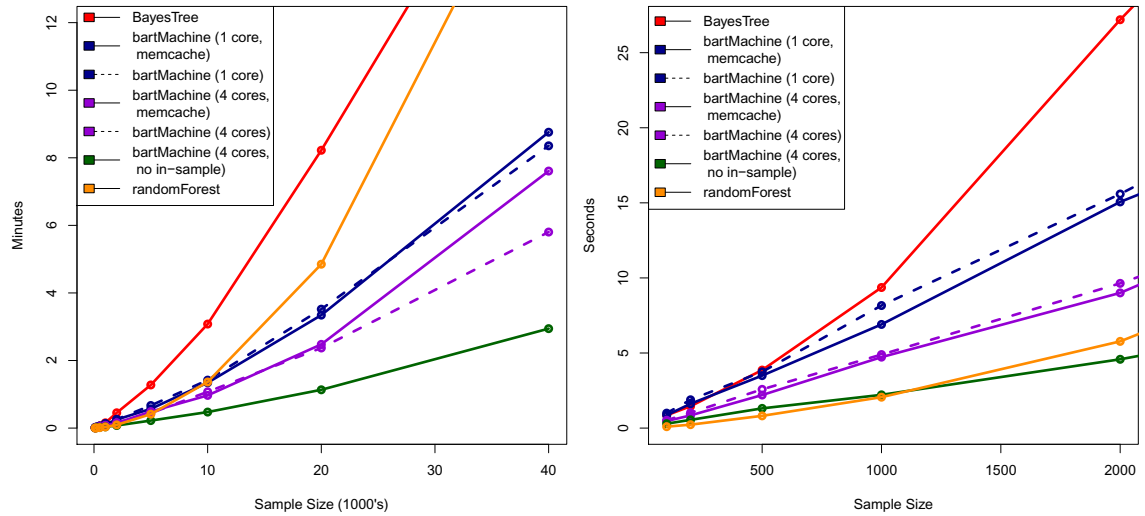
defaulted to “on.”

3. Careful calculations in Appendix A.4.1 eliminate many unnecessary computations. For instance, the likelihood ratios are only functions of the squared sum of responses and no longer require computing the sum of the responses squared.

Figure 8.1 displays model creation speeds for different values of n on a linear model with $p = 20$, normally distributed covariates, $\beta_1, \dots, \beta_{20} \stackrel{iid}{\sim} U(-1, 1)$, and standard normal noise. Note that we do not vary p as it was already shown in Chipman et al. (2010) that BART’s computation time is largely unaffected by the dimensionality of the problem (relative to the influence of sample size). We include results for BART using `BayesTree` (Chipman et al., 2010), `bartMachine` with one and four cores, the `memcache` option on and off, as well as four cores, `memcache` off and computation of in-sample statistics off (all with $m = 50$ trees). We also include Random Forests via the package `randomForest` (Liaw and Wiener, 2002) with its default settings.

We first note that Figure 8.1a demonstrates that the `bartMachine` model creation runtime is approximately linear in n . There is about a 30% speed-up when using four cores instead of one. The `memcache` enhancement should be turned off only with sample sizes larger than $n = 20,000$. Noteworthy is the 50% reduction in time of constructing the model when not computing in-sample statistics. In-sample statistics are computed by default because the user generally wishes to see them. Also, for the purposes of this comparison, `BayesTree` models compute the in-sample statistics by necessity since the trees are not saved. The `randomForest` implementation becomes slower just after $n = 1,000$ due to its reliance on a greedy exhaustive search at each node.

Figure 8.1b displays results for smaller sample sizes ($n \leq 2,000$) that are often encountered in practice. We observe the `memcache` enhancement provides about a 10% speed improvement. Thus, if memory is an issue, it can be turned off with little



(a) Large sample sizes

(b) Small sample sizes

Figure 8.1: Model creation times as a function of sample size for a number of settings of `bartMachine`, `BayesTree` and `RandomForests`. Simulations were run on a quad-core 3.4GHz Intel i5 desktop with 24GB of RAM running the Windows 7 64bit operating system.

performance degradation.

8.3.2 Missing Data in BART

`bartMachine` implements a native method for incorporating missing data into both model creation and future prediction with test data. The details are given in Chapter 9 of this document (Kapelner and Bleich, 2013a) but we provide a brief summary here.

There are a number of ways to incorporate missingness into tree-based methods (see Ding and Simonoff, 2010 for a review). The method implemented here is known as “Missing Incorporated in Attributes” (MIA, Twala et al., 2008, section 2) which natively incorporates missingness by augmenting the nodes’ splitting rules to (a) also

handle sorting the missing data to the left or right and (b) use missingness *itself* as a variable to be considered in a splitting rule. Algorithm 3 summarizes these new splitting rules as they are implemented within the package.

Implementing MIA into the BART procedure is straightforward. These new splitting rules are sampled uniformly during the GROW or CHANGE steps. For example, a splitting rule might be “ $\mathbf{x}_j < c$ or \mathbf{x}_j is missing.” To account for splitting on missingness itself, we create dummy vectors of length n for each of the p attributes, denoted $\mathbf{M}_1, \dots, \mathbf{M}_p$, which assume the value 1 when the entry is missing and 0 when the entry is present. The original training matrix is then augmented with these dummies, giving the opportunity to select missingness *itself* when choosing a new splitting rule during the grow or change steps. Note that this can increase the number of predictors by up to a factor of 2. We illustrate the building a BART model with missing data in Section 8.4.9. As described in Chipman et al. (2010, Section 6), BART’s runtime increases negligibly in the number of covariates and this has been our experience using the augmented training matrix.

Algorithm 2 The MIA choices for all attributes $j \in \{1, \dots, p\}$ and all split points x_{ij}^* where $i \in \{1, \dots, n\}$ during a GROW or CHANGE step in `bartMachine`.

- 1: If x_{ij} is missing, send it \leftarrow ; if it is present and $x_{ij} \leq x_{ij}^*$, send it \leftarrow , otherwise \rightarrow .
 - 2: If x_{ij} is missing, send it \rightarrow ; if it is present and $x_{ij} \leq x_{ij}^*$, send it \leftarrow , otherwise \rightarrow .
 - 3: If x_{ij} is missing, send it \leftarrow ; if it is present, send it \rightarrow .
-

8.3.3 Principled Variable Selection

Our package also implements the variable selection procedures developed in Bleich et al. (2013), which is best applied to data problems where the number of covariates

influencing the response is small relative to the total number of covariates. We give a brief summary of the procedures here.

In order to select variables, we make use of the “variable inclusion proportions,” the proportion of times each predictor is chosen as a splitting rule divided by the total number of splitting rules appearing in the model (see Section 8.4.6 for more details). The variable selection procedure can be outlined as follows:

1. Compute the model’s variable inclusion proportions.
2. Permute the response vector, thereby breaking the relationship between the covariates and the response. Rebuild the model and compute the “null” variable inclusion proportions. Repeat this a number of times to create a null permutation distribution.
3. Three selection rules are can be used depending on the desired stringency of selection:
 - (a) Local Threshold: Include a predictor \mathbf{x}_k if its variable inclusion proportion exceeds the $1 - \alpha$ quantile of its own null distribution.
 - (b) Global Max Threshold: Include a predictor \mathbf{x}_k if its variable inclusion proportion exceeds the $1 - \alpha$ quantile of the distribution of the maximum of the null variable inclusion proportions from each permutation of the response.
 - (c) Global SE Threshold: Select \mathbf{x}_k if its variable inclusion proportion exceeds a threshold based from its own null distribution mean and SD with a global multiplier shared by all predictors.

The Local procedure is the least stringent in terms of selection and the Global Max procedure the most. The Global SE procedure is a compromise, but behaves more similarly to the Global Max. Bleich et al. (2013) demonstrate that the best

procedure depends on the underlying sparsity of the problem, which is often unknown. Therefore, the authors include an additional procedure that chooses the best of these thresholds via cross-validation and this method is also implemented in `bartMachine`. Examples of these procedures for variable selection are provided in Section 8.4.10.

8.4 Regression Features

We illustrate the package features by using both real and simulated data, focusing first on regression problems.

8.4.1 Computing parameters

We first set some computing parameters. We allow up to 5GB of RAM for the Java heap (although we never used more than 1GB during this paper’s exploration)³ and we set the number of computing cores available for use to 4.

```
> library(bartMachine)
> set_bart_machine_num_cores(4)
> init_java_for_bart_machine_with_mem_in_mb(5000)
```

The following Sections 8.4.2 – 8.4.10 use a dataset obtained from UCI (Bache and Lichman, 2013). The $n = 201$ observations are automobiles and the goal is to predict each automobile’s price from 25 features (15 continuous and 10 nominal), first explored by Kibler et al. (1989).⁴ This dataset also contains missing data.

³Note that the maximum amount of memory can be set only *once* at the beginning of the R session (a limitation of `rJava` since only one Java Virtual Machine can be initiated per session), but the number of cores can be respecified at any time.

⁴We first preprocess the data. We first drop one of the nominal predictors (car company) due to too many categories (22). We then coerce two of the of the nominal predictors to be continuous. Further, the response variable, price, was logged to reduce right skew in its distribution.

8.4.2 Model Building

We are now ready to construct a `bartMachine` model. The default hyperparameters generally follow the recommendations of Chipman et al. (2010) and provide a ready-to-use algorithm for many data problems. Our hyperparameter settings are $m = 50$,⁵ $\alpha = 0.95$, $\beta = 2$, $k = 2$, $q = 0.9$, $\nu = 3$, probabilities of the GROW / PRUNE / CHANGE steps is 39% / 39% / 44%. We set the number of burn-in Gibbs samples to 250 and number of post-burn-in samples to 1,000. We default the missing data feature to be off. We default the covariates to be equally important *a priori*. Other parameters and their defaults can be found in the package's online manual. Below is a default `bartMachine` model. Here, \mathbf{X} denotes automobile attributes and \mathbf{y} denotes the log price of the automobile.

```
> bart_machine = build_bart_machine(X, y)
Building BART for regression ... evaluating in sample data...done
```

If one wishes to see the iterations of the Gibbs sampler of Equation A.7, the flag `verbose` can be set to "TRUE". One can see more debug information from the Java program by setting the flag `debug_log` to true and the program will print to `unnamed.log` in the current working directory. We now inspect the model object to query its in-sample performance and to be reminded of the input data and model hyperparameters.

Since the response was considered continuous, we employ BART for regression. The dimensions of the design matrix are given. Note that we dropped 45 observations that contained missing data (which we will retain in Section 8.4.9). We then have

⁵In contrast to Chipman et al. (2010), we recommend this default as a good starting point rather than $m = 200$ due to our experience experimenting with the "RMSE by number of trees" feature found in later in this section. Performance is often similar and computational time and memory requirements are dramatically reduced.


```

> bart_machine
Bart Machine v1.0b for regression

training data n = 160 and p = 46
built in 1 secs on 4 cores, 50 trees, 250 burn in and 1000 post. samples

sigsq est for y beforehand: 0.014
avg sigsq estimate after burn-in: 0.00886

in-sample statistics:
L1 = 9.03
L2 = 0.8
rmse = 0.07
Pseudo-Rsq = 0.9741
p-val for shapiro-wilk test of normality of residuals: 0.01785
p-val for zero-mean noise: 0.97692

```

Figure 8.2: The summary for the default `bartMachine` model built with the automobile data

a recording of the MSE for the OLS model and our average estimate of σ_e^2 . We are then given in-sample statistics on error. Pseudo- R^2 is calculated via $1 - SSE/SST$. Also provided are outputs from tests of the error distribution being mean centered and normal. In this case, we cannot conclude normality of the residuals using the Shapiro-Wilk test.

We can also obtain out-of-sample statistics to assess level of overfitting by using k-fold cross validation. Using 10 folds we find:

```

> k_fold_cv(X, y, k_folds = 10)
$L1_err          $L2_err          $rmse          $PseudoRsq
[1] 22.63155      [1] 5.202831      [1] 0.1803266    [1] 0.831917

```

The Pseudo- R^2 being lower out-of-sample versus in-sample suggests evidence that BART is slightly overfitting (note that the training sample during cross-validation is 10% smaller).

It may also be of interest how the number of trees m affects performance. One can examine how out-of-sample predictions vary by the number of trees via

```
> rmse_by_num_trees(bart_machine, num_replicates = 20)
```

and the output is shown in Figure 8.3.

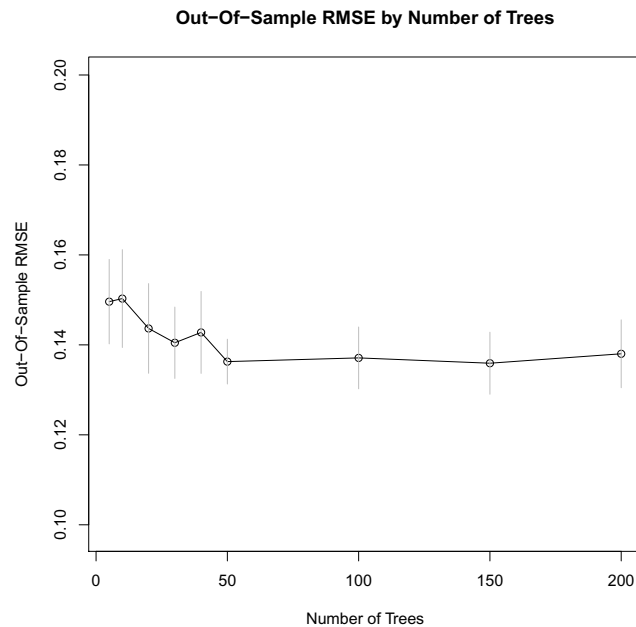


Figure 8.3: Out-of-sample predictive performance by number of trees

It seems that increasing $m > 50$ does not result in any substantial increase in performance. We can now try to build a better `bartMachine` by grid-searching over a set of hyperparameter combinations, including m (for more details, see `BART-cv` in Chipman et al., 2010). The default grid search is small and it can be customized by the user.

```
> bart_machine_cv = build_bart_machine_cv(X, y)
```

```
...
```

```
BART CV win: k: 2 nu, q: 10, 0.75 m: 200
```

This function returns the “winning” model, which is the one with lowest out-of-sample RMSE over a 5-fold cross-validation. Here, the cross-validated `bartMachine` model has slightly better in-sample performance ($L1 = 8.18$, $L2 = 0.68$ and Pseudo- $R^2 = 0.978$) as well as slightly better out-of-sample performance:

```
> k_fold_cv(X, y, k_folds = 10, k = 2, nu = 3, q = 0.9, num_trees = 200)
$L1_err          $L2_err          $rmse          $PseudoRsq
[1] 21.21557      [1] 4.517916      [1] 0.1680386    [1] 0.8540439
```

Predictions are handled with the `predict` function:

```
predict(bart_machine_cv, X[1 : 14, ])
  9.479963  9.775766  9.799110 10.050041  9.659138  9.697902  9.873622
  9.931972  8.550436  8.688874  8.794351  8.647986  8.690300  9.029066
```

We also include a convenience method `bart_predict_for_test_data` that will predict and return out-of-sample error metrics when the test outcomes are known.

8.4.3 Model Destroying

As noted in the introduction to Section 8.3, `bartMachine` objects persist in Java for the entirety of an R session. These model objects can use a substantial amount of RAM and it is prudent to release this memory when it is no longer needed. Simply removing a `bartMachine` in R does *not* destroy the Java object and release the RAM (resulting in a memory leak).

Therefore, we provide a utility function `destroy_bart_machine` that cleans up the Java object. This function should be called when a `bartMachine` object is no longer needed and *before* removing or overwriting the R variable. Since we no longer are using the original `bart_machine` object, we now release its memory via:

```
destroy_bart_machine(bart_machine)
```

8.4.4 Assumption Checking

The package includes features that assess the plausibility of the BART model assumptions. Checking the mean-centeredness of the noise is addressed in the summary output of Figure 8.2 and is simply a one-sample t -test of the average residual value against a null hypothesis of true mean zero. We assess both normality and heteroskedasticity via:

```
> check_bart_error_assumptions(bart_machine_cv)
```

This will display a window similar to Figure 8.4 which contains a QQ-plot (to assess normality) as well as a residual-by-predicted plot (to assess homoskedasticity). It appears that the errors are most likely normal and homoskedastic.

In addition to the model assumptions, BART requires convergence of its Gibbs sampler. Figure 8.5 displays four types of convergence diagnostics.

8.4.5 Credible Intervals and Prediction Intervals

An advantage of BART is that if we believe the priors and model assumptions, the Bayesian probability model and corresponding burned-in Gibbs samples provide the approximate posterior distribution of $f(\mathbf{x})$. Thus, one can compute uncertainty estimates via quantiles of the posterior samples. These provide Bayesian “credible intervals” which are intervals for the conditional expectation function, $\mathbb{E}[\mathbf{y} \mid \mathbf{X}]$.

Another useful uncertainty interval can be computed for individual predictions by combining uncertainty from the conditional expectation function with the systematic, homoskedastic normal noise produced by \mathcal{E} . Since we have draws from the posterior of the conditional expectation distribution and concomitant draws from the posterior of the variance distribution, we can simulate the distribution of the response by drawing many realizations from $\mathcal{N}(f(\mathbf{x})_n, \sigma_n^2)$ for each the post-burn-in samples

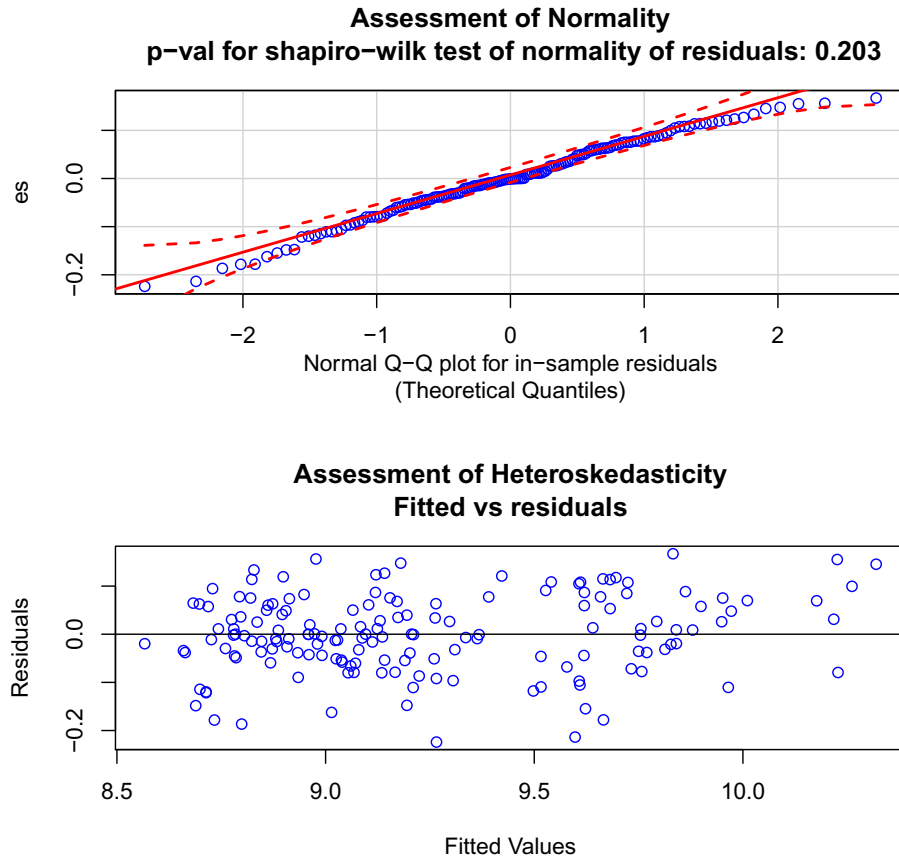


Figure 8.4: Test of normality of errors using QQ-plot and the Shapiro-Wilk test (top), residual plot to assess heteroskedasticity (bottom).

and aggregating them. The prediction interval is then provided by the appropriate quantiles of the posterior samples.

Below is an example of how both types of intervals are computed in the package (for the 100th observation of the training data):

```
> calc_credible_intervals(bart_machine_cv, X[100, ], ci_conf = 0.95)
      ci_lower_bd ci_upper_bd
[1,]    8.725202    8.971687
> calc_prediction_intervals(bart_machine_cv, X[100, ], pi_conf = 0.95)
      pi_lower_bd pi_upper_bd
```

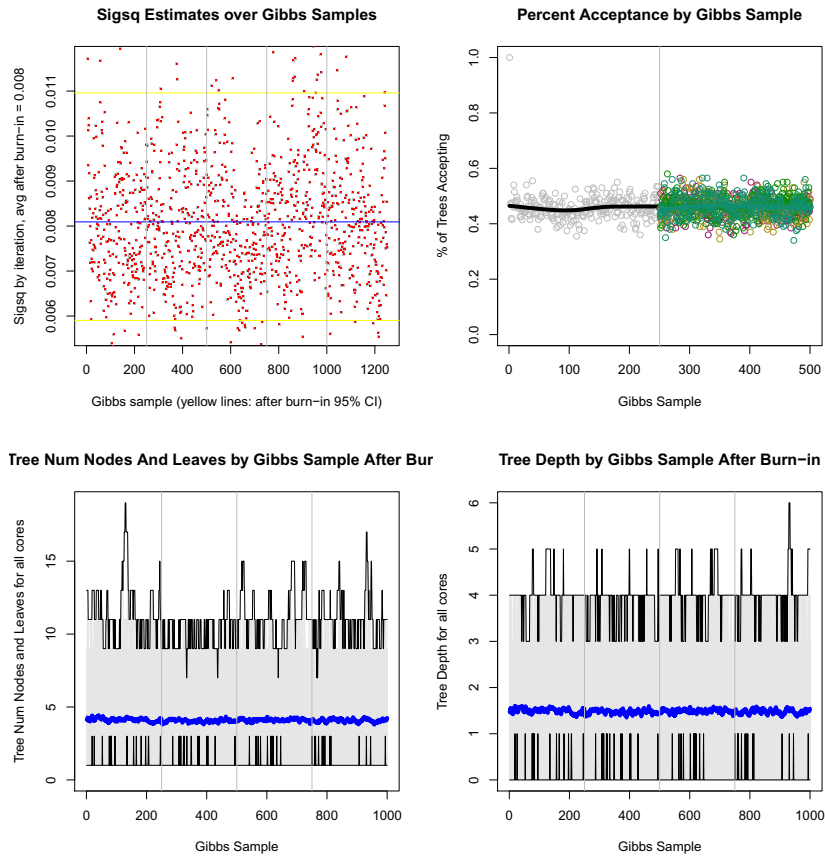


Figure 8.5: Convergence diagnostics for the cross-validated `bartMachine` model. Top left: σ^2 by Gibbs sample. Samples to the left of the first vertical grey line are burn-in from the first computing core's Gibbs sample chain. The four subsequent plots separated by grey lines are the post-burn-in samples from each of the four computing cores employed during model construction. Top right: percent acceptance of Metropolis-Hastings proposals across the m trees where each point plots one iteration. Points before the grey vertical line illustrate burn-in samples and points after illustrate post burn-in. Each computing core is colored differently. Bottom left: average number of leaves across the m trees by sample (post burn-in only where computing cores separated by vertical grey lines). Bottom right: average tree depth across the m trees by sample (post burn-in only where computing cores separated by vertical grey lines).

```
[1,]      8.631243      9.06353
```

Note that the prediction intervals are wider than the credible intervals because they reflect the uncertainty from the error term.

We can then plot these intervals in sample:

```
> plot_y_vs_yhat(bart_machine_cv, credible_intervals = TRUE)
> plot_y_vs_yhat(bart_machine_cv, prediction_intervals = TRUE)
```

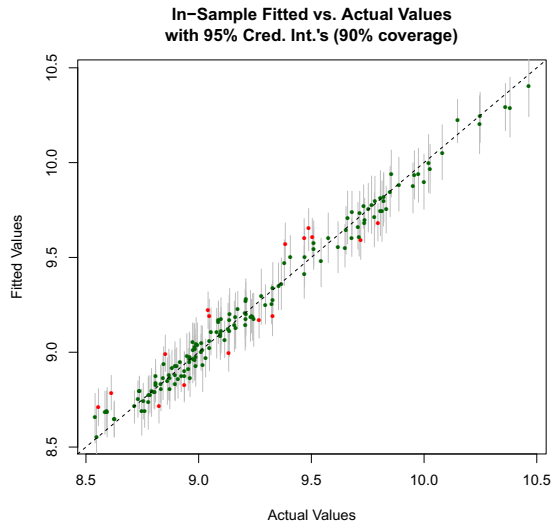
Figure 8.6a shows how our prediction fared against the original response (in-sample) with 95% credible intervals. Figure 8.6b shows the same prediction versus the original response plot now with 95% prediction intervals.

8.4.6 Variable Importance

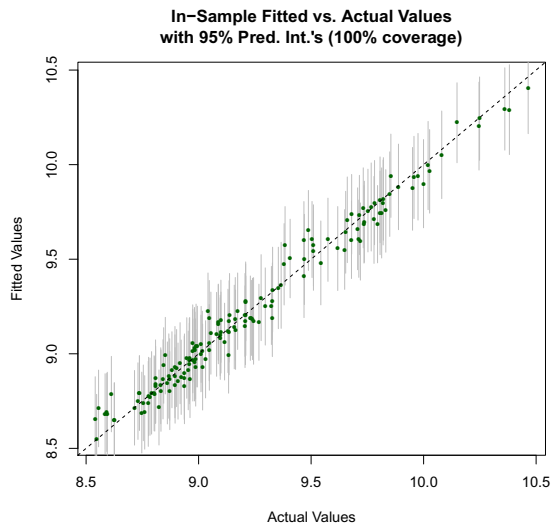
After a BART model is built, it is natural to ask the question: which variables are most important? This is assessed by examining the splitting rules in the m trees across the post burn-in gibbs samples which are known as “inclusion proportions” (Chipman et al., 2010). The inclusion proportion for any given predictor represents the proportion of times that variable is chosen as a splitting rule out of all splitting rules among the posterior draws of the sum-of-trees model. Figure 8.7 illustrates the inclusion proportions for all variables obtained via:

```
> investigate_var_importance(bart_machine_cv,
  num_replicates_for_avg = 20)
```

Actual selection of variables *significantly* affecting the response is addressed conceptually in Section 8.3.3 and examples are provided in Section 8.4.10.



(a) Segments illustrate credible intervals



(b) Segments illustrate prediction intervals

Figure 8.6: Fitted versus actual response values for the automobile dataset. Segments are 95% credible intervals (a) or 95% prediction intervals (b). Green dots indicate the true response is within the stated interval and red dots indicate otherwise. Note that the percent coverage in (a) is not expected to be 95% because the response includes a noise term.

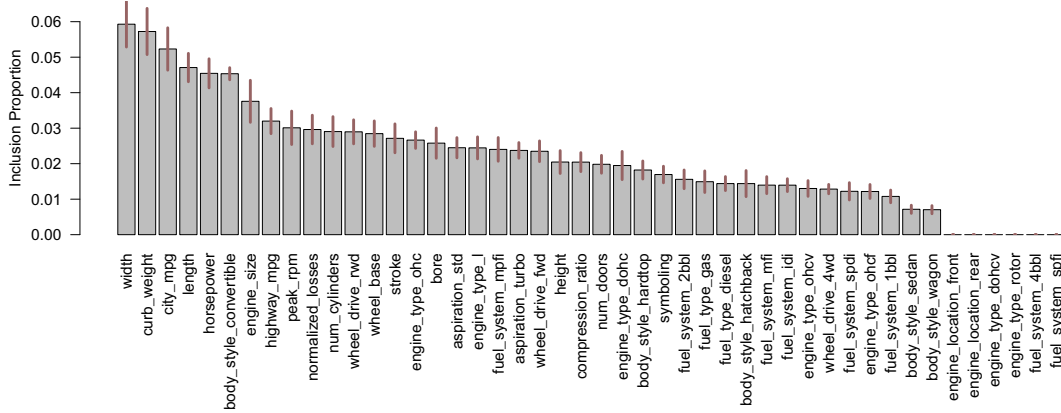


Figure 8.7: Average variable inclusion proportions in the cross-validated `bartMachine` model for the automobile data averaged over 100 model constructions to obtain stable estimates across many posterior modes in the sum-of-trees distribution (as recommended in Bleich et al., 2013). The segments atop the bars represent 95% confidence intervals. The eight predictors with inclusion proportions of zero feature identically one value (after missing data was dropped).

8.4.7 Variable Effects

It is also natural to ask: does \mathbf{x}_j affect the response, controlling for other variables in the model? This is roughly analogous to the t -test in ordinary least squares regression of no linear effect of \mathbf{x}_j on \mathbf{y} while controlling for \mathbf{x}_{-j} . The null hypothesis here is the same but the linearity constraint is relaxed. To test this, we employ a permutation approach where we record the observed Pseudo- R^2 from the `bartMachine` model built with the original data. Then we permute the \mathbf{x}_j th column, thereby destroying any relationship between \mathbf{x}_j and \mathbf{y} , construct a new duplicate `bartMachine` model from this permuted design matrix and record a “null” Pseudo- R^2 . We then repeat this process to obtain a null distribution of Pseudo- R^2 ’s. Since the alternative hypothesis is that \mathbf{x}_j has an effect on \mathbf{y} in terms of predictive power, our p_{val} is the proportion of null Pseudo- R^2 ’s greater than the observed Pseudo- R^2 , making our procedure a

natural one-sided test. Note, however, that this test is conditional on the BART model and its selected priors being true, similar to the assumptions of the linear model.

If we wish to test if a set of covariates $A \subset \{\mathbf{x}_1, \dots, \mathbf{x}_p\}$ affect the response after controlling for other variables, we repeat the procedure outlined in the above paragraph by permuting the columns of A in every null sample. This is roughly analogous to the partial F -test in ordinary least squares regression.

If we wish to test if *any* of the covariates matter in predicting \mathbf{y} , roughly analogous to the omnibus F -test in ordinary least squares regression, we simply permute \mathbf{y} during the null sampling. This procedure breaks the relationship between the response and the predictors but does not alter the existing associations between predictors.

At $\alpha = 0.05$, Figure 8.8a demonstrates an insignificant effect of the variable `width` of car on price. Even though `width` is putatively the “most important” variable as measured by proportions of splits in the posterior sum-of-trees model (Figure 8.7), note that this is largely an easy prediction problem with many collinear predictors. Figure 8.8b shows the results of a test of the putatively most important categorical variable, `body style` (which involves permuting the categories, then dummifying the levels to preserve the structure of the variable). We find a marginally significant effect ($p = 0.0495$). A test of the top ten most important variables is convincingly significant (Figure 8.8c). For the omnibus test, Figure 8.8d illustrates an extremely statistically significant result, as would be expected. The code to run these tests is shown below (output suppressed).

```
> cov_importance_test(bart_machine_cv, covariates = c("width"))
> cov_importance_test(bart_machine_cv, covariates = c("body_style"))
> cov_importance_test(bart_machine_cv, covariates = c("width",
  "curb_weight", "city_mpg", "length", "horsepower", "body_style",
  "engine_size", "highway_mpg", "peak_rpm", "normalized_losses"))
> cov_importance_test(bart_machine_cv)
```

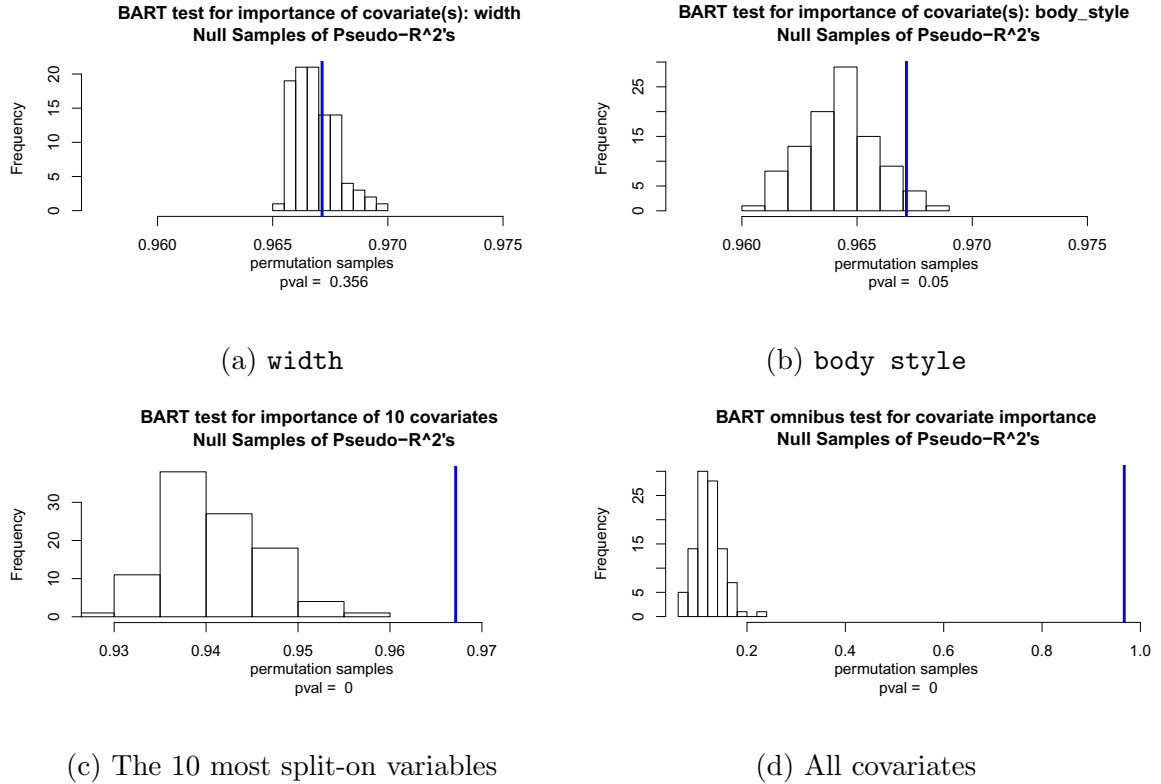


Figure 8.8: Tests of covariate importance conditional on the cross-validated `bartMachine` model. All tests performed with 100 null samples.

8.4.8 Partial Dependence

A data analyst may also be interested in understanding how \mathbf{x}_j affects the response on average, after controlling for other predictors. This can be examined using Friedman (2001)'s Partial Dependence Function (PDP),

$$f_j(\mathbf{x}_j) = \mathbb{E}_{\mathbf{x}_{-j}} [f(\mathbf{x}_j, \mathbf{x}_{-j})] = \int f(\mathbf{x}_j, \mathbf{x}_{-j}) dP(\mathbf{x}_{-j}), \quad (8.4)$$

where \mathbf{x}_{-j} denotes all variables except \mathbf{x}_j . The PDP of predictor \mathbf{x}_j gives the average value of f when \mathbf{x}_j is fixed and \mathbf{x}_{-j} varies over its marginal distribution, $dP(\mathbf{x}_{-j})$. As

neither the true model f nor the distribution of the predictors $dP(\mathbf{x}_{-j})$ are known, we estimate Equation 8.4 by computing

$$\hat{f}_j(\mathbf{x}_j) = \frac{1}{n} \sum_{i=1}^n \hat{f}(\mathbf{x}_j, \mathbf{x}_{-j,i}) \quad (8.5)$$

where n is the number of observations in the training data and \hat{f} denotes the `bartMachine` model. Since BART provides an estimated posterior distribution, we can plot credible bands for the PDP function. In Equation 8.5, the \hat{f} can be replaced with a function that calculates the q th quantile of the post-burned-in Gibbs samples for $\hat{\mathbf{y}}$. Figure 8.9a plots the PDP along with the 2.5%ile and the 97.5%ile for the variable `horsepower`. By varying over most of the range of `horsepower`, the price is predicted to increase by about \$1000. Figure 8.9b plots the PDP along with the 2.5%ile and the 97.5%ile for the variable `stroke`. This predictor seemed to be relatively unimportant according to Figure 8.7 and the PDP confirms this, with a very small, yet nonlinear average partial effect. The code for both plots is below.

```
> pd_plot(bart_machine_cv, j = "horsepower")
> pd_plot(bart_machine_cv, j = "stroke")
```

8.4.9 Incorporating Missing Data

The procedure for incorporating missing data was introduced in Section 8.3.2. We now build a `bartMachine` model using this procedure below:

```
> bart_machine = build_bart_machine(X, y, use_missing_data = TRUE,
  use_missing_data_dummies_as_covars = TRUE)
> bart_machine
Bart Machine v1.0b for regression
```

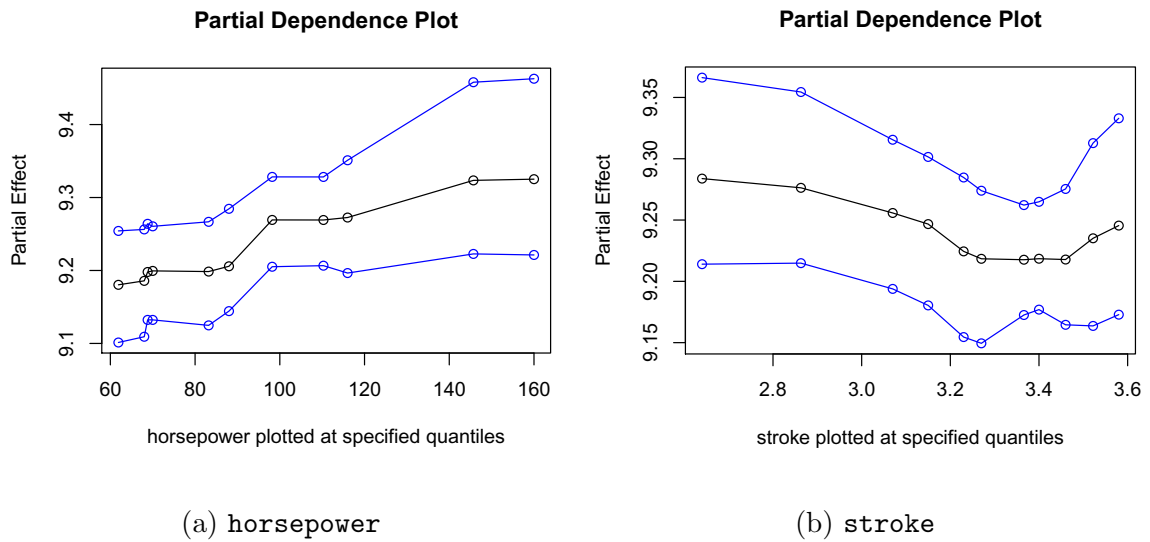


Figure 8.9: PDPs plotted in black and 95% credible intervals plotted in blue for variables in the automobile dataset. Points plotted are at the 5%ile, 10%ile, 20%ile, . . . , 90%ile and 95%ile of the values of the predictor. Lines plotted between the points approximate the PDP by linear interpolation.

Missing data feature ON

training data n = 201 and p = 50

built in 1.3 secs on 1 core, 50 trees, 250 burn-in and 1000 post. samples

sig² est for y beforehand: 0.016

avg sig² estimate after burn-in: 0.01055

in-sample statistics:

L1 = 12.77

L2 = 1.28

rmse = 0.08

```
Pseudo-Rsq = 0.9746
p-val for shapiro-wilk test of normality of residuals: 0.6638
p-val for zero-mean noise: 0.95693
```

Note that we now use the complete data set including the 41 observations for which there were missing features. Also note that p has now increased from 46 to 51. The five new predictors are dummy variables which indicate missingness constructed from the predictors which exhibited missingness (due to the `use_missing_data_dummies_as_covars` parameter being set to true). These variables are important if missingness *itself* shifts the response, as was the case in models explored in Chapter 9 of this document (Kapelner and Bleich, 2013a). One way to test if this type of missingness is important is to run a covariate test (Section 8.4.7) on these new 5 covariates:

```
> cov_importance_test(bart_machine, covariates = c("M_normalized_losses",
  "M_bore", "M_stroke", "M_horsepower", "M_peak_rpm"))
BART test for importance of 5 covariates....p_val = 0.673
```

From the p -value, we cannot conclude that these variables have an effect on the response. This seems reasonable given that only 5 predictors featured missingness among only 40 observations. This suggests that it may be appropriate to shrink the model by leaving out these missing dummies, but still allow for missingness to be incorporated into the split rules:

```
> bart_machine = build_bart_machine(X, y, use_missing_data = TRUE)
```

The procedure of Section 8.3.2 also natively incorporates missing data during prediction. Missingness will yield larger credible intervals. In the example below, we suppose that the `curb_weight` and `symboling` values were unavailable for 20th automobile.

```

> x_star = X[20, ]
> calc_credible_intervals(bart_machine, x_star, ci_conf = 0.95)
      ci_lower_bd ci_upper_bd
[1,]    8.650093    8.824515
> x_star[c("curb_weight", "symboling")] = NA
> calc_credible_intervals(bart_machine, x_star, ci_conf = 0.95)
      ci_lower_bd ci_upper_bd
[1,]    8.622582    8.978313

```

8.4.10 Variable Selection

In this section we demonstrate the principled variable selection procedure introduced in Section 8.3.3. The following code will select variables based on the three thresholds and also displays the plot in Figure 8.10.⁶

```

> vs = var_selection_by_permute_response_three_methods(bart_machine,
      bottom_margin = 10, num_permute_samples = 10)
> vs$important_vars_local_names
"curb_weight" "city_mpg" "engine_size" "horsepower"
"length"      "width"      "num_cylinders" "body_style_convertible"
"wheel_base"  "peak_rpm" "highway_mpg"  "wheel_drive_fwd"
> vs$important_vars_global_max_names
"curb_weight" "city_mpg" "engine_size" "horsepower" "length"
> vs$important_vars_global_se_names
"curb_weight" "city_mpg" "engine_size" "horsepower" "length"

```

⁶By default, variable selection is performed individually on dummy variables for a factor. The variable selection procedures return the permutation distribution and an aggregation of the dummy variables' inclusion proportions can allow for variable selection to be performed on an entire factor.

```
"width"          "num_cylinders" "wheel_base"  "wheel_drive_fwd"
```

Usually, “Global Max” and “Global SE” perform similarly, as they are both more stringent in selection. However, in many situations it will not be clear to the data analyst which threshold is most appropriate. The “best” procedure can be chosen via cross-validation on out-of-sample RMSE as follows:

```
var_selection_by_permute_response_cv(bart_machine)
```

```
$best_method
```

```
[1] "important_vars_local_names"
```

```
$important_vars_cv
```

```
[1] "body_style_convertible" "city_mpg"          "curb_weight"
```

```
[4] "engine_size"           "engine_type_ohc"  "horsepower"
```

```
[7] "length"                "num_cylinders"   "peak_rpm"
```

```
[10] "wheel_base"           "wheel_drive_fwd"  "wheel_drive_rwd"
```

```
[13] "width"
```

On this dataset, the “best” approach (as defined by out-of-sample prediction error) is the “Local” procedure.

The following sections (8.4.11 and 8.4.12) demonstrate additional features using Friedman (1991)’s function:

$$\mathbf{y} = 10\sin(\pi\mathbf{x}_1\mathbf{x}_2) + 20(\mathbf{x}_3 - .5)^2 + 10\mathbf{x}_4 + 5\mathbf{x}_5 + \mathcal{E}, \quad \mathcal{E} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}). \quad (8.6)$$

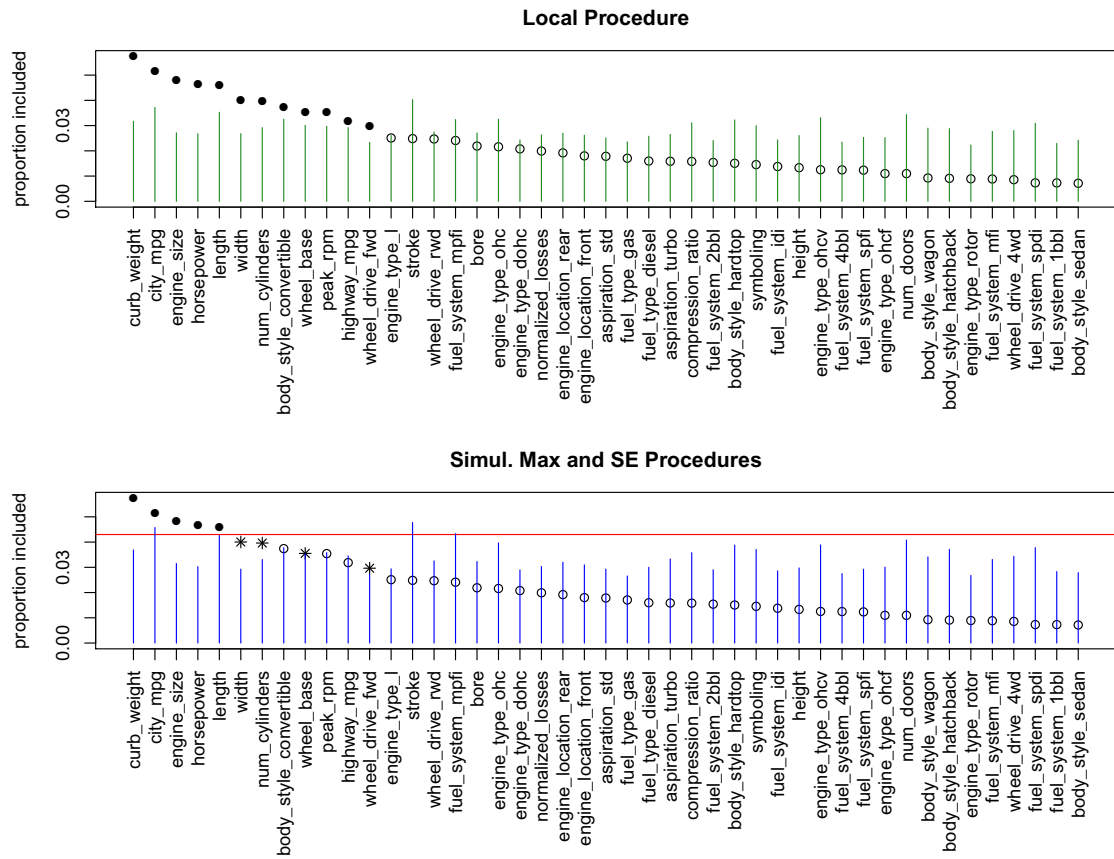


Figure 8.10: Visualization of the three variable selection procedures outlined in Section 8.3.3 with $\alpha = 0.05$. The top plot illustrates the “Local” procedure. The green lines are the threshold levels determined from the permutation distributions that must be exceeded for a variable to be selected. The plotted points are the variable inclusion proportions for the observed data (averaged over five duplicate `bartMachine` models). If the observed value is higher than the green bar, the variable is included and is displayed as a solid dot; if not, it is not included and it is displayed as an open dot. The bottom plot illustrates both the “Global SE” and “Global Max” thresholds. The red line is the cutoff for “Global Max” and variables pass this threshold are displayed as solid dots. The blue lines represent the thresholds for the “Global SE” procedure. Variables that exceed this cutoff but not the “Global Max” threshold are displayed as asterisks. Open dots exceed neither threshold.

8.4.11 Informed Prior Information on Covariates

Bleich et al. (2013) propose a method for incorporating informed prior information about the predictors into **BART**. This can be achieved by modifying the prior on the splitting rules as well as the corresponding calculations in the Metropolis-Hastings step. In particular, covariates believed to influence the response can be proposed more often as candidates for splitting rules. Useful prior information can aid in both variable selection and prediction tasks. We illustrate the impact of a correctly informed prior in the context of the Friedman function (Equation 8.6). We include the 5 predictors which influence the response as well as 95 that do not.

All that is required is a specification of relative weights for each predictor. These are converted internally to probabilities. We assign 5 times the weight to the 5 true covariates of the model relative to the 95 useless covariates.

```
> prior = c(rep(5, times = 5), rep(1, times = 95))
```

We now sample 500 observations from the Friedman function and construct a default `bartMachine` model as well as a `bartMachine` model with the informed prior and compare their performance on a test set of another 500 observations.

```
> bart_machine = build_bart_machine(X, y)
> bart_machine_informed = build_bart_machine(X, y, cov_prior_vec = prior)

> bart_predict_for_test_data(bart_machine, Xtest, ytest)$rmse
[1] 1.661159
> bart_predict_for_test_data(bart_machine_informed, Xtest, ytest)$rmse
[1] 1.232925
```

There is a substantial improvement in out-of-sample predictive performance when a properly informed prior is used.

Note that we recommend use of the prior vector to down-weight the indicator variables that result from dummifying factors so that the total set of dummy variables has the same weight as a continuous covariate.

8.4.12 Interaction Effect Detection

In Section 8.4.6, we explored using variable inclusion proportions to understand the relative influences of different covariates. A similar procedure can be carried out for examining interaction effects within a BART model. This question was initially explored in Damien et al. (2013) where an interaction was considered to exist between two variables if they both appeared in at least one splitting rule in a given tree. We refine the definition of an interaction as follows.

We first begin with a $p \times p$ matrix of zeroes. Within a given tree, for each split rule variable j , we look at all split rule variables of child nodes, k , and we increment the j, k element of the matrix. Hence variables are considered to interact in a given tree *only if* they appear together in a contiguous downward path from the root node to a terminal node. Note that a variable may interact with itself (when fitting a linear effect, for instance). Since there is no order between the parent and child, we then add the j, k counts together with the k, j counts (if $j \neq k$). Summing across trees and Gibbs samples gives the total number of interactions for each pair of variables from which relative importance can be assessed.

We demonstrate interaction detection on the Friedman function using 10 covariates using the code below:

```
> interaction_investigator(bart_machine, num_replicates_for_avg = 25,  
  num_var_plot = 10, bottom_margin = 5)
```

Shown in Figure 8.11 are the ten most important interactions in the model. The illustration is averaged over many model constructions to obtain stable estimates

across many posterior modes in the sum-of-trees distribution. Notice that the interaction between \mathbf{x}_1 and \mathbf{x}_2 dominates all other terms, as BART is correctly capturing the single true interaction effect in Equation 8.6. Choosing which of these interactions *significantly* affect the response is not addressed in this paper. The methods suggested in Section 8.3.3 may be applicable here and we consider this to be fruitful future work.

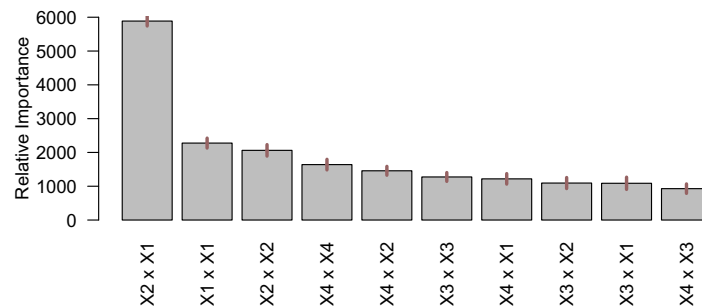


Figure 8.11: The top 10 average variable interaction counts (termed “relative importance”) in the default `bartMachine` model for the Friedman function data averaged over 25 model constructions. The segments atop the bars represent 95% confidence intervals.

8.5 Classification Features

In this section we highlight the features that differ from the regression case when the response is dichotomous. The illustrative dataset consists of 332 Pima Indians obtained from the UCI repository. Of the 332 subjects, 109 were diagnosed with diabetes, the binary response variable. There are seven continuous predictors which are body metrics such as blood pressure, glucose concentration, etc. and there is no missing data.

Building a `bartMachine` model for classification has the same computing parameters except that q, ν cannot be specified since there is no longer a prior on σ^2 (see Section 8.2.3). We first build a cross-validated model below.

```
> bart_machine_cv = build_bart_machine_cv(X, y)
... BART CV win: k: 3 m: 50
> bart_machine_cv
Bart Machine v1.0b for classification

training data n = 332 and p = 7
built in 0.5 secs on 4 cores, 50 trees, 250 burn-in and 1000 post. samples

confusion matrix:
```

	predicted No	predicted Yes	model errors
actual No	211.000	12.00	0.054
actual Yes	41.000	68.00	0.376
use errors	0.163	0.15	0.160

Classification models have an added hyperparameter, `prob_rule_class`, which is the rule for determining if the probability estimate is great enough to be classified into the positive category. We can see above that the `bartMachine` at times predicts “NO” for true “YES” outcomes and we suffer from a 37.6% error rate for this outcome. We can try to mitigate this error by lowering the threshold to increase the number of “YES” labels predicted:

```
> build_bart_machine(X, y, prob_rule_class = 0.3)
Bart Machine v1.0b for classification
```

training data n = 332 and p = 7

built in 0.5 secs on 4 cores, 50 trees, 250 burn-in and 1000 post. samples

confusion matrix:

	predicted No	predicted Yes	model errors
actual No	178.000	45.000	0.202
actual Yes	12.000	97.000	0.110
use errors	0.063	0.317	0.172

This lowers the model error to 11% for the “YES” class, but at the expense of increasing the error rate for the “NO” class. We encourage the user to cross-validate this rule based on the appropriate objective function for the problem at hand.

We can also check out-of-sample statistics:

```
> oos_stats = k_fold_cv(X, y, k_folds = 10)
```

```
> oos_stats$confusion_matrix
```

	predicted No	predicted Yes	model errors
actual No	203.000	20.000	0.090
actual Yes	47.000	62.000	0.431
use errors	0.188	0.244	0.202

Note that it is possible to predict both class labels and probability estimates for given observations:

```
> predict(bart_machine_cv, X[1 : 2, ], type = "prob")
```

```
[1] 0.6253160 0.1055975
```

```
> predict(bart_machine_cv, X[1 : 2, ], type = "class")
```

```
[1] "Yes" "No"
```

When using the covariate tests of Section 8.4.7, total misclassification error becomes the statistic of interest instead of Pseudo- R^2 . The p value is calculated now as the proportion of null samples with *higher* misclassification error. Figure 8.12 illustrates the test showing that predictor `age` seems to matter in the prediction of `Diabetes`, controlling for other predictors.

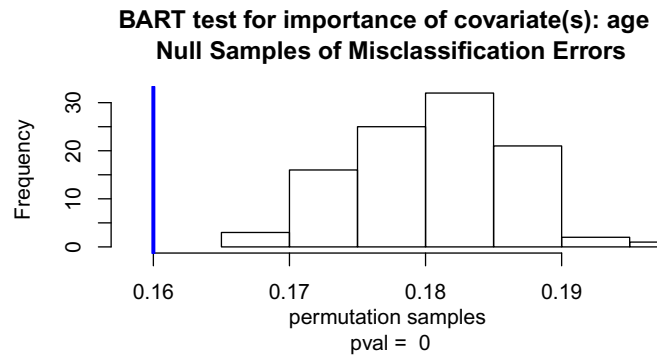


Figure 8.12: Test of covariate importance for predictor `age` on whether or not the subject will contract `Diabetes`.

The partial dependence plots of Section 8.4.8 are now scaled as probit of the probability estimate. Figure 8.13 illustrates that as glucose increases, the probability of contracting `Diabetes` increases linearly on a probit scale.

Credible intervals are implemented for classification `bartMachine` and are displayed on the probit scale. Note that the prediction intervals of Section 8.4.5 do not exist for classification.

```
> calc_credible_intervals(bart_machine_cv, X[1 : 2, ])
      ci_lower_bd ci_upper_bd
[1,]  0.34865355  0.8406097
[2,]  0.01686486  0.2673171
```

Other functions work similarly to regression except those that plot the responses and those that explicitly depend on RMSE as an error metric.

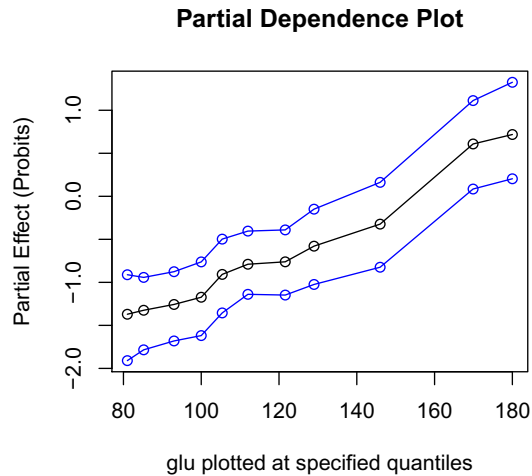


Figure 8.13: PDP for predictor `glu`. The blue lines are 95% credible intervals.

8.6 Discussion

This article introduced `bartMachine`, a new R package which implements Bayesian Additive Regression Trees. The goal of this package is to provide a fast, extensive and user-friendly implementation accessible to a wide range of data analysts, and increase the visibility of BART to a broader statistical audience. We hope we have provided organized, well-documented open-source code and we encourage the community to make innovations on this package.

Replication

The code for `bartMachine` is located at <http://github.com/kapelner/bartMachine> under the MIT license. Results, tables, and figures found in this paper can be replicated via the scripts located in the `bart_package_paper` folder within this git repository.

Acknowledgements

We thank Richard Berk, Andreas Buja, Zachary Cohen, Ed George, Alex Goldstein, Shane Jensen, Abba Krieger, and Robert DeRubeis for helpful discussions. We thank Simon Urbanek for his generous help with rJava.

Incorporating Missingness into BART*

Abstract

We present a method for incorporating missing data into general forecasting problems which use non-parametric statistical learning. We focus on a tree-based method, Bayesian Additive Regression Trees (BART), enhanced with “Missingness Incorporated in Attributes,” an approach recently proposed for incorporating missingness into decision trees. This procedure extends the native partitioning mechanisms found in tree-based models and does not require imputation. Simulations on generated models and real data indicate that our procedure offers promise for both selection model and pattern mixture frameworks as measured by out-of-sample predictive accuracy. We also illustrate BART’s abilities to incorporate missingness into uncertainty intervals. Our implementation is readily available in the R package `bartMachine`.

*Joint work with Justin Bleich

9.1 Introduction

This article addresses prediction problems where covariate information is missing during model construction and is also missing in future observations for which we are obligated to generate a forecast. Our aim is to innovate a non-parametric statistical learning extension which incorporates missingness into *both* the training and the forecasting phases. In the spirit of non-parametric learning, we wish to incorporate the missingness in both phases automatically, without the need for pre-specified modeling.

We limit our focus to tree-based statistical learning, which has demonstrated strong predictive performance and has consequently received considerable attention in recent years. State-of-the-art algorithms include Random Forests (RF, Breiman, 2001a), stochastic gradient boosting (Friedman, 2002), and Bayesian Additive and Regression Trees (BART, Chipman et al., 2010), the algorithm of interest in this study. Popular implementations of these methods do not incorporate covariate missingness *natively* without relying on either imputation or a complete case analysis of observations with no missing information.

Previous simulations and real data set applications have indicated that BART is capable of achieving excellent predictive accuracy. Unlike most competing techniques, BART is composed of a probability model, rather than a procedure that is purely “algorithmic” (Breiman, 2001b). BART presents an alternative approach to statistical learning for those comfortable with the Bayesian framework. This framework provides certain advantages, such as built-in estimates of uncertainty in the form of credible intervals as well as the ability to incorporate prior information on covariates (Bleich et al., 2013). However, no means for incorporating missing data in BART has been published to date. Our goal here is to develop a principled way of adapting BART’s machinery to incorporate missing data that takes advantage of the Bayesian

framework.

Our proposed method, henceforth named **BARTm**, modifies the recursive partitioning scheme during construction of the decision trees to incorporate missing data into splitting rules. By relying on the Metropolis-Hastings algorithm embedded in **BART**, our method attempts to send missing data to whichever of the two daughter nodes increases overall model likelihood. Missingness *itself* also becomes a valid splitting criterion.

During model construction, taking advantage of this modified set of splitting rules does not require imputation, which relies on assumptions that cannot be easily verified. Our approach is equally viable for continuous and nominal covariate data and both selection and pattern mixture models. The latter models do not assume that missing data is necessarily free of information; the data may have gone missing for a reason crucial to the response function and therefore crucial to our forecast. **BARTm** is able to exploit this relationship when appropriate.

Since missingness is handled natively within the algorithm, **BARTm** can generate predictions on future data with missing entries as well. Additionally, **BART**'s Bayesian framework also naturally provides estimates of uncertainty in the form of credible intervals. The amount of uncertainty increases with the amount of information lost due to missingness; thereby missingness is appropriately incorporated into the standard error of the prediction. Also, our proposed procedure has negligible impact on the runtime during both model construction and prediction phases.

In Sections 9.2.1 - 9.2.3, we provide a framework for statistical learning with missingness with a focus on decision trees. We give a brief overview of the **BART** algorithm in Section 9.2.4 and explain the internals of **BARTm** in Section 9.3. We then demonstrate **BARTm**'s predictive performance on generated models in Section 9.4 as well as real data with a variety of missing data scenarios in Section 9.5. We conclude with Section 9.6. **BARTm** can be found in the R package `bartMachine` which is available

on CRAN (Kapelner and Bleich, 2013a).

9.2 Background

9.2.1 A Framework for Missing Data in Statistical Learning

Consider p covariates $\mathbf{X} := [X_1, \dots, X_p]$, a continuous response \mathbf{Y} and an unknown function f where $\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon}$. When \mathbf{Y} is binary, we use a similar framework with an appropriate link function encapsulating f . We denote $\boldsymbol{\varepsilon}$ as the noise in the response unexplained by f . The goal of statistical learning is to use the *training set*, $[\mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}}]$ which consists of n observations drawn from the population $\mathbb{P}(\mathbf{Y}, \mathbf{X})$, to produce an estimate, \hat{f} , the best guess of $\mathbb{E}[\mathbf{Y} | \mathbf{X}]$. This function estimate can then be used to generate predictions on future test observations with an unknown response. We denote these future observations as \mathbf{X}_* which we assume are likewise drawn from the same population as the training set.

Missingness is one of the scourges of data analysis, plaguing statistical learning by causing missing entries in both the training matrix $\mathbf{X}_{\text{train}}$ as well as missing entries in the future records, \mathbf{X}_* . In the statistical learning context, the training set is defined by observations which do not exhibit missingness in their response, $\mathbf{y}_{\text{train}}$. Records with missing responses cannot be used to build models for estimation of f . “Imputing missing values in the response” for the new \mathbf{X}_* is equivalent to “prediction” and is the primary goal of statistical learning. Thus, “missingness” considered in this paper is missingness *only* in $\mathbf{X}_{\text{train}}$ and \mathbf{X}_* . We denote missingness in the $p_M \leq p$ features of \mathbf{X} which suffer from missingness as $\mathbf{M} := [M_1, \dots, M_{p_M}]$, binary vectors where 1 indicates missing and 0 indicates present, and covariates that are present with $\mathbf{X}_{\text{obs}} := [X_{\text{obs}_1}, \dots, X_{\text{obs}_p}]$. The main goal of statistical learning with missingness is to estimate $\mathbb{E}[\mathbf{Y} | \mathbf{X}_{\text{obs}}, \mathbf{M}]$.

We now frame missing data models in statistical learning using the canonical framework of selection and pattern-mixture models (Little, 1993). Conditional on \mathbf{X} , *selection models* factor the full data likelihood as

$$\mathbb{P}(\mathbf{Y}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \mathbb{P}(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) \mathbb{P}(\mathbf{M} \mid \mathbf{X}, \boldsymbol{\gamma}) \quad (9.1)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are parameter vectors and are assumed distinct. The first term on the right hand side reflects that the marginal likelihood for the response $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta})$ is independent of missingness. The second term on the right conventionally conditions on \mathbf{Y} . In the forecasting paradigm, missingness is *assumed independent* of the response because \mathbf{Y} is often yet to be realized and thus its unknown value should not influence \mathbf{M} , the missingness of the previously realized covariates.

Conditional on \mathbf{X} , *pattern mixture models* partition the full data likelihood as

$$\mathbb{P}(\mathbf{Y}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \mathbb{P}(\mathbf{Y} \mid \mathbf{M}, \mathbf{X}, \boldsymbol{\theta}) \mathbb{P}(\mathbf{M} \mid \mathbf{X}, \boldsymbol{\gamma}). \quad (9.2)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are parameter vectors and again assumed distinct. The difference between the above and Equation 9.1 is the marginal likelihood of the response is now a function of \mathbf{M} . This means there can be different response models under different patterns of missingness in the p_M covariates.

In both selection and pattern-mixture paradigms, the term on the right is the *missing data mechanism* (MDM), which traditionally is the mechanism controlling missingness in the response. In our framework however, the MDM controls missingness only in \mathbf{X} : the covariates (and parameters $\boldsymbol{\gamma}$) create missingness within themselves which inevitably needs to be incorporated during model construction and forecasting. Thus, the MDM is conceptually equivalent in both the selection and pattern mixture

paradigms.

The conceptual difference between the selection and pattern mixture models in the statistical learning framework can be envisioned as follows. Imagine the full covariates \mathbf{X} are realized but due to the MDM, \mathbf{X} is latent and we instead observe \mathbf{X}_{obs} and \mathbf{M} . In the selection paradigm, \mathbf{Y} is realized only from the full covariates via $\mathbb{P}(\mathbf{Y} | \mathbf{X}, \boldsymbol{\theta})$. However, in the pattern-mixture paradigm, both \mathbf{X} and \mathbf{M} intermix to create many collated response models $\{\mathbb{P}(\mathbf{Y} | \mathbf{X}, \boldsymbol{\theta}, \mathbf{M} = m)\}_{m \in \mathcal{M}}$ corresponding to different points in \mathbf{M} -space. Thus, under our assumptions, selection models are a subset of pattern mixture models. Note that pattern-mixture models are chronically under-identified and difficult to explicitly model in practice. We address why our proposed method is well-suited to handle prediction problems under this framework in Section 9.3.

We now present Little and Rubin (2002)'s taxonomy of MDM's which are traditionally framed in the selection model paradigm but here apply to both paradigms: (1) missing completely at random (MCAR), (2) missing at random (MAR) and (3) not missing at random (NMAR). MCAR is a mechanism that generates missingness in covariate j without regard to the value of X_j itself nor the values and missingness of any other covariates, denoted \mathbf{X}_{-j} . MCAR is exclusively determined by exogenous parameter(s) $\boldsymbol{\gamma}$. The MAR mechanism generates missingness without regard to X_j , its own value, but can depend on values of other attributes \mathbf{X}_{-j} as well as $\boldsymbol{\gamma}$. The NMAR mechanism features the additional dependence on the value of X_j itself as well as unobserved covariates (note that explicit dependence on unobserved covariates was not explored as MDM's in this paper). We summarize these mechanisms in Table 9.1. In our framework, each of the $p_M \leq p$ covariates with missingness, denoted as X_j 's, are assumed to have their own MDM $_j$. Thus, the full MDM for the whole covariate space, $\mathbb{P}(\mathbf{M} | \mathbf{X}, \boldsymbol{\gamma})$, can be arbitrarily convoluted, exhibiting combinations of MCAR, MAR and NMAR among its p_M covariates and each MDM $_j$ relationship

may be highly non-linear with complicated interactions.

MDM	$\mathbb{P}(\mathbf{M}_j \mid X_{j,\text{miss}}, \mathbf{X}_{-j,\text{miss}}, \mathbf{X}_{-j,\text{obs}}, \boldsymbol{\gamma}) = \dots$
MCAR	$\mathbb{P}(\mathbf{M}_j \mid \boldsymbol{\gamma})$
MAR	$\mathbb{P}(\mathbf{M}_j \mid \mathbf{X}_{-j,\text{miss}}, \mathbf{X}_{-j,\text{obs}}, \boldsymbol{\gamma})$
NMAR	(does not simplify)

Table 9.1: MDM models in the context of statistical learning. \mathbf{M}_j is an indicator vector which takes the value one when the j^{th} covariate is missing for the i th observation. $\mathbf{X}_{-j,\text{obs}}$ are the observed values of the other covariates, besides j . $\mathbf{X}_{-j,\text{miss}}$ are the values of the other covariates, besides j , which are not observed because they are missing.

We conclude this section by emphasizing that in the non-parametric statistical learning framework where predictive performance is the objective, there is no need for explicit inference of $\boldsymbol{\theta}$ (which may have unknown structure and arbitrary, possibly infinite, dimension). Instead, the algorithm performs “black-box” estimation of the data generating process such that the output \hat{f} estimates the $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}_{\text{obs}}, \mathbf{M}]$ function. Thus, if we can successfully estimate this conditional expectation function directly, then accurate forecasts can be obtained. This is the approach that BARTm takes.

9.2.2 Strategies for Incorporating Missing Data

A simple strategy for incorporating missingness into model building is to simply ignore the observations in $\mathbf{X}_{\text{train}}$ that contain at least one missing measurement. This is called “list-wise deletion” or “complete case analysis.” It is well known that complete case analysis will be unbiased for MCAR and MAR selection models where missingness does not depend on the response when the target of estimation is $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$.

However, when forecasting, the data analyst must additionally be guaranteed that \mathbf{X}_* has no missing observations, since it is not possible to generate forecasts for these cases.

By properly modeling missingness, incomplete cases can be used and more information about $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$ becomes available, potentially yielding higher predictive performance. One popular strategy is to guess or “impute” the missing entries. These guesses are then used to “fill in” the holes in $\mathbf{X}_{\text{train}}$ and \mathbf{X}_* . The imputed $\mathbf{X}_{\text{train}}$ is then used *as if* it were the real covariate data when constructing \hat{f} and the imputed \mathbf{X}_* is then used as if it were the real covariate data during forecasting. To carry out imputation, the recommended strategy is to properly model the predictive distribution and then draws from the model are used to fill in the missing entries. *Multiple imputation* involves imputing many times and averaging over the results from each imputation (Rubin, 1978). In statistical learning, a prediction could be calculated by averaging the predictions from many \hat{f} 's built from many imputed $\mathbf{X}_{\text{train}}$'s and then further averaging over many imputed \mathbf{X}_* 's. In practice, having knowledge of both the missing data mechanism and each probability model is very difficult and has usually given way to nonparametric methods such as k -nearest neighbors (Troyanskaya et al., 2001) for continuous covariates and saturated multinomial modeling (Schafer, 1997) for categorical covariates. The widely used R package `randomForest` (Liaw and Wiener, 2002) imputes via “hot-decking” (Little and Rubin, 2002).

A more recent approach, `MissForest` (Stekhoven and Bühlmann, 2012), fits non-parametric imputation models for any combination of continuous and categorical input data, even when the response is unobserved. In this unsupervised procedure (i.e., no response variable needed), initial guesses for the imputed values are made. Then, for each attribute with missingness, the observed values of that attribute are treated as the response and a RF model is fit using the remaining attributes as predictors. Predictions for the missing values are made via the trained RF and serve as

updated imputations. The process proceeds iteratively through each attribute with missingness and then repeats until a stopping criterion is achieved. The authors argue that their procedure intrinsically constitutes multiple imputation due to Random Forest’s averaging over many unpruned decision trees. The authors also state that their method will perform particularly well when “the data include complex interactions or non-linear relations between variables of unequal scales and different type.” Although no explicit reference is given to Little and Rubin (2002)’s taxonomy in their work, we expect `MissForest` to perform well in situations generally well-suited for imputation, namely, the MCAR and MAR selection models discussed in Section 9.2.1. `MissForest` would not be suited for NMAR MDMs as imputation values for X_j can only be modeled from \mathbf{X}_{-j} in their implementation. Additionally, implementing `MissForest` would not be recommended for pattern-mixture scenarios because imputation is insufficient to capture differing response patterns.

Since BART is composed primarily of a sum-of-regression-trees model, we now review strategies for incorporating missing data in tree-based models.

9.2.3 Missing data in Binary Decision Trees

Binary decision trees are composed of a set of connecting nodes. *Internal nodes* contain a *splitting rule*, for example, $\mathbf{x}_j < \mathbf{c}$, where \mathbf{x}_j is the *splitting attribute* and \mathbf{c} is the *splitting value*. An observation that satisfies the rule is passed to the left daughter node otherwise it is passed to the right daughter node. This partitioning proceeds until an observation reaches a *terminal node*. Terminal nodes (also known as *leaves*) do not have splitting rules and instead have *leaf values*. When an observation “lands” in a terminal node it is assigned the leaf value of the terminal node in which it has landed. In *regression trees*, this leaf value is a real number, and is the estimate of the response y for the given covariates. Thus, regression trees are a nonparametric

fitting procedure where the estimate of f is a partition of predictor space into various hyperrectangles. Regression trees are well-known for their ability to approximate complicated response surfaces containing both nonlinearities and interaction effects.

There are many different ways to build decision trees. Many classic approaches rely on a greedy procedure to choose the best splitting rule at each node based on some pre-determined criterion. Once the construction of the tree is completed, the tree is then pruned back to prevent overfitting.

Previous efforts to handle missingness in trees include surrogate variable splitting (Therneau and Atkinson, 1997), “Missing Incorporated in Attributes” (MIA, Twala et al., 2008, section 2) and many others (see Ding and Simonoff, 2010 and Twala, 2009). MIA, the particular focus for this work, is a procedure that natively uses missingness when greedily constructing the rules for the decision tree’s internal nodes. We summarize the procedure in Algorithm 3 and we explain how the expanded set of rules is injected into the BART procedure in Section 9.3.

Algorithm 3 *Splitting rule choices during construction of a new tree branch in MIA.*

The algorithm chooses one of the following three rules for all splitting attributes and all splitting values c . Since there are p splitting attributes and at most $n - 1$ unique values to split on, the greedy splitting algorithm with MIA checks $2(n - 1)p + p$ possible splitting rules at each iteration instead of the classic $(n - 1)p$.

- 1: If x_{ij} is present and $x_{ij} \leq c$, send this observation left (\leftarrow); otherwise, send this observation right (\rightarrow). If x_{ij} is missing, send this observation left (\leftarrow).
 - 2: If x_{ij} is present and $x_{ij} \leq c$, send this observation left (\leftarrow); otherwise, send this observation right (\rightarrow). If x_{ij} is missing, send this observation right (\rightarrow).
 - 3: If x_{ij} is missing, send this observation left (\leftarrow); if it is present, regardless of its value, send this observation right (\rightarrow).
-

There are many advantages of the MIA approach. First, MIA has the ability to

model complex MAR and NMAR relationships, as evidenced in both Twala et al. (2008) and the results of Sections 9.4 and 9.5. Since missingness is integrated into the splitting rules, forecasts can be made without imputing when \mathbf{X}_* contains missingness.

Another strong advantage of MIA is the ability to split on feature missingness (line 3 of Algorithm 3). This splitting rule choice allows for the tree to better capture pattern mixture models where missingness directly influences the response model. Generally speaking, imputation ignores pattern mixture models; missingness is only viewed as holes to be filled-in and forgotten.

Due to these benefits as well as conceptual simplicity, we chose to implement MIA-within-BART, denoted “BARTm”, when enhancing BART to handle missing data.

9.2.4 BART

Bayesian Additive Regression Trees is a combination of many regression trees estimated via a Bayesian model. Imagine the true response function can be approximated by the sum of m trees with additive normal and homoskedastic noise:

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon} \approx \mathfrak{T}_1^{\text{leaf}}(\mathbf{X}) + \mathfrak{T}_2^{\text{leaf}}(\mathbf{X}) + \dots + \mathfrak{T}_m^{\text{leaf}}(\mathbf{X}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n). \quad (9.3)$$

The notation, $\mathfrak{T}^{\text{leaf}}$, denotes both structure and splitting rules (\mathfrak{T}) as well as leaf values (leaf). Note that BART can be adapted for classification problems by using a probit link function and a data augmentation approach relying on latent variables (Albert and Chib, 1993).

BART can be distinguished from other purely algorithmic ensemble-of-trees models by its full Bayesian model, consisting of both a set of independent priors and likelihoods. Its posterior distribution is estimated via Gibbs Sampling (Geman and

Geman, 1984) with a Metropolis-Hastings step (Hastings, 1970).

There are three regularizing priors within the BART model which are designed to prevent overfitting. The first prior, placed on the tree structure is designed to prevent trees from growing too deep, thereby limiting the complexity that can be captured by a single tree. The second prior is placed on the leaf value parameters (the predicted values found in the terminal nodes) and is designed to shrink the leaf values towards the overall center of the response's empirical distribution. The third prior is placed on the variance of the noise σ^2 and is designed to curtail overfitting by introducing noise into the model if it begins to fit too snugly. Our development of BARTm uses the default hyperparameters recommended in the original work (Chipman et al., 2010). For those who do not favor a pure Bayesian approach, these priors can be thought of as tuning parameters.

In addition to the regularization priors, BART imposes an agnostic prior on the splitting rules within the decision tree branches. First, for a given branch, the splitting attribute is uniformly drawn from the set $\{x_1, \dots, x_p\}$ of variables available at the branch. The splitting value is then selected by drawing uniformly from the available values conditional on the splitting attribute j . Selecting attributes and values from a uniform discrete distribution represents a digression from the approach used in decision tree algorithms of greedily choosing splits based on some splitting criteria. Extending this prior allows for BART to incorporate MIA, which is discussed in Section 9.3.

To generate draws from the posterior distribution, each tree is fit iteratively, holding the other $m - 1$ trees constant, by using only the portion of the response left unfitted. To sample trees, changes to the tree structure are proposed then accepted or rejected via a Metropolis-Hastings step. The tree proposals are equally-likely alterations: growing a leaf by adding two daughter nodes, pruning two twin leaves (rendering their parent node into a leaf), or changing a splitting rule. Following the

tree sampling, the posterior for the leaf value parameters are Gibbs sampled. The likelihood of the predictions in each node is assumed to be normal. Therefore, the normal-normal conjugacy yields the canonical posterior normal distribution. After sampling all tree changes and terminal node parameters, the variance parameter σ^2 is Gibbs sampled. By model assumption, the likelihood for the errors is normal and the conjugacy with the inverse-gamma prior yields the canonical posterior inverse-gamma.

Usually around 1,000 Metropolis-within-Gibbs iterations are run as “burn-in” until σ^2 converges (by visual inspection). Another 1,000 or so are sampled to obtain “burned-in” draws from the posterior, which define the BART model. Forecasts are then obtained by dropping the observations of \mathbf{X}_* down the collection of sampled trees within each burned-in Gibbs sample. A point prediction \hat{y} is generated by summing the posterior leaf values across the trees as in Equation A.5. *Credible intervals*, which are intervals containing a desired percentage (e.g. 95%) of the posterior probability mass for a Bayesian parameter of interest, can be computed via the desired empirical quantiles over the burned-in samples.

For a thorough description about the internals of BART see Chipman et al. (2010) and Kapelner and Bleich (2013a).

9.3 Missing Incorporated in Attributes within BART

Implementing BARTm is straightforward. We previously described the prior on the splitting rules within the decision tree branches as being discrete uniform on the possible splitting attributes and discrete uniform on the possible splitting values. To account for Lines 1 and 2 in the MIA procedure (Algorithm 3), the splitting attribute x_j and split value are proposed as explained in Section 9.2.4, but now we additionally propose a direction (left or right with equal probability) for records to be sent when the records have with missing values in x_j . A possible splitting rule would therefore

be “ $x_{ij} < c$ and move left if x_{ij} is missing.” To account for Line 3 in the algorithm, splitting on missingness itself, we create dummy vectors of length n for each of the p_M attributes with missingness, denoted $\mathbf{M}_1, \dots, \mathbf{M}_{p_M}$, which assume the value 1 when the entry is missing and 0 when the entry is present. We then augment the original training matrix together with these dummies and use the augmented training matrix, $\mathbf{X}'_{\text{train}} := [\mathbf{X}_{\text{train}}, M_1, \dots, M_{p_M}]$, as the training data in the BARTm algorithm. Once again, the prior on splitting rules is the same as the original BART but now with the additional consideration that the direction of missingness is equally likely left or right conditional on the splitting attribute and value. But why should this algorithm yield good predictive performance under the framework discussed in Section 9.2.1?

We expect BARTm to exhibit greater predictive performance over MIA in classical decision trees for two reasons. First, BARTm’s sum-of-trees model offers much greater fitting flexibility of interactions and non-linearities compared to a single tree; this flexibility will explore models that the data analyst may not have thought of. Additionally, due to the greedy nature of decision trees, once a split is chosen, the direction in which missingness is sent cannot be reversed. BARTm can alter its trees by pruning and regrowing nodes or changing splitting rules. These proposed modifications to the trees are accepted or rejected stochastically using the Metropolis-Hastings machinery depending on how strongly the proposed move increases the model’s likelihood.

We hypothesize that BARTm’s stochastic search for splitting rules allows observations with missingness to be grouped with observations having similar response values. Due to the Metropolis-Hastings step, only splitting rules and corresponding groupings that increase overall model likelihood $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}, \mathbf{M})$ will be accepted. In essence, BARTm is “feeling around” predictor space for a location where the missing data would most increase the overall marginal likelihood. For selection models, since splitting rules can depend on any covariate including the covariate with missing data, it should be possible to generate successful groupings for the missing data under both

MAR and NMAR mechanisms.

As a simple MAR example, suppose there are two covariates X_1 and X_2 and a MAR mechanism where X_2 is increasingly likely to go missing for large values of X_1 . BARTm can partition this data in two steps to increase overall likelihood: (1) A split on a large value of X_1 and then (2) a split on M_2 . As a simple NMAR example, suppose a mechanism where X_2 is more likely to be missing for large values of X_2 . BARTm can select splits of the form “ $x_2 > c$ and x_2 is missing” with c large. Here, the missing data is appropriately kept with larger values of X_2 and overall likelihood should be increased.

When missingness does not depend on any other covariates, it should be more difficult to find appropriate ways to partition the missing data, and we hypothesize that BARTm will be least effective for selection models with MCAR MDMs. We believe this is due to the regularization prior on the depths of the trees coupled with the fact that all missing data must move to the same daughter node. Thus, the trees do not grow deep enough to create sufficiently complex partitioning schemes to handle the MCAR mechanism. Additionally, we hypothesize that BARTm has potential to perform well on pattern mixture models due to the partitioning nature of the regression tree. BARTm can partition the data based on different patterns of missingness by creating splits based on missingness itself. Then, underneath these splits, different submodels for the different patterns can be constructed. If missingness is related to the response, there is a good chance BARTm will find it and exploit it, yielding accurate forecasts.

Another motivation for adapting MIA to BART arises from computational concerns. BART is a computationally intensive algorithm, but its runtime increases negligibly in the number of covariates (see Chipman et al., 2010, Section 6). Hence, BARTm leaves the computational time virtually unchanged with the addition of the p_M new missingness dummy covariates. Another possible strategy would be to develop an iterative imputation procedure using BART similar to that in Stekhoven and Bühlmann

(2012) or a model averaging procedure using a multiple imputation framework, but we believe these approaches would be substantially more computationally intensive.

9.4 Generated Data Simulations

9.4.1 A Simple Pattern Mixture Model

We begin with an illustration of **BARTm**'s ability to directly estimate $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}_{\text{obs}}, \mathbf{M}]$ and additionally provide uncertainty intervals. We consider the following nonlinear response surface:

$$\mathbf{Y} = g(X_1, X_2, X_3) + \mathbf{B}M_3 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_e^2), \quad B \stackrel{iid}{\sim} \mathcal{N}(\mu_b, \sigma_b^2), \quad (9.4)$$

$$g(X_1, X_2, X_3) = X_1 + X_2 + 2X_3 - X_1^2 + X_2^2 + X_1X_2$$

$$[X_1, X_2, X_3] \stackrel{iid}{\sim} \mathcal{N}_3 \left(\mathbf{0}, \sigma_x^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{bmatrix} \right),$$

where $\sigma_x^2 = 1$, $\rho_1 = 0.2$, $\rho_2 = 0.4$, $\sigma_e^2 = 1$, $\mu_b = 10$ and $\sigma_b^2 = 0.5$. Note that the pattern mixture model is induced by missingness in X_3 . Under this missingness pattern, the response is offset by B , a draw from a normal distribution. Figure 9.1a displays the $n = 500$ sample of the response from the model colored by M_3 to illustrate the separation of the two response patterns. We choose the following jointly NMAR MDM for X_2 and X_3 which were chosen to be simple for the sake of ensuring that the illustration is clear. The next section features more realistic mechanisms.

$$1 : X_2 \text{ is missing with probability } 0.3 \text{ if } X_2 \geq 0 \quad (9.5)$$

2 : X_3 is missing with probability 0.3 if $X_1 \geq 0$.

If the BARTm model assumptions hold and is successfully able to estimate the conditional expectation function $\mathbb{E}[Y | \mathbf{X}_{\text{obs}}, \mathbf{M}]$, then the true $\mathbb{E}[Y | \mathbf{X}_{\text{obs}}, \mathbf{M}]$ is highly likely to be contained within a 95% credible interval for the prediction. We first check to see whether BARTm can capture the correct response when $\mathbf{X}_{\text{train}}$ has missing entries but \mathbf{X}_* does not. Predicting on $\mathbf{x}_* = [0 \ 0 \ 0]$ should give $\mathbb{E}[Y | \mathbf{X} = \mathbf{x}_*] = 0$ for the prediction. Figure 9.1b illustrates that BARTm captures the expected value within its 95% credible interval.

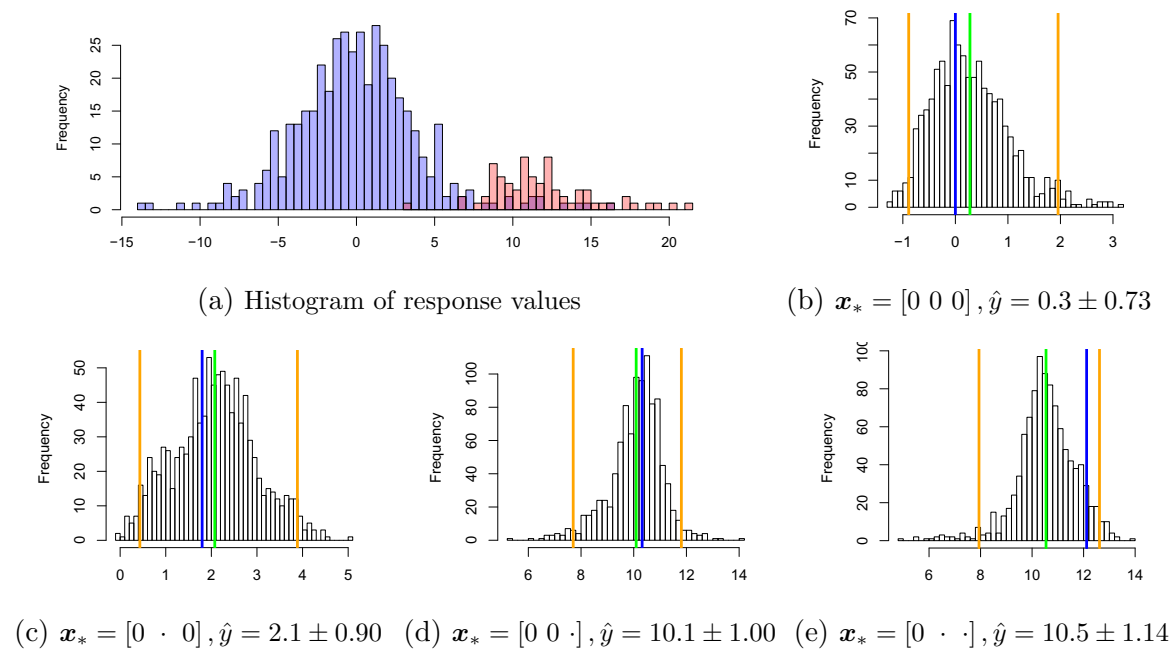


Figure 9.1: (a) A $n = 500$ sample of the responses of the model in Equation 9.4. Colored in blue are the responses when X_3 is present and red are responses when X_3 is missing. (b-e) 1,000 burned-in posterior draws from a BARTm model for different values of \mathbf{x}_* drawn from the data generating process found in Equation 9.4. The green line is BARTm's forecast \hat{y} (the average of the posterior burned-in samples). The blue line is the true model expectation. The two yellow lines are the bounds of the 95% credible interval for $\mathbb{E}[Y | \mathbf{X}_{\text{obs}} = \mathbf{x}_*, \mathbf{M} = \mathbf{m}_*]$.

Next we explore how well BARTm estimates the conditional expectation when missingness occurs within the new observation \mathbf{x}_* . We examine how BARTm handles missingness in attribute X_2 by predicting on $\mathbf{x}_* = [0 \cdot 0]$ where the “.” denotes missingness. By Equation 9.5, X_2 is missing 30% of the time if X_2 itself is greater than 0. By evaluating the moments of a truncated normal distribution, it follows that BARTm should guess $\mathbb{E}[X_2 + X_2^2 \mid X_2 > 0] = \sqrt{2/\pi} + 1 \approx 1.80$. Figure 9.1c indicates that BARTm’s credible interval captures this expected value. Note the larger variance of the posterior distribution relative to Figure 9.1b reflecting the higher uncertainty due to x_{*2} going missing. This larger interval is a great benefit of MIA. As the trees are built during the Gibbs sampling, the splitting rules on X_2 are accompanied by a protocol for missingness: missing data will flow left or right in the direction that increases model likelihood and this direction is chosen with randomness. Thus, when \mathbf{x}_* is predicted with x_{*2} missing, missing records flow left and right over the many burned-in Gibbs samples creating a wider distribution of predicted values, and thus a wider credible interval. This is an important point — BARTm can give a rough estimate of how much information is lost when values in new records become missing by looking at the change in the standard error of a predicted value. Note that if BART’s hyperparameters are considered “tuning parameters,” the credible intervals’ endpoints are not interpretable. However, the relative lengths of the intervals still signify different levels of forecast confidence to the practitioner.

We next consider how BARTm performs when X_3 is missing by predicting on $\mathbf{x}_* = [0 \ 0 \cdot]$. By Equation 9.5, BARTm should guess $\mathbb{E}[X_3 \mid X_1 > 0] = .4\sqrt{2/\pi} \approx .32$ (which follows directly from the properties of the conditional distribution of bivariate normal distribution, recalling that $\text{Corr}[X_1, X_3] = 0.4$). When X_3 is missing, there is a different response pattern, and the response is shifted up by B . Since $\mathbb{E}[B] = 10$, BARTm should predict approximately 10.32. The credible interval found in Figure 9.1d indicates that BARTm’s credible interval again covers the conditional expectation.

Finally, we consider the case where X_2 and X_3 are simultaneously missing. Predicting on $\mathbf{x}_* = [0 \cdot \cdot]$ has a conditional expectation of $\mathbb{E}[X_2 + X_2^2 \mid X_2 > 0] + \mathbb{E}[X_3 \mid X_1 > 0] + \mathbb{E}[B] \approx 12.12$. Once again, the posterior draws displayed in Figure 9.1e indicate that BARTm reasonably estimates the conditional expectation. Note that the credible interval here is wider than in Figure 9.1d due to the additional missingness of X_2 .

9.4.2 Selection Model Performance

In order to gauge BARTm's out-of-sample predictive performance on selection models and to evaluate the improvement over model-building on complete cases, we construct the same model as Equation 9.4 withholding the offset B (which previously induced the pattern mixture). Thus $\mathbf{Y} = g(X_1, X_2, X_3) + \boldsymbol{\mathcal{E}}$. We imposed three scenarios illustrating performance under the following missingness mechanisms. The first is MCAR; X_1 is missing with probability γ . The second is MAR; X_3 is missing according to a non-linear probit model depending on the other two covariates:

$$\mathbb{P}(M_3 = 1 \mid X_1, X_2) = \Phi(\gamma_0 + \gamma_1 X_1 + \gamma_1 X_2^2). \quad (9.6)$$

The last is NMAR; X_2 goes missing according to a similar non-linear probit model this time depending on itself and X_1 :

$$\mathbb{P}(M_2 = 1 \mid X_1, X_2) = \Phi(\gamma_0 + \gamma_1 X_1^2 + \gamma_1 X_2). \quad (9.7)$$

For each simulation, we set the number of training observations to $n = 250$ and simulate 500 times. Additionally, each simulation is carried out with different levels of missing data, approximately $\{0, 10, \dots, 70\}$ percent of rows have at

least one missing covariate entry. For the MCAR dataset, the corresponding $\gamma = \{0, 0.03, 0.07, 0.11, 0.16, 0.26, 0.33\}$ and for both the MAR and NMAR datasets it was $\gamma_0 = -3$ and $\gamma_1 = \{0, 0.8, 1.4, 2.0, 2.7, 4.0, 7.0, 30\}$.

We record results for four different scenarios: (1) $\mathbf{X}_{\text{train}}$ and \mathbf{X}_* contain missingness (2) $\mathbf{X}_{\text{train}}$ contains missingness and \mathbf{X}_* is devoid of missing data (in this case, \mathbf{X}_* is generated without the MDM to maintain a constant number of rows). (3) only complete cases of $\mathbf{X}_{\text{train}}$ are used to build the model but \mathbf{X}_* contains missingness and (4) only complete cases of $\mathbf{X}_{\text{train}}$ are used to build the model and \mathbf{X}_* is devoid of missing data.

We make a number of hypotheses about the relationship between the predictive performance of using incomplete cases (all observations) compared to the complete case performance. As we discussed in Section 9.3, **BARTm** should be able model the expectation of the marginal likelihood in selection models, thus we expect models built with incomplete cases to predict better than models that were built with only the complete cases. The complete case models suffer from performance degradation for two main reasons (1) these models are built with a smaller sample size and hence their estimate of $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}_{\text{obs}}, \mathbf{M}]$ is higher in bias and variance (2) the lack of missingness during the training phase does not allow the model to learn how to properly model the missingness, resulting in the missing data being filtered randomly from node to node during forecasting. These hypotheses are explored in Figure 9.2 by comparing the solid blue and solid red lines.

Further, during forecasting, we expect \mathbf{X}_* samples with incomplete cases to have worse performance than the full \mathbf{X}_* samples (devoid of missingness) simply because missingness is equivalent to information loss. However, for the NMAR model, we expect prediction performance on \mathbf{X}_* without missingness to eventually, as the amount of missingness increases, be beaten by the predictive performance on \mathbf{X}_* with missingness. Eventually there will be so much missingness in X_2 that (1) the trained

model on missingness will only be able to create models by using M_2 and expect M_2 in the future X_* and (2) the trained model on complete cases will never observe the response of the function where X_2 went missing. These hypotheses are explored in Figure 9.2 by comparing the solid lines to the dashed lines within the same color.

The results for the four scenarios under the three MDM's comport with our hypotheses. The solid red line is uniformly higher than the solid blue line confirming degradation for complete-case model forecasting on data with missingness. The dotted lines are lower than their solid counterparts indicating that providing more covariate information yields higher predictive accuracy. The one exception is for NMAR. After the number of rows with missingness is more than 40%, forecasts on complete-cases only begin to perform worse than the forecast data with missingness for models built with missingness (BARTm).

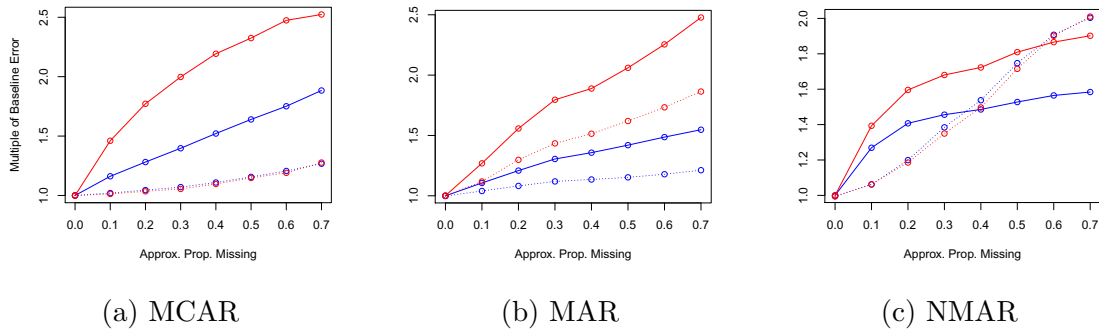


Figure 9.2: Simulation results of the response model for the three MDM's explained in the text. The y -axis measures the multiple of out-of-sample root mean square error (oosRMSE) relative to the performance under the absence of missingness. **Blue** lines correspond to the two scenarios where BART was built with all cases in X_{train} and **red** lines correspond to the two scenarios where BART was built with only the complete cases of X_{train} . Solid lines correspond to the two scenarios where X_* included missing data and dotted lines correspond to the two scenarios where the MDM was turned off in X_* .

In conclusion, for this set of simulations, **BARTm** performs better than **BART** models that ignore missingness in the training phase. The next section demonstrates **BARTm**'s performance in a real data set and compares its performance to a non-parametric statistical learning procedure that relies on imputation.

9.5 Real Data Example

The Boston Housing data (BHD) measures 14 features about housing in the $n = 506$ census tracts in Boston in 1970. For model building, the response variable is usually taken to be the median home value. For this set of simulations, we evaluate the performance of three procedures (1) **BARTm** (2) **RF** with $\mathbf{X}_{\text{train}}$ and \mathbf{X}_* imputed via **MissForest** and (3) **BART** with $\mathbf{X}_{\text{train}}$ and \mathbf{X}_* imputed via **MissForest**. Note that in these simulations we assume *a priori* that \mathbf{X}_* will have missing data. Thus, the complete-case comparisons a la Section 9.4.2 were not possible. We gauge out-of-sample predictive performance as measured by the oosRMSE for the three procedures on the simulation scenarios described in Table 9.2.

Similar to Section 9.4.2, each simulation is carried out with different levels of missing data, approximately $\{0, 10, 20, \dots, 70\}$ percent of rows have at least one missing covariate entry. For the MCAR scenario, $\gamma = \{0, 0.02, 0.04, 0.07, 0.10, 0.13, 0.17\}$, for the MAR scenario and pattern mixture scenario, $\gamma_1 = \{0, 1.3, 1.5, 1.7, 2.1, 2.6, 3.1, 3.8\}$ and $\gamma_0 = -3$ and for the NMAR scenario $\gamma_1 = \{0, 3.3, 3.6, 3.9, 4.1, 4.3, 4.6, 4.8\}$ and $\gamma_0 = -3$. Similar to Section 9.4.1, we induce a pattern mixture model by creating a normally distributed offset based on missingness (we create two such offsets here). Here, we choose μ_b to be 25% of the range in y and σ_b to be $\mu_b/4$. These values are arbitrarily set for illustration purposes. It is important to note that the performance gap of **BARTm** versus **RF** with imputation can be arbitrarily increased by making μ_b larger.

Scenario	Description
Selection Model MCAR	<code>rm</code> , <code>crim</code> , <code>lstat</code> , <code>nox</code> and <code>tax</code> are each missing w.p. γ
Selection Model MAR	<code>rm</code> and <code>crim</code> are missing according to the following models: $\mathbb{P}(\mathbf{M}_{\text{rm}} = 1) = \Phi(\gamma_0 + \gamma_1(\text{indus} + \text{lstat} + \text{age}))$ $\mathbb{P}(\mathbf{M}_{\text{crim}} = 1) = \Phi(\gamma_0 + \gamma_1(\text{nox} + \text{rad} + \text{tax}))$
Selection Model NMAR	<code>rm</code> and <code>crim</code> are missing according to the following models: $\mathbb{P}(\mathbf{M}_{\text{rm}} = 1) = \Phi(\gamma_0 + \gamma_1(\text{rm} + \text{lstat}))$ $\mathbb{P}(\mathbf{M}_{\text{crim}} = 1) = \Phi(\gamma_0 + \gamma_1(\text{crim} + \text{nox}))$
Pattern Mixture	The MAR selection model above and two offsets: (1) if $\mathbf{M}_{\text{rm}} = 1$, the response is increased by $\mathcal{N}(\mu_b, \sigma_b^2)$ (2) if $\mathbf{M}_{\text{crim}} = 1$, the response is decreased by $\mathcal{N}(\mu_b, \sigma_b^2)$

Table 9.2: Missingness scenarios for the BHD simulations. Monospace `codes` are names of covariates in the BHD. Note that `rm` has sample correlations with `indus`, `lstat` and `age` of -0.39, -0.61 and -0.24 and `crim` has sample correlations with `nox`, `rad`, and `tax` of 0.42, 0.63 and 0.58. These high correlations should allow for imputations that perform well.

For each scenario and each level of missing data, we run 500 simulations. In each simulation, we first draw missingness via the designated scenario found in Table 9.2. Then, we randomly partition 80% of the 506 BHD observations (now with missingness) as $\mathbf{X}_{\text{train}}$ and the remaining 20% as \mathbf{X}_* . We build all three models (BARTm, RF with `MissForest` and BART with `MissForest`) on $\mathbf{X}_{\text{train}}$, forecast on \mathbf{X}_* and record the oosRMSE. Thus, we integrate over idiosyncrasies that could be found in a single draw from the MDM and idiosyncrasies that could be found in a single train-test partition. When using `MissForest` during training, we impute values for the missing entries in $\mathbf{X}_{\text{train}}$ using $[\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}]$ column-binded together. To obtain forecasts, we impute the missing values in \mathbf{X}_* using $[\mathbf{X}_{\text{train}}, \mathbf{X}_*]$ row-binded together then predict

using the bottom rows (i.e. those corresponding to the imputed test data). Note that we use `MissForest` in both `RF` and `BART` to ensure that the difference in predictive capabilities of `BART` and `RF` are not driving the results.

For the MCAR selection model, we hypothesize that the `MissForest`-based imputation procedures will outperform `BARTm` due to the conceptual reasons discussed in Section 9.3. For the MAR selection model, we hypothesize similar performance between `BARTm` and both `MissForest`-based imputation procedures, as both MIA and imputation are designed to perform well in this scenario. In the NMAR selection model and pattern mixture model, we hypothesize that `BARTm` will outperform both `MissForest`-based imputation procedures, as `MissForest` (1) cannot make use of the values in the missingness columns it is trying to impute and (2) cannot construct different submodels based on missingness. Although imputation methods are not designed to handle these scenarios, it is important to run this simulation to ensure that `BARTm`, which *is* designed to succeed in these scenarios, has superior out-of-sample predictive performance.

The results displayed in Figure 9.3 largely comport with our hypotheses. Methods based on `MissForest` perform better on the MCAR selection model scenario (Figure 9.3a) and `BARTm` is stronger in the NMAR scenario (Figure 9.3c) and pattern mixture scenario (Figure 9.3d). It is worth noting that in the MAR selection model scenario (Figure 9.3b), `BARTm` begins to outperform the imputation-based methods once the percentage of missing data becomes greater than 20%. The performance of the imputation-based algorithms degrades rapidly here, while `BARTm`'s performance remains fairly stable, even with 70% of the rows having at least one missing entry. In conclusion, `BARTm` will generally perform better than `MissForest` because it is not “limited” to what can be imputed from the data on-hand. This advantage generally grows with the amount of missingness.

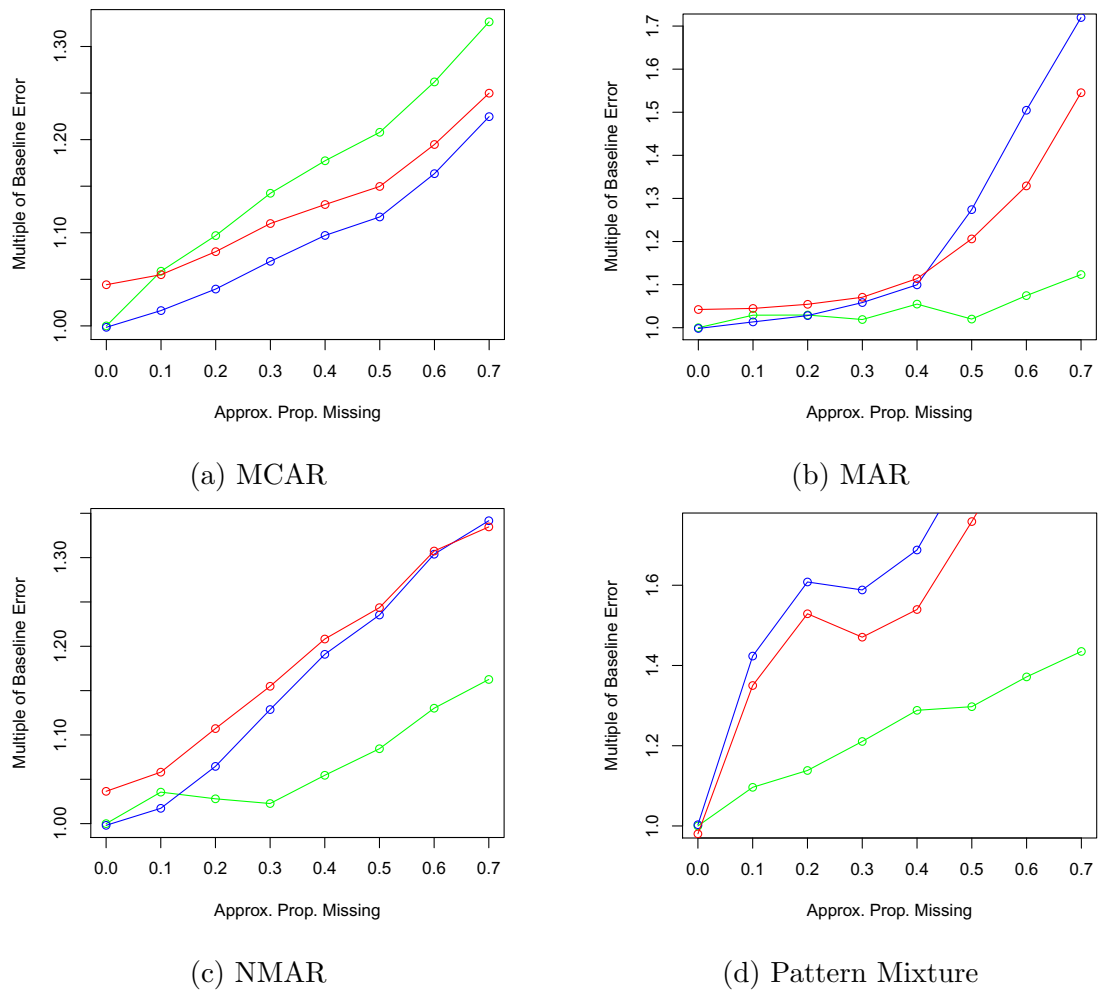


Figure 9.3: Simulations for different probabilities of missingness across the four simulated missing data scenarios in the BHD. The y-axis is oosRMSE relative to BART’s oosRMSE on the full dataset. Lines in **green** plot BARTm’s performance, lines in **red** plot RF-with-MissForest’s performance, and lines in **blue** plot BART-with-MissForest’s performance. Note that the MissForest-based imputation might perform worse in practice because here we allow imputation of the entire test set. In practice, it is likely that test observations appear sequentially.

9.6 Discussion

We propose a means of incorporating missing data into statistical learning for prediction problems where missingness may appear during both the training and forecasting phases. Our procedure, **BARTm**, implements “missing incorporated in attributes” (MIA), a technique recently explored for use in decision trees, into Bayesian Additive Regression Trees, a newly developed tree-based statistical learning algorithm for classification and regression. MIA natively incorporates missingness by sending missing observations to one of the two daughter nodes. Due to the Bayesian framework and the Metropolis-Hastings sampling, missingness is incorporated into splitting rules which are chosen to increase overall model likelihood. This innovation allows missingness itself to be used as a legitimate value within splitting criteria, resulting in no need for imputing in the training or new data and no need to drop incomplete cases.

For the simulations explored in this article, **BARTm**’s performance was superior to models built using complete cases, especially when missingness appeared in the test data as well. Additionally, **BARTm** provided higher predictive performance on the MAR selection model relative to **MissForest**, a non-parametric imputation technique. We also observe promising performance on NMAR selection models and pattern mixture models in simulations. To the best of our knowledge, there is no clear evidence of other techniques that will exhibit uniformly better predictive performance in both selection and pattern mixture missingness models. Additionally, **BARTm**’s Bayesian nature provides informative credible intervals reflecting uncertainty when the forecasting data has missing covariates.

Although the exploration in this article was focused on regression, we have observed **BARTm** performing well in binary classification settings. **BARTm** for both classification and regression is implemented in the R package **bartMachine**.

Due to MIA’s observed promise, we recommend it as a viable strategy to handle

missingness in other tree-based statistical learning methods. Future work should also consider exploration of methods that combine imputation with MIA appropriately, in order to enhance predictive performance for MCAR missing data mechanisms.

Supplementary Materials

Simulated results, tables, and figures can be replicated via the scripts located at <http://github.com/kapelner/bartMachine> (the home of the CRAN package `bartMachine`) in the `missing_data_paper` folder.

Acknowledgements

We thank Richard Berk, Dana Chandler, Ed George, Dan Heitjan, Shane Jensen, and José Zubizarreta for helpful discussions. Adam Kapelner acknowledges support from the National Science Foundation's Graduate Research Fellowship Program.

A.1 Supplement for Chapter 2

A.1.1 Detailed Experimental Design

This section details exact screens shown to users in the experimental groups. The worker begins by encountering the HIT on the MTurk platform (see Figure A.1).



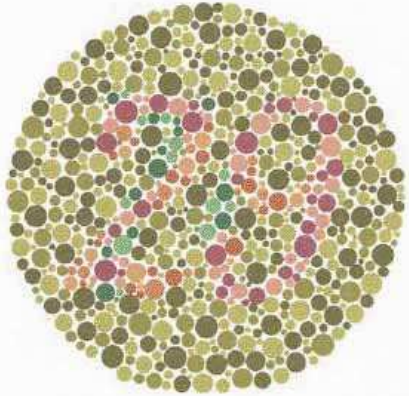
Figure A.1: The HIT as initially encountered on MTurk. Note: we used an alias in order to appear as a non-corporate and non-institutional employer.

The worker can then click on the HIT and they see the “preview screen” which describes the HIT (not shown) with text. In retrospect, a flashy image enticing the worker into the HIT would most likely have increased throughput. If the worker

chooses to accept, they are immediately directed to a multi-purpose page which hosts a colorblindness test, demographic survey, and an audio test for functioning speakers (see Figure A.2). Although many tasks require workers to answer questions before working, we avoided asking too many survey-like questions to avoid appearing as an experiment.

Since the tasks you will perform require you to be able to differentiate color, we have to ask you a few questions that will determine if you may be colorblind.

1. Look at the below image.



Do you see a number? If so, enter it into this box:

2. Are you male or female?
 Male Female

3. Have you ever had trouble differentiating between reds and greens?
 Yes No

4. Have you ever had trouble differentiating between blues and yellows?
 Yes No

5. How old are you?


6. Listen to the following sound clip () and enter the word below:

Figure A.2: The colorblindness test.

At this point, the worker is randomized into one of the three treatments and

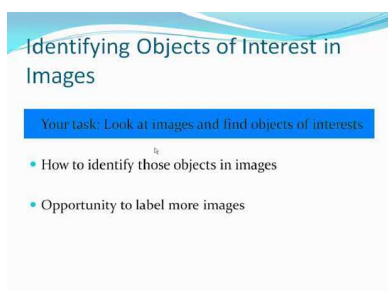
transitioned to the “qualification test.” The page displays an instructional video varying by treatment which they cannot fast-forward. Screenshots of the video are shown in Figures A.3, A.4, and A.5.¹

We include the verbatim script for the videos below. Text that differs between treatments is typeset in square brackets separated by a slash. The text before the slash in red belongs to the meaningful treatment and the text following the slash in blue belongs to both the zero-context and shredded treatments.

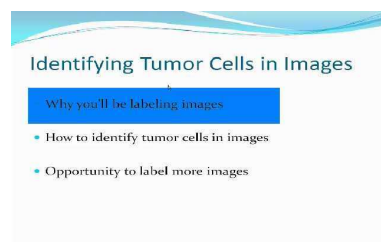
Thanks for participating in this task. [Your job will be to help identify tumor cells in images and we appreciate your help. / In this task, you’ll look at images and find objects of interest.]

In this video tutorial, we’ll explain [three / two] things:

[First, why you’re labeling the images, which is to help researchers identify tumorous cancer cells. Next, we’ll show you how to identify those tumor cells. / First, we’ll show you how to identify objects of interest in images.] [Finally, / Then,] we’ll explain how after labeling your first image you’ll have a chance to label some more.



(a) Zero-context / Shredded treatments



(b) Meaningful treatment

Figure A.3: Opening screen of training video.

Now we’re ready to learn how to identify [tumor cells / objects of interest] in images. Some example pictures of the [tumor cells / objects of interest] you’ll be identifying can be found at the bottom left. Each [tumor cell / object of interest] is blue and circular and surrounded by a red border.

When you begin each image, the magnification will be set to the lowest resolution. This gives you an overview of all points on the image, but you’ll need to zoom in and out in order to make the most precise clicks in the center of the [tumor cells / objects of interest].

Let’s scroll through the image and find some [tumor cells / objects of interest] to identify.

¹We thank Rob Cohen who did an excellent job narrating both scripts.

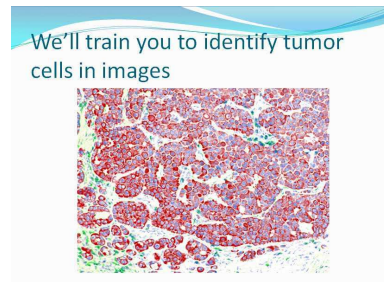
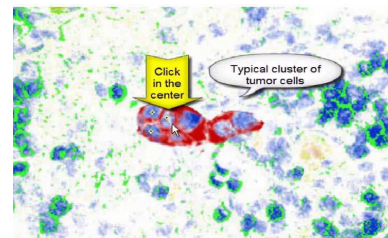
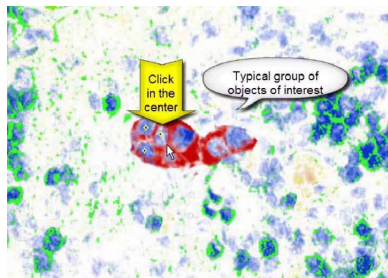


Figure A.4: Examples of meaningful cues which are not present in the Zero-context and Shredded treatment instructional video.



(a) Zero-context / Shredded treatments

(b) Meaningful treatment

Figure A.5: Describing the training process.

Here's a large cluster of [tumor cells / objects of interest]. To identify them, it is very important to click as closely to the center as possible on each [cell / object]. If I make a mistake and don't click in the center, I can undo the point by right-clicking.

Notice that this [cell / point] isn't entirely surrounded by red, [probably because the cell broke off]. Even though it's not entirely surrounded by red, we still want to identify it as a [tumor cell / object of interest].

In order to ensure that you've located all [tumor cells / objects of interest], you should use the thumbnail view in the top right. You can also use the magnification buttons to zoom out.

It looks like we missed a cluster of [tumor cells / objects of interest] at the bottom. Let's go identify those points.

Remember once again, that if you click on something that is not a [tumor cell / object of interest], you can unclick by right-clicking.

Using the scroll bars, we'll navigate to the other points ... and here's some more to the left ... Now that we think we've identified all points, let's zoom out to be sure and scroll around.

Before submitting, we should be sure of three things: (1) That we've identified all [tumor cells / objects of interest] (2) That we've clicked in the center of each one (3) That we haven't clicked on

anything that's not a [tumor cell / object of interest].

Once we've done that, we're ready to submit.

Finally, after you complete your first image, you'll have an opportunity to label additional images as part of this HIT.

The first images you label will pay more to compensate for training.

After that, as part of this HIT you'll have the chance to identify as many additional images as you like as long as you aren't taking more than 15 minutes per image.

Although you can label unlimited images in this HIT, you won't be able to accept more HITs. This is to give a variety of turkers an opportunity to identify the images.

[Thank you for your time and effort. Advances in the field of cancer and treatment prevention rely on the selfless contributions of countless individuals such as yourself.]

Then, workers must take a quiz (see Figure A.6). During the quiz, they can watch the video freely (which was rarely done).

Upon passing, they began labeling their first image (see Figure A.7). The training interface includes the master training window where workers can create and delete points and scroll across the entire image. To the left, there is a small image displaying example tumor cells. Above the master window, they have zoom in / out buttons. And on the top right there is a thumbnail view of the overall image.

Participants were given 15 minutes to mark an image. Above the training window, we displayed a countdown timer that indicated the amount of time left. The participant's total earnings was also prominently displayed atop. On the very top, we provided a submit button that allowed the worker to submit results at any time.

Each image had the same 90 cells from various-sized clusters. The cell clusters were selected for their unambiguous examples of cells, thereby eliminating the difficulty of training the difficult-to-identify tumor cells. In each image, the same clusters were arranged and rotated haphazardly, then pasted on one of five different believable backgrounds using Adobe Photoshop. Those clusters were then further rotated to create a set of ten images. This setup guarantees that the difficulty was relatively the same image-image. Images were displayed in random order for each worker, repeating

Please answer the below questions. Once you answer these questions, you will be qualified to help identify tumor cells.

- 1 - You should adjust the magnification in order to...
 - Make a prettier picture
 - Make the tumors exactly 10 pixels across
 - Find tumors and make clicks as close to the center as possible

- 2 - When you are clicking on tumor cells, how many times do you click on the tumor?
 - Once
 - Twice
 - As many dots as you can fit inside the tumor

- 3 - When you incorrectly click on an area, you should...
 - Give up the HIT
 - Reload the page
 - Use the right mouse button

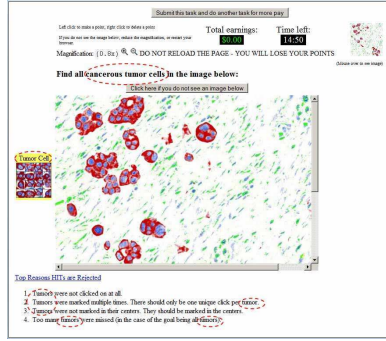
- 4 - What will happen if you don't accurately click on the tumor?
 - Scientists who are depending on you to identify tumors will not have accurate results
 - Your HIT may be rejected
 - You will not be allowed to do additional HITs with us
 - All of the above.

- 5 - Your HIT will be rejected if..
 - You do not click on all the tumors
 - You don't click in the center of the tumor
 - You click multiple times on the same tumor or on things that are NOT tumors
 - All of the above.

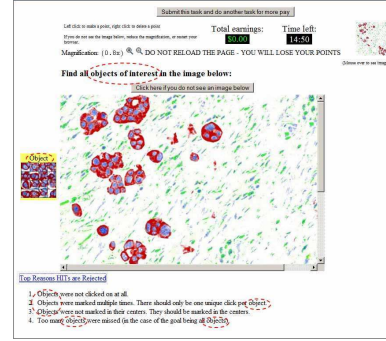
- 6 - As part of this HIT, how many images will you have the chance to identify assuming you correctly label the tumors?
 - As many as you want within a 4hr period (as long as you complete each image within 15 minutes)
 - You can label up to 4 more images
 - You must submit since you can only label one image

Begin Task

Figure A.6: The quiz after watching the training video for the meaningful treatment. In the zero-context and shredded treatments, all mention of “tumor cells” are replaced by “objects of interest.” The shredded treatment has an additional question asking them to acknowledge that they are working on a test system and their work will be discarded. Green indicates a correct response; red indicates an incorrect response.



(a) Zero-context / Shredded treatments



(b) Meaningful treatment

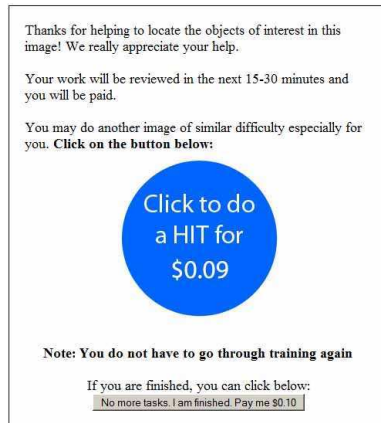
Figure A.7: The training interface as seen by workers. The meaningful interface reminds the subjects in 8 places that they are identifying tumor cells. The zero-context interface only says “objects of interest” and the shredded condition in addition has a message in red indicating that their points will not be saved (unshown). The circles around each point were *not* visible to participants. We display them to illustrate the size of a 10-pixel radius.

after each set of ten (repetition was not an issue since it was rare for a participant to label more than ten).

After the worker is finished labeling, the worker presses submit and they are led to an intermediate page which asks if they would like to label another image and the new wage is prominently displayed (see Figure A.8). In the meaningful treatment, we add one last cue of meaning — a stock photo of a researcher to emphasize the purpose of the task. In the shredded treatment, we append the text “NONE of your points will be saved because we are testing our system, but you will still be paid.” If the worker wishes to continue, they are led to another labeling task; otherwise, they are directed to the post manipulation check survey shown in figure A.9.

The program ensures that the worker is being honest. We require them to find more than 20% of the cells (the workers were unaware that we were able to monitor their accuracy). If they are found cheating on three images, they are deemed *fraud-*

ulent and not allowed to train more images. Since payment is automatic, this is to protect us from a worker depleting our research account. In practice, this happened rarely and was not correlated with treatment.



(a) Zero-context / Shredded treatments



(b) Meaningful treatment

Figure A.8: The landing page after a labeling task is completed. At this point, workers are asked if they'd like to label another image or quit and be paid what they've earned so far.

The majority of data (70%) was collected over three days (January 5-8, 2010). However, since Indians in our sample had high attrition, we collected the remaining 30% from January 29 - February 2. In order to be balanced on time-of-day, we posted tasks so that the local time in the US and India would be as similar as possible. TK -we don't need this anymore, it's not common to talk about dates of data collection anyway.

Thanks for all of your work. In order to improve our HIT, please complete the following OPTIONAL feedback form.

If you do not want to fill in the survey, please click here:

[Submit to MTurk.](#)

Please rate how much you agree/disagree with the following statements (where 1-strongly disagree, 5-strongly agree):

The task was fun/enjoyable <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5
I liked that the task seemed to be useful and had a good purpose <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5
I felt good completing the task <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5
The task seemed a lot more meaningful than the average MTurk HITs <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5
The task was well-designed and respected my efforts and work more than the average MTurk HITs <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5
Any other comments: <input type="text"/>
Submit my survey

Figure A.9: The survey a subject fills out upon completion of the task.

A.1.2 A Technical Guide to Running Field Experiments on Mechanical Turk

Institutional Review Board (IRB) Requirements

This study requires the use of deception in order to observe social preferences in a natural environment and thus is not exempted under category 2's survey procedures. The issue is you cannot give the subjects an initial consent form indicating that they are part of an experiment.

Upon waiving the requirement of consent, the IRB will most likely require you to issue a debrief statement to your subjects stating that they were part of an experiment, a blurb about the purpose of the experiment, and contact information to your institution's IRB. In order for the experiment to work properly, you can only issue the debriefing *after* data collection is completed. Otherwise, MTurk subjects can communicate to each other that this is a study and this may be a problem for internal validity.

Engineering Required

The implementation of an experiment of similar scale to the one we describe in the paper requires no more than two weeks of full-time work for an experienced software engineer.

It is critical that the engineer be fluent in "front-end" design. The front-end is what your subjects will use throughout the experiment and it can be highly dynamic, responding to the individual participant's actions. MTurk tasks are rendered in HTML and CSS. Javascript controls the dynamism and AJAX provides smooth client-server communication. Although not used in this study, the front-end can become even more fancy by implementing Adobe Flash or Microsoft Silverlight applets.

An MTurk experiment also requires a back-end web stack consisting of an http server, a database, and a server-side platform. The back-end's function is to render webpages and store the experiment's data. We recommend Ruby on Rails 3.1 because of its rapid development speed, inexpensive and instant deployment to rented space in "the Cloud," and because it is free and open-source.

Since the server must communicate with MTurk, it is convenient for the engineer to also have experience using the Amazon web services application programming interface (MTurk's API) as well as CRON jobs on Linux.

We recommend setting up two cron jobs for effective experimentation:

- (a) A CRON that creates HITs every 15 minutes. In order for subjects to see your task, it must remain fresh on Amazon's main page. Creating tasks every 15 min with short expiration times (an hour or so) will allow for maximum throughput.
- (b) Another CRON that automatically pays workers. We recommend doing this every hour or so. It's important that this be done automatically including the bonuses and rejections otherwise it will be cumbersome for the experimenter to check through each and every assignment.

Apart from technical skills, the engineer should also have knowledge of the principles of experimental design.

A.1.3 Instructions for Replication of this Study

- (a) Obtain the source from http://github.com/kapelner/breaking_monotony.
- (b) Create a Ruby on Rails hosting account on EngineYard cloud at <http://cloud.engineyard.com> and configure for Rails 3.1.
- (c) Modify the `PersonalInformation.rb` file by adding your MTurk account information, server IP address, and the administrator login and password.

- (d) Deploy the application and setup the two CRON jobs recommended in the previous section.

A.2 Supplement for Chapter 4

A.2.1 Demographics Questions

Questions and answer choices will be displayed in random order to reduce possible satisficing bias. The fade-in *a la* Kapelner and Chandler (2010) will not be used here to reduce burden to the subject. Questions marked “NFC” are a subset of the “need for cognition” survey, a scale which measures to what degree subjects desire to think abstractly and expend mental effort (see Cacioppo et al., 1984). Questions marked “BIG5” measure attributes on the well-accepted NEO five-factor model of personality traits (see McCrae and Costa, 1987); we employ a highly modified survey that includes only one question per factor.

- 1) Age (natural number ≥ 18)
- 2) Gender (binary: 1 for male)
- 3) Urban or Rural location (binary: 1 for urban)
- 4) Highest Education Attainment (categorical: high school, some college, bachelors, masters, doctorate)
- 5) Marital Status (categorical: married, divorced, single, widowed, in a relationship)
- 6) Employment Status (categorical: full-time, part-time, self-employed, unemployed)
- 7) Number of children (natural number)
- 8) Annual Personal Income (approximated to the nearest dollar)
- 9) Religion by Birth (categorical: Protestant Christian, Roman Catholic, Evangelical Christian, Jewish, Muslim, Hindu, Buddhist, Other)

- 10) Belief in a divine being (binary: 1 for belief)
- 11) Race (categorical: White, Hispanic, Black, Asian, Native American, Other)
- 12) Would you want to participate in an academic study? (binary: 1 for participate)
- 13) NFC: I prefer my life to be filled with puzzles I must solve (binary: 1 for prefer)
- 14) NFC: It's enough for me that something gets the job done; I don't care how or why it works (binary: prefer)
- 15) BIG5: I am more (i) inventive and curious (ii) consistent and cautious (binary: 1 for i)
- 16) BIG5: I am more (i) efficient and organized (ii) easy-going and more careless (binary: 1 for i)
- 17) BIG5: I am more (i) outgoing and energetic (ii) solitary and reserved (binary: 1 for i)
- 18) BIG5: I am more (i) friendly and compassionate (ii) cold and unkind (binary: 1 for i)
- 19) BIG5: I am more (i) sensitive and nervous (ii) secure and confident (binary: 1 for i)

A.2.2 Experimental Covariates

In addition to the demographic questions, we will also collect as much information as we could during the experiment about subject behavior:

- 1) Time (in seconds) it takes the subject to accept the HIT and begin the task.
- 2) Time (in seconds) it takes for the subject to fill in the demographic questions.

- 3) Time (in seconds) it takes the subject to do the experimental question.
- 4) Number of page reloads on the experimental question.
- 5) Number of characters in the free-response comments box after the experiment is completed.

A.2.3 Full Text of the Studies

Below is the full text of each of the studies. Text displayed for subjects in the *treatment* group is colored red and text displayed for subjects in the *control* group is colored green.

(a) Framing Study

You are lying on the beach on a hot day. All you have to drink is ice water. For the last hour you have been thinking about how much you would enjoy a nice cold bottle of your favorite brand of beer.

A companion gets up to go make a phone call and offers to bring back a beer from the only nearby place where beer is sold, **a fancy resort hotel** / **a small, run-down grocery store**

He says that the beer might be expensive and so asks how much you are willing to pay for the beer. He says that he will buy the beer if it costs as much or less than the price you state. But if it costs more than the price you state he will not buy it.

You trust your friend, and there is no possibility of bargaining with **the bartender.** / **store owner.** What price do you tell him?

(b) Priming Study

As Jesus started on his way, a man ran up to him and fell on his knees before him. “Good teacher,” he asked, “what must I do to inherit eternal life?”

“Why do you call me good?” Jesus answered. “No one is good-except God alone. You know the commandments: You shall not murder, you shall not commit adultery, you shall not steal, you shall not give false testimony, you shall not defraud, honor your father and mother.”

“Teacher,” he declared, “all these I have kept since I was a boy.”

Jesus looked at him and loved him. “One thing you lack,” he said. “Go, sell everything you have and give to the poor, and you will have treasure in heaven. Then come, follow me.”

At this the man’s face fell. He went away sad, because he had great wealth.

Jesus looked around and said to his disciples, “How hard it is for the rich to enter the kingdom of God!”

Some species of fish are viviparous. In such species the mother retains the eggs and nourishes the embryos. Typically, viviparous fish have a structure analogous to the placenta seen in mammals connecting the mother’s blood supply with that of the embryo.

Examples of viviparous fish include the surf-perches, spltfins, and lemon shark. Some viviparous fish exhibit oophagy, in which the developing embryos eat other eggs produced by the mother. This has been observed primarily among sharks, such as the shortfin mako and porbeagle, but is known for a few bony fish as well, such as the halfbeak *Nomorhamphus ebrardtii*.

Intrauterine cannibalism is an even more unusual mode of vivipary, in

which the largest embryos eat weaker and smaller siblings. This behavior is also most commonly found among sharks, such as the grey nurse shark, but has also been reported for *Nomorhamphus ebrardtii*. Aquarists commonly refer to ovoviviparous and viviparous fish as live-bearers although they are known by other colloquial names as well.

You have just been randomly paired with another Turker. He or she is seeing the same thing you are right now and you can choose to cooperate with each other or not to cooperate. If both of you choose to cooperate, you will both be awarded a bonus of 10 cents.

However, if you choose to cooperate and the other person doesn't, you make a bonus of 20 cents and the other person gets nothing.

If you both choose to not cooperate with each other at the same time, you both will get 4 cents as a bonus.

Would you like to cooperate with this other person?"

(c) Sunk Cost

You paid handsomely for tickets / **Your friend gave you tickets for free** for a theater production that received rave reviews. You've been wanting to see this show for quite some time and it only plays once in your city.

However, during the night of the show, the weather turned for the worst. There's heavy rain and some hail. Your city's mayor is not recommending anyone drive on the roads. Despite the inclement weather, the show is still going to be performed.

How likely are you to go? 0 means "you're not going" and 9 means "you're definitely going" while 5 would mean that "you're unsure about going."

A.3 Supplement for Chapter 5

An additional strategy is to match after the experiment has completed. We choose to dynamically allocate using the stratification procedure found in Section 3. Once all n subjects are allocated, we then calculate all $\binom{n}{2}$ Mahalanobis-squared distances according to Equation 2. Note that we have the luxury here of using all n subjects to calculate \mathbf{S}^{-1} . Then, using the $F_{p,n-p}$ approximation, we convert the squared distances to nominal probabilities. We then use λ as a cutoff: if the probability distance between two subjects is greater than λ , we set this distance to ∞ . We then use Hansen and Klopfer (2006)'s `optmatch` package in R to do an optimal matching based on this enhanced distance matrix. The optimal matches found become the *matched* data and the data that was not optimally matched becomes the *unmatched* data. The unmatched data is the analogue of the reservoir during the *on-the-fly* matching as explained in Section 2.2. Using these two subsets of the data, we can now generate estimates according to Equations 4 and 6. Over many repeated samples, we can estimate power, just like among the other competitors.

Figures A.10 to A.17 show the power results of Section 3's simulations for many different values of λ . We include the post-matching results in these figures under "SPM."

Table A.1 displays results for bias including post-matching results.

		Sample Average Absolute Bias					
Allocation		Scenario NL		Scenario LI		Scenario ZE	
n	Method	Classic	Linear	Classic	Linear	Classic	Linear
50	CR	0.704	0.635	0.752	0.404	0.402	0.395
	Efron's BCD	0.685	0.654	0.730	0.395	0.391	0.412
	Stratification	0.612	0.556	0.509	0.395	0.380	0.385
	Minimization	0.618	0.601	0.519	0.392	0.387	0.382
	Seq. Matching	0.466	0.466	0.478	0.434	0.418	0.455
	Seq. Post Matching	0.552	0.552	0.504	0.419	0.400	0.427
100	CR	0.515	0.454	0.530	0.287	0.280	0.283
	Efron's BCD	0.502	0.447	0.534	0.279	0.285	0.282
	Stratification	0.427	0.409	0.362	0.270	0.276	0.277
	Minimization	0.441	0.424	0.340	0.275	0.278	0.277
	Seq. Matching	0.318	0.322	0.324	0.298	0.286	0.295
	Seq. Post Matching	0.388	0.420	0.350	0.295	0.272	0.291
200	CR	0.369	0.321	0.376	0.198	0.196	0.193
	Efron's BCD	0.340	0.321	0.379	0.201	0.197	0.195
	Stratification	0.304	0.284	0.255	0.197	0.196	0.198
	Minimization	0.315	0.299	0.240	0.194	0.193	0.196
	Seq. Matching	0.210	0.221	0.210	0.199	0.195	0.207
	Seq. Post Matching	0.290	0.313	0.242	0.200	0.198	0.194

Table A.1: Average absolute bias of sequential matching and post matching using stratification to allocate versus competitors by scenario and testing procedure. Both matching methods used $\lambda = 0.10$ (other values yielded similar results) and the Z approximation of Equations 4 and 6. Exact tests are not shown because they do not admit an estimate, only a p value.

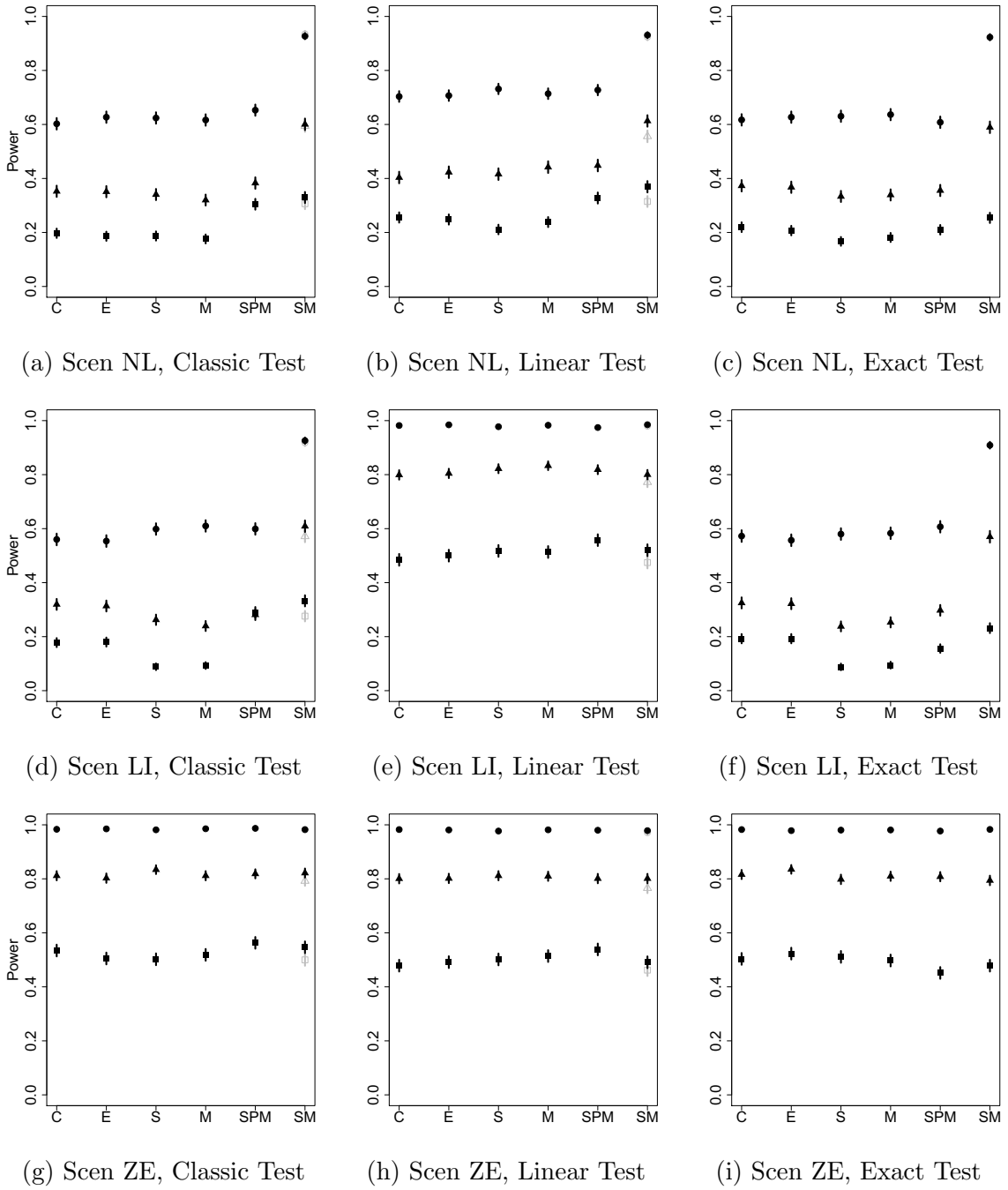


Figure A.10: Power illustrated for matching parameter $\lambda = 1\%$. We also include the post matching procedure of Appendix A.3 as “SPM.” See paper, Figure 1 for full caption information.

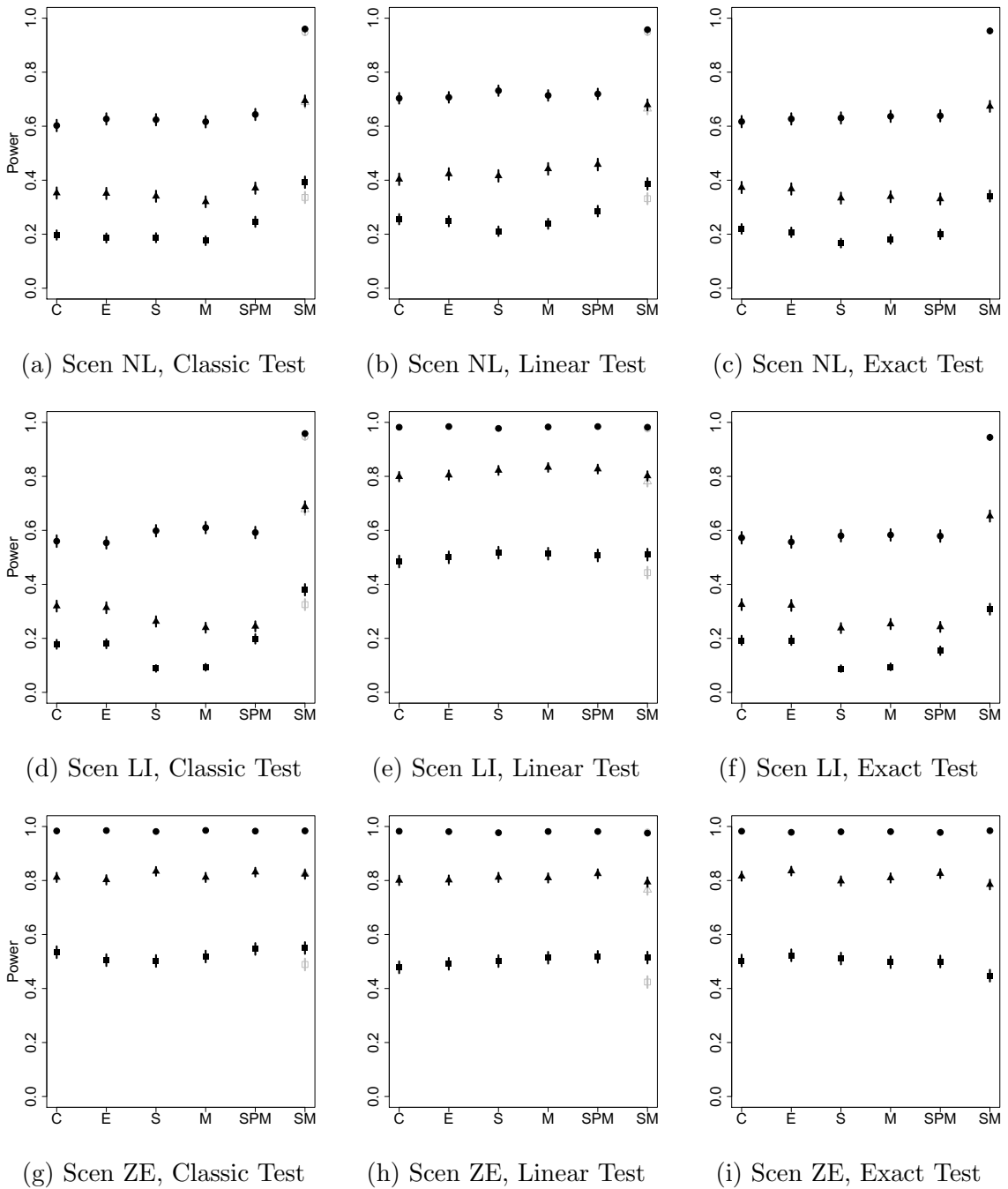


Figure A.11: Power illustrated for matching parameter $\lambda = 2.5\%$. We also include the post matching procedure of Appendix A.3 as “SPM.” See paper, Figure 1 for full caption information.

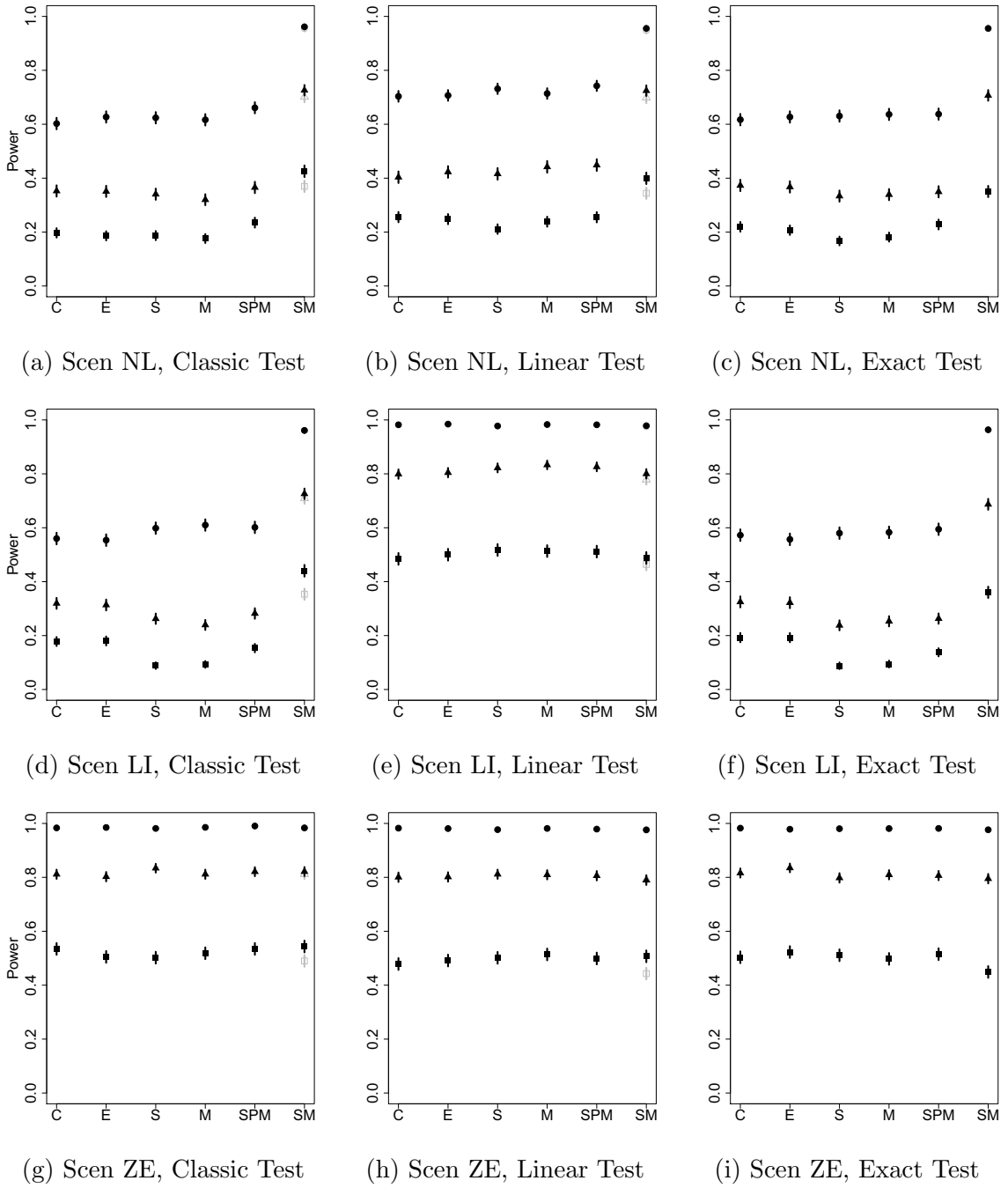


Figure A.12: Power illustrated for matching parameter $\lambda = 5\%$. We also include the post matching procedure of Appendix A.3 as “SPM.” See paper, Figure 1 for full caption information.

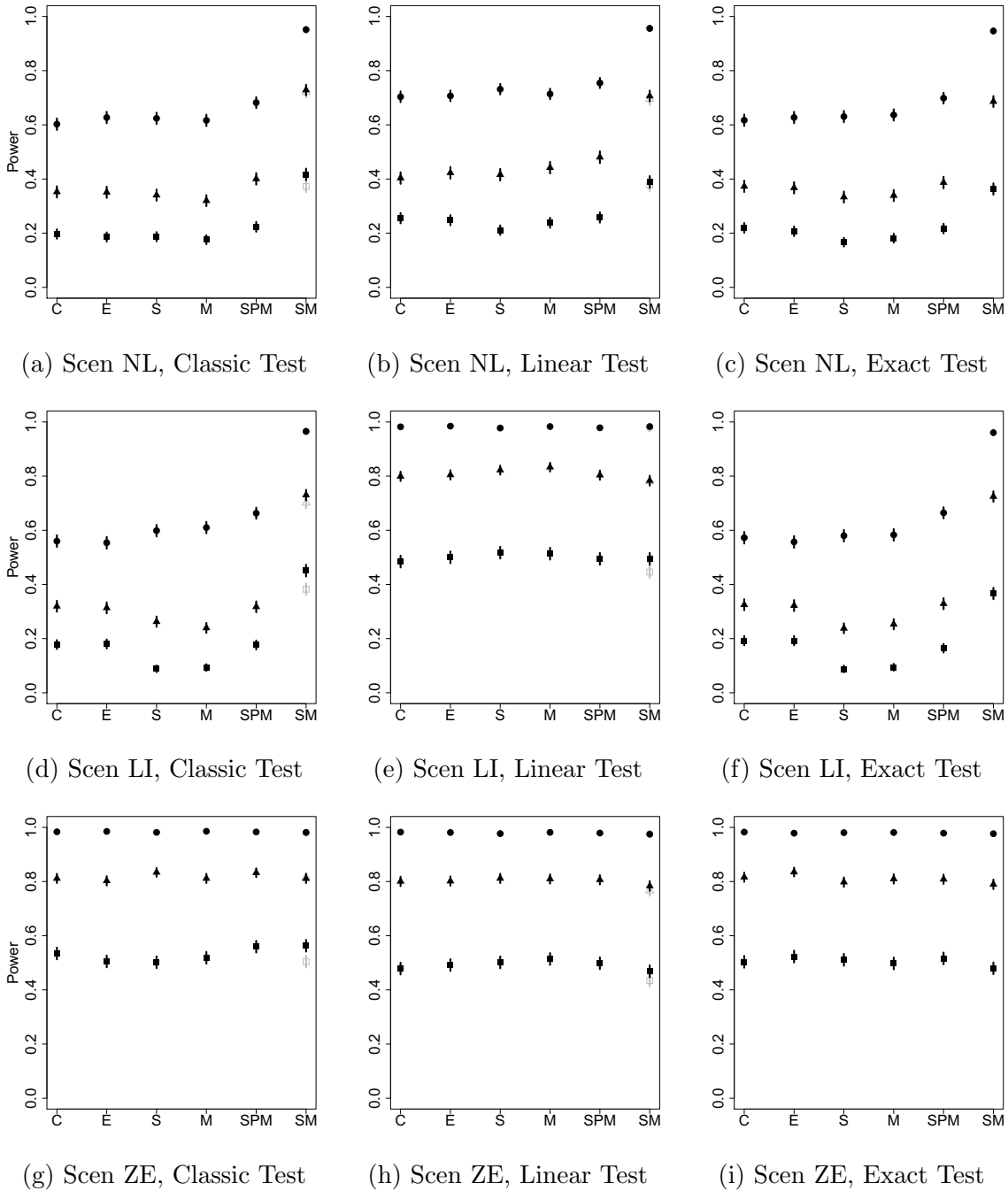


Figure A.13: Power illustrated for matching parameter $\lambda = 7.5\%$. We also include the post matching procedure of Appendix A.3 as “SPM.” See paper, Figure 1 for full caption information.

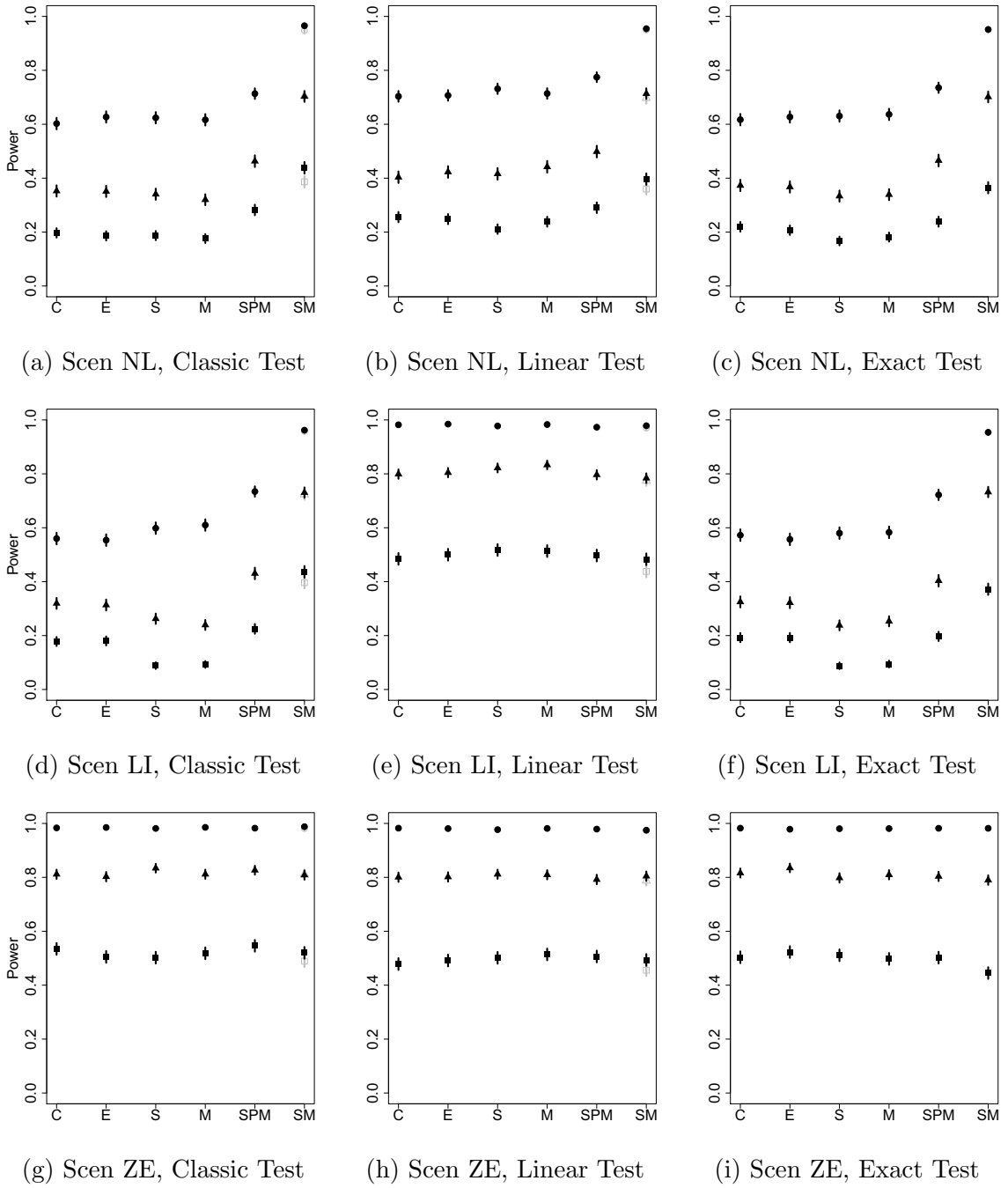


Figure A.14: Power illustrated for matching parameter $\lambda = 10\%$. We also include the post matching procedure of Appendix A.3 as “SPM.” See paper, Figure 1 for full caption information.

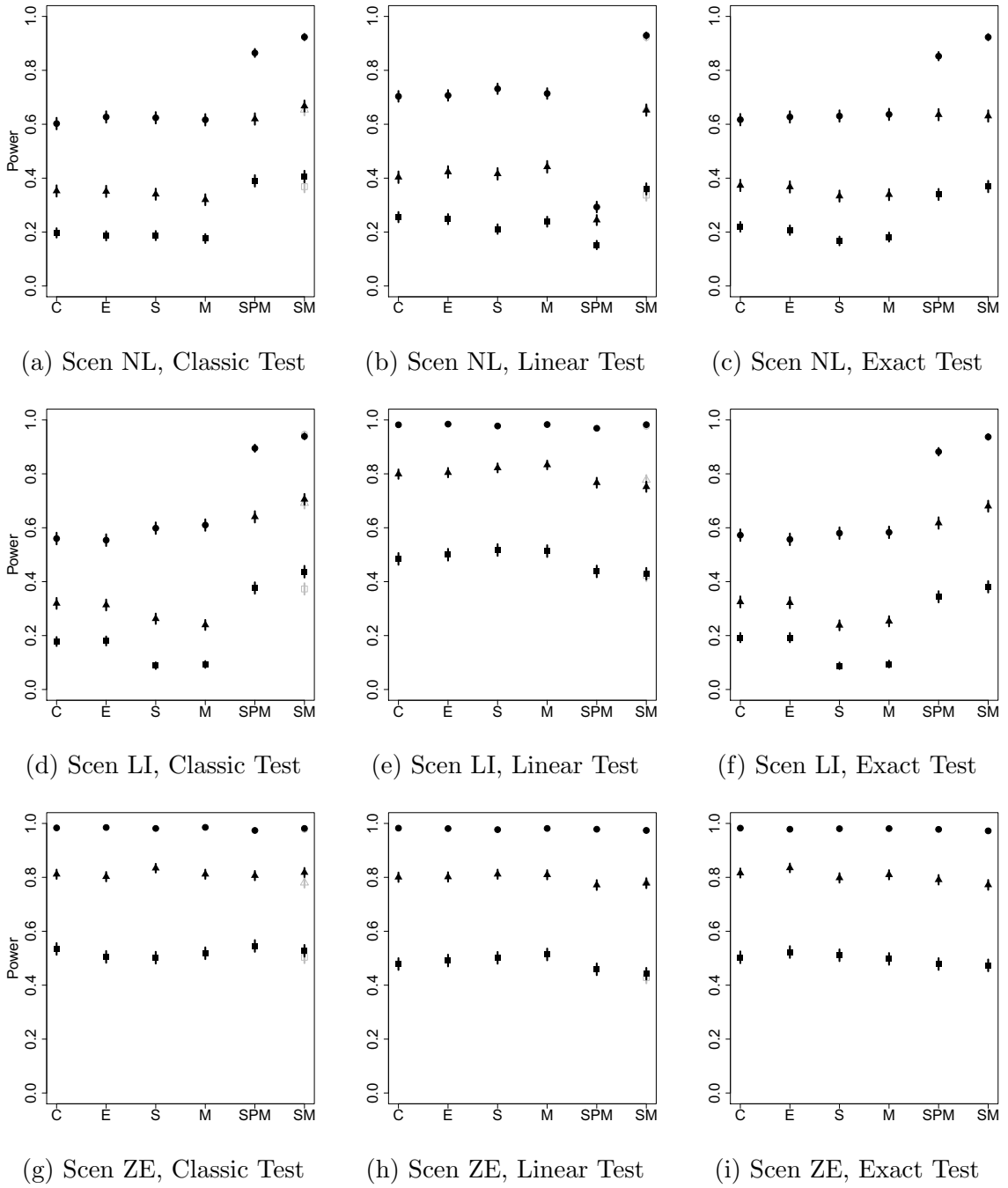


Figure A.15: Power illustrated for matching parameter $\lambda = 20\%$. We also include the post matching procedure of Appendix A.3 as “SPM.” See paper, Figure 1 for full caption information.

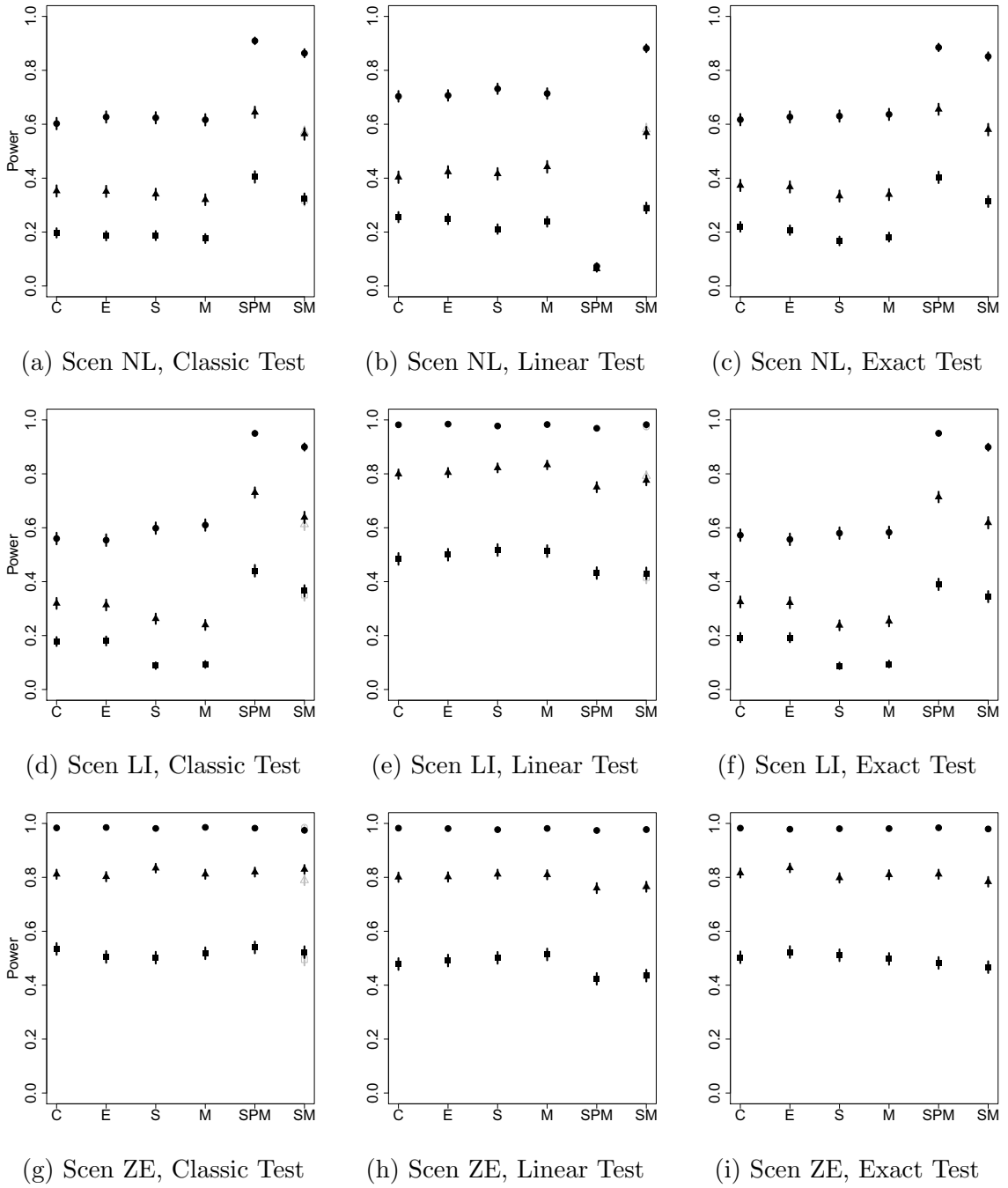


Figure A.16: Power illustrated for matching parameter $\lambda = 35\%$. We also include the post matching procedure of Appendix A.3 as “SPM.” See paper, Figure 1 for full caption information.

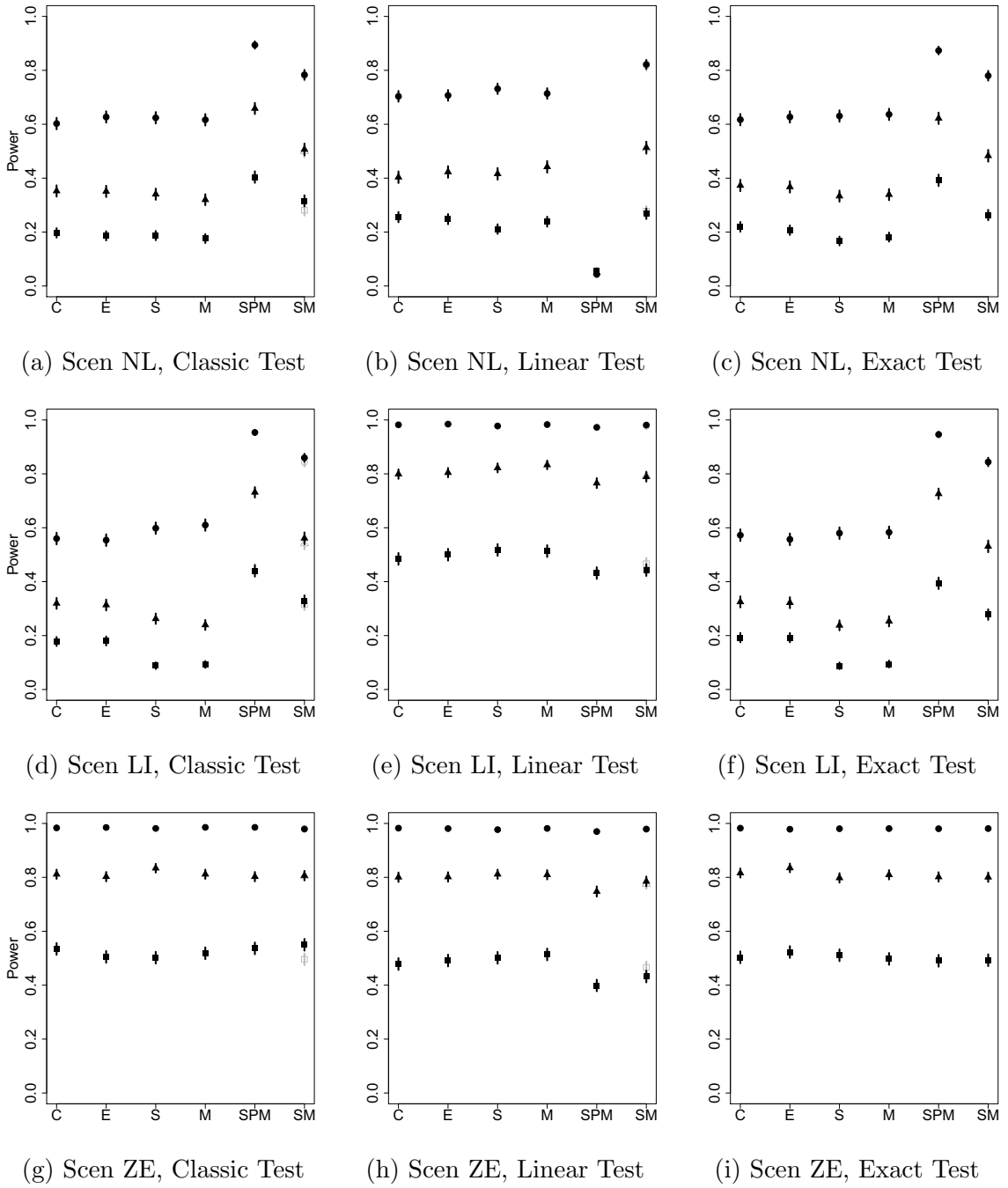


Figure A.17: Power illustrated for matching parameter $\lambda = 50\%$. We also include the post matching procedure of Appendix A.3 as “SPM.” See paper, Figure 1 for full caption information.

A.4 Supplement for Chapter 8

A.4.1 Sampling New Trees

This section provides details on the implementation of Equation A.7 (steps 1, 3, ..., $2m - 1$), the Metropolis-Hastings step for sampling new trees. Recall from Section 8.2.2 that trees can be altered via growing new daughter nodes from an existing terminal node, pruning two terminal nodes such that their parent becomes terminal, or changing the splitting rule in a node.

Below is the Metropolis ratio (Gelman et al., 2004, p.291) where the parameter sampled is the tree and the data is the responses unexplained by other trees denoted by \mathbf{R} . We denote the new, proposal tree with an asterisk and the original tree without the asterisk.

$$r = \frac{\mathbb{P}(\mathfrak{T}_* \rightarrow \mathfrak{T}) \mathbb{P}(\mathfrak{T}_* | \mathbf{R}, \sigma^2)}{\mathbb{P}(\mathfrak{T} \rightarrow \mathfrak{T}_*) \mathbb{P}(\mathfrak{T} | \mathbf{R}, \sigma^2)} \quad (\text{A.1})$$

We accept a draw from the posterior distribution of trees if a draw from a standard uniform distribution is less than the value of r . Immediately we note that it is difficult (if not impossible) to calculate the posterior probabilities for the trees themselves. Instead, we employ Bayes' Rule,

$$\mathbb{P}(\mathfrak{T} | \mathbf{R}, \sigma^2) = \frac{\mathbb{P}(\mathbf{R} | \mathfrak{T}, \sigma^2) \mathbb{P}(\mathfrak{T} | \sigma^2)}{\mathbb{P}(\mathbf{R} | \sigma^2)},$$

and plug the result into Equation A.15 to obtain:

$$r = \underbrace{\frac{\mathbb{P}(\mathbb{T}_* \rightarrow \mathbb{T})}{\mathbb{P}(\mathbb{T} \rightarrow \mathbb{T}_*)}}_{\text{transition ratio}} \times \underbrace{\frac{\mathbb{P}(\mathbf{R} | \mathbb{T}_*, \sigma^2)}{\mathbb{P}(\mathbf{R} | \mathbb{T}, \sigma^2)}}_{\text{likelihood ratio}} \times \underbrace{\frac{\mathbb{P}(\mathbb{T}_*)}{\mathbb{P}(\mathbb{T})}}_{\text{tree structure ratio}}.$$

Note that the probability of the tree structure is independent of σ^2 .

The goal of this section is to explicitly calculate r for all possible tree proposals — GROW, PRUNE and CHANGE. For each proposal, the calculations are organized into separate sections detailing each of the three ratios — transition, likelihood and tree structure. Note that our actual implementation uses the following expressions in log form for numerical accuracy.

A.4.2 Grow Proposal

Transition Ratio

Transitioning from the original tree to a new tree involves growing two daughter nodes from a current terminal node:

$$\begin{aligned} \mathbb{P}(\mathbb{T} \rightarrow \mathbb{T}_*) &= \mathbb{P}(\text{GROW}) \mathbb{P}(\text{selecting } \eta \text{ to grow from}) \times & (A.2) \\ &\mathbb{P}(\text{selecting the } j\text{th attribute to split on}) \times \\ &\mathbb{P}(\text{selecting the } i\text{th value to split on}) \\ &= \mathbb{P}(\text{GROW}) \frac{1}{b} \frac{1}{p_{\text{adj}}(\eta)} \frac{1}{n_{j\text{-adj}}(\eta)}. \end{aligned}$$

We chose one of the current b terminal nodes which we denote the η th node, and then we pick an attribute and split point. $p_{\text{adj}}(\eta)$ denotes the number of predictors left available to split on. This can be less than p if certain predictors do not have

two or more unique values once the data reaches the η th node. For example, this regularly occurs if a dummy variable was split on in some node higher up in the lineage. $n_{j\text{-adj}}(\eta)$ denotes the number of *unique* values left in the p th attribute after adjusting for parents' splits.

Transitioning from the new tree back to the original tree involves pruning that node:

$$\mathbb{P}\left(\mathbb{T}_* \rightarrow \mathbb{T}\right) = \mathbb{P}(\text{PRUNE}) \mathbb{P}(\text{selecting } \eta \text{ to prune from}) = \mathbb{P}(\text{PRUNE}) \frac{1}{w_2^*}$$

where w_2^* denotes the number of second generation internal nodes (nodes with two terminal daughter nodes) in the new tree. Thus, the full transition ratio is:

$$\frac{\mathbb{P}\left(\mathbb{T}_* \rightarrow \mathbb{T}\right)}{\mathbb{P}\left(\mathbb{T} \rightarrow \mathbb{T}_*\right)} = \frac{\mathbb{P}(\text{PRUNE})}{\mathbb{P}(\text{GROW})} \frac{b p_{\text{adj}}(\eta) n_{j\text{-adj}}(\eta)}{w_2^*}.$$

Note that when there are no variables with more two or more unique values, the probability of GROW is set to zero and the step will be automatically rejected.

Likelihood Ratio

To calculate the likelihood, the tree structure determines which responses fall into which of the b terminal nodes. Thus,

$$\mathbb{P}\left(R_1, \dots, R_n \mid \mathbb{T}, \sigma^2\right) = \prod_{\ell=1}^b \mathbb{P}\left(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} \mid \sigma^2\right)$$

where each term on the right hand side is the probability of responses in one of the b terminal nodes, which are independent by assumption. The R_ℓ 's denote the data

in the ℓ th terminal node and where n_ℓ denotes how many observations are in each terminal node and $n = \sum_{\ell=1}^b n_\ell$.

We now find an analytic expression for the node likelihood term. Remember, if the mean in each terminal node, which we denote μ_ℓ , was known, then we would have $R_{\ell_1}, \dots, R_{\ell_{n_\ell}} | \mu_\ell, \sigma^2 \stackrel{iid}{\sim} \mathcal{N}(\mu_\ell, \sigma^2)$. BART requires μ_ℓ to be margined out, allowing the Gibbs sampler in Equation A.7 to avoid dealing with jumping between continuous spaces of varying dimensions (Chipman et al., 2010, page 275). Recall that one of the BART model assumptions is a prior on the average value of $\mu \sim \mathcal{N}(0, \sigma_\mu^2)$ and thus,

$$\mathbb{P}(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} | \sigma^2) = \int_{\mathbb{R}} \mathbb{P}(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} | \mu_\ell, \sigma^2) \mathbb{P}(\mu_\ell; \sigma_\mu^2) d\mu_\ell$$

which can be shown via completion of the square or convolution to be

$$\begin{aligned} \mathbb{P}(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} | \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n_\ell/2}} \sqrt{\frac{\sigma^2}{\sigma^2 + n_\ell\sigma_\mu^2}} \times \\ &\exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_\ell} (R_{\ell_i} - \bar{R}_\ell)^2 - \frac{\bar{R}_\ell^2 n_\ell^2}{n_\ell + \frac{\sigma^2}{\sigma_\mu^2}} + n_\ell \bar{R}_\ell^2 \right)\right) \end{aligned} \quad (\text{A.3})$$

where \bar{R}_ℓ denotes the mean response in the node and R_{ℓ_i} denotes the observations $i = 1 \dots n_\ell$ in the node.

Since the likelihoods are solely determined by the terminal nodes, the proposal tree differs from the original tree by only the selected node to be grown, denoted by ℓ , which becomes two daughters after the GROW step denoted by ℓ_L and ℓ_R . Hence, the likelihood ratio becomes:

$$\frac{\mathbb{P}(\mathbf{R} | \mathfrak{F}_*, \sigma^2)}{\mathbb{P}(\mathbf{R} | \mathfrak{F}, \sigma^2)} = \frac{\mathbb{P}(R_{\ell_{L,1}}, \dots, R_{\ell_{L,n_{\ell,L}}} | \sigma^2) \mathbb{P}(R_{\ell_{R,1}}, \dots, R_{\ell_{R,n_{\ell,R}}} | \sigma^2)}{\mathbb{P}(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} | \sigma^2)} \quad (\text{A.4})$$

Plugging Equation A.16 into Equation A.17 three times yields the ratio for the GROW step:

$$\sqrt{\frac{\sigma^2 (\sigma^2 + n_{\ell} \sigma_{\mu}^2)}{(\sigma^2 + n_{\ell_L} \sigma_{\mu}^2) (\sigma^2 + n_{\ell_R} \sigma_{\mu}^2)}} \exp \left(\frac{\sigma_{\mu}^2}{2\sigma^2} \left(\frac{(\sum_{i=1}^{n_{\ell_L}} R_{\ell_L,i})^2}{\sigma^2 + n_{\ell_L} \sigma_{\mu}^2} + \frac{(\sum_{i=1}^{n_{\ell_R}} R_{\ell_R,i})^2}{\sigma^2 + n_{\ell_R} \sigma_{\mu}^2} - \frac{(\sum_{i=1}^{n_{\ell}} R_{\ell,i})^2}{\sigma^2 + n_{\ell} \sigma_{\mu}^2} \right) \right)$$

where n_{ℓ_L} and n_{ℓ_R} denote the number of data points in the newly grown left and right daughter nodes.

Tree Structure Ratio

In Section 8.2.1 we discussed the prior on the tree structure (where the splits occur) as well as the tree rules. For the entire tree,

$$\mathbb{P} \left(\frac{\mathfrak{F}}{\mathfrak{F}} \right) = \prod_{\eta \in H_{\text{terminals}}} (1 - \mathbb{P}_{\text{SPLIT}}(\eta)) \prod_{\eta \in H_{\text{internals}}} \mathbb{P}_{\text{SPLIT}}(\eta) \prod_{\eta \in H_{\text{internals}}} \mathbb{P}_{\text{RULE}}(\eta)$$

where $H_{\text{terminals}}$ denotes the set of terminal nodes and $H_{\text{internals}}$ denotes the internal nodes.

Recall that the probability of splitting on a given node η is $\mathbb{P}_{\text{SPLIT}}(\eta) = \alpha / (1 + d_{\eta})^{\beta}$. The probability is controlled by two hyperparameters, α and β , and d_{η} is the depth (number of parent generations) of node η . When assigning a rule, recall that BART picks from all available attributes and then from all available unique split points. Using the notation from the transition ratio section, $\mathbb{P}_{\text{RULE}}(\eta) = 1/p_{\text{adj}}(\eta) \times 1/n_{j\text{-adj}}(\eta)$.

Once again, the original tree features a node η that was selected to be grown. The proposal tree differs with two daughter nodes denoted η_L and η_R . We can now form the ratio:

$$\begin{aligned}
\frac{\mathbb{P}\left(\frac{\mathfrak{F}}{\mathfrak{F}_*}\right)}{\mathbb{P}\left(\frac{\mathfrak{F}}{\mathfrak{F}}\right)} &= \frac{(1 - \mathbb{P}_{\text{SPLIT}}(\eta_L))(1 - \mathbb{P}_{\text{SPLIT}}(\eta_R)) \mathbb{P}_{\text{SPLIT}}(\eta) \mathbb{P}_{\text{RULE}}(\eta)}{(1 - \mathbb{P}_{\text{SPLIT}}(\eta))} \\
&= \frac{\left(1 - \frac{\alpha}{(1 + d_{\eta_L})^\beta}\right) \left(1 - \frac{\alpha}{(1 + d_{\eta_R})^\beta}\right) \frac{\alpha}{(1 + d_\eta)^\beta} \frac{1}{p_{\text{adj}}(\eta)} \frac{1}{n_{j \cdot \text{adj}}(\eta)}}{1 - \frac{\alpha}{(1 + d_\eta)^\beta}} \\
&= \alpha \frac{\left(1 - \frac{\alpha}{(2 + d_\eta)^\beta}\right)^2}{\left((1 + d_\eta)^\beta - \alpha\right) p_{\text{adj}}(\eta) n_{j \cdot \text{adj}}(\eta)}
\end{aligned}$$

The last line follows from algebra and using the fact that the depth of the grown nodes is the depth of the parent node incremented by one ($d_{\eta_L} = d_{\eta_R} = d_\eta + 1$).

A.4.3 Prune Proposal

A prune proposal is the “opposite” of a grow proposal. Prune selects a node with two daughters and removes them. Thus, each ratio will be approximately the inverse of the ratios found in the previous section concerning the grow proposal. Note also that prune steps are not considered in trees that consist of a single root node.

Transition Ratio

We begin with transitioning from the original tree to the proposal tree:

$$\mathbb{P}\left(\frac{\mathfrak{F}}{\mathfrak{F}} \rightarrow \frac{\mathfrak{F}}{\mathfrak{F}_*}\right) = \mathbb{P}(\text{PRUNE}) \mathbb{P}(\text{selecting } \eta \text{ to prune from}) = \mathbb{P}(\text{PRUNE}) \frac{1}{w_2}$$

where w_2 denotes the number of parent nodes that have two daughters but no grand-daughters. To transition in the opposite direction, we are obligated to grow from node η . This is similar to Equation A.2 except the proposed tree has one less terminal node

due to the pruning of the original tree, resulting in a $1/(b-1)$ term:

$$\mathbb{P}\left(\mathbb{T}_* \rightarrow \mathbb{T}\right) = \mathbb{P}(\text{GROW}) \frac{1}{b-1} \frac{1}{p_{\text{adj}}(\eta^*)} \frac{1}{n_{j^* \cdot \text{adj}}(\eta^*)}.$$

Thus, the transition ratio is:

$$\frac{\mathbb{P}\left(\mathbb{T}_* \rightarrow \mathbb{T}\right)}{\mathbb{P}\left(\mathbb{T} \rightarrow \mathbb{T}_*\right)} = \frac{\mathbb{P}(\text{GROW})}{\mathbb{P}(\text{PRUNE})} \frac{w_2}{(b-1)p_{\text{adj}}(\eta^*)n_{j^* \cdot \text{adj}}(\eta^*)}.$$

Likelihood Ratio

This is simply the inverse of the likelihood ratio for the grow proposal:

$$\begin{aligned} \frac{\mathbb{P}\left(\mathbf{R} \mid \mathbb{T}_*, \sigma^2\right)}{\mathbb{P}\left(\mathbf{R} \mid \mathbb{T}, \sigma^2\right)} &= \sqrt{\frac{(\sigma^2 + n_{\ell_L} \sigma_\mu^2) (\sigma^2 + n_{\ell_R} \sigma_\mu^2)}{\sigma^2 (\sigma^2 + n_\ell \sigma_\mu^2)}} \times \\ &\exp\left(\frac{\sigma_\mu^2}{2\sigma^2} \left(\frac{(\sum_{i=1}^{n_\ell} R_{\ell,i})^2}{\sigma^2 + n_\ell \sigma_\mu^2} - \frac{(\sum_{i=1}^{n_{\ell_L}} R_{\ell_L,i})^2}{\sigma^2 + n_{\ell_L} \sigma_\mu^2} - \frac{(\sum_{i=1}^{n_{\ell_R}} R_{\ell_R,i})^2}{\sigma^2 + n_{\ell_R} \sigma_\mu^2}\right)\right). \end{aligned}$$

Tree Structure Ratio

This is also simply the inverse of the tree structure ratio for the grow proposal:

$$\frac{\mathbb{P}\left(\mathbb{T}_*\right)}{\mathbb{P}\left(\mathbb{T}\right)} = \frac{\left((1 + d_\eta)^\beta - \alpha\right) p_{\text{adj}}(\eta^*) n_{j^* \cdot \text{adj}}(\eta^*)}{\alpha \left(1 - \frac{\alpha}{(2+d_\eta)^\beta}\right)^2}.$$

A.4.4 Change

A change proposal involves picking an internal node and changing its rule by picking both a new available predictor to split on and a new valid split value among values of the selected predictor. Although this could be implemented for use in any internal node in the tree, for simplicity we limit our implementation to *singly* internal nodes: those that have two terminal daughter nodes and thus, no grand-daughters.

Transition Ratio

The transition to a proposal tree is below:

$$\begin{aligned} \mathbb{P} \left(\mathfrak{T} \rightarrow \mathfrak{T}_* \right) &= \mathbb{P}(\text{CHANGE}) \mathbb{P}(\text{selecting node } \eta \text{ to change}) \times \\ &\quad \mathbb{P}(\text{selecting the new attribute to split on}) \times \\ &\quad \mathbb{P}(\text{selecting the new value to split on}) \end{aligned}$$

When calculating the ratio, the first three terms are shared in both numerator and denominator. The probability of selecting the new value to split on will differ as different split features have different numbers of unique values available. Thus we are left with

$$\frac{\mathbb{P} \left(\mathfrak{T}_* \rightarrow \mathfrak{T} \right)}{\mathbb{P} \left(\mathfrak{T} \rightarrow \mathfrak{T}_* \right)} = \frac{n_{j^* \cdot \text{adj}}(\eta^*)}{n_{j \cdot \text{adj}}(\eta)}$$

where $n_{j^* \cdot \text{adj}}(\eta^*)$ is the number of split values available under the proposal tree's splitting rule and $n_{j \cdot \text{adj}}(\eta)$ is the number of split values available under the original tree's splitting rule.

Likelihood Ratio

The proposal tree differs from the original tree only in the two daughter nodes of the selected change node. These two terminal nodes have the unexplained responses apportioned differently. Denote $R_{1.}$ as the residuals of the first daughter node and $R_{2.}$ as the unexplained responses in the second daughter node. Thus we begin with:

$$\frac{\mathbb{P}(\mathbf{R} | \mathbb{F}_*, \sigma^2)}{\mathbb{P}(\mathbf{R} | \mathbb{F}, \sigma^2)} = \frac{\mathbb{P}(R_{1^*,1}, \dots, R_{1^*,n_{1^*}} | \sigma^2) \mathbb{P}(R_{2^*,1}, \dots, R_{2^*,n_{2^*}} | \sigma^2)}{\mathbb{P}(R_{1,1}, \dots, R_{1,n_1} | \sigma^2) \mathbb{P}(R_{2,1}, \dots, R_{2,n_2} | \sigma^2)}$$

where the responses denoted with an asterisk are the responses in the proposal tree's daughter nodes.

Substituting Equation A.16 four times and using algebra, the following expression is obtained for the ratio:

$$\sqrt{\frac{\left(\frac{\sigma^2}{\sigma_\mu^2} + n_1\right) \left(\frac{\sigma^2}{\sigma_\mu^2} + n_2\right)}{\left(\frac{\sigma^2}{\sigma_\mu^2} + n_1^*\right) \left(\frac{\sigma^2}{\sigma_\mu^2} + n_2^*\right)}} \times \exp\left(\frac{1}{2\sigma^2} \left(\frac{(\sum_{i=1}^{n_{1^*}} R_{1^*,i})^2}{n_{1^*} + \frac{\sigma^2}{\sigma_\mu^2}} + \frac{(\sum_{i=1}^{n_{2^*}} R_{2^*,i})^2}{n_{2^*} + \frac{\sigma^2}{\sigma_\mu^2}} - \frac{(\sum_{i=1}^{n_1} R_{1,i})^2}{n_1 + \frac{\sigma^2}{\sigma_\mu^2}} - \frac{(\sum_{i=1}^{n_2} R_{2,i})^2}{n_2 + \frac{\sigma^2}{\sigma_\mu^2}} \right)\right)$$

which simplifies to

$$\exp\left(\frac{1}{2\sigma^2} \left(\frac{(\sum_{i=1}^{n_{1^*}} R_{1^*,i})^2 - (\sum_{i=1}^{n_{1^*}} R_{1,i})^2}{n_1 + \frac{\sigma^2}{\sigma_\mu^2}} + \frac{(\sum_{i=1}^{n_{2^*}} R_{2^*,i})^2 - (\sum_{i=1}^{n_{2^*}} R_{2,i})^2}{n_2 + \frac{\sigma^2}{\sigma_\mu^2}} \right)\right)$$

if the number of responses in the children do not change in the proposal ($n_1 = n_1^*$ and $n_2 = n_2^*$).

Tree Structure Ratio

The proposal tree has the same structure as the original tree. Thus we only need to take into account the changed node's daughters:

$$\frac{\mathbb{P}\left(\frac{\mathbb{P}_*}{\mathbb{P}}\right)}{\mathbb{P}\left(\frac{\mathbb{P}}{\mathbb{P}_*}\right)} = \frac{(1 - \mathbb{P}_{\text{SPLIT}}(\eta_{1*})) (1 - \mathbb{P}_{\text{SPLIT}}(\eta_{2*})) \mathbb{P}_{\text{SPLIT}}(\eta_*) \mathbb{P}_{\text{RULE}}(\eta_*)}{(1 - \mathbb{P}_{\text{SPLIT}}(\eta_1) (1 - \mathbb{P}_{\text{SPLIT}}(\eta_2))) \mathbb{P}_{\text{SPLIT}}(\eta) \mathbb{P}_{\text{RULE}}(\eta)}.$$

The probability of splits remain the same because the daughter nodes are at the same depths. Thus we only need to consider the ratio of the probability of the rules. Once again, the probability of selecting the new value to split on will differ as different split features have different numbers of unique values available. We are left with

$$\frac{\mathbb{P}\left(\frac{\mathbb{P}_*}{\mathbb{P}}\right)}{\mathbb{P}\left(\frac{\mathbb{P}}{\mathbb{P}_*}\right)} = \frac{n_{j \cdot \text{adj}}(\eta)}{n_{j^* \cdot \text{adj}}(\eta^*)}.$$

Note that this is the inverse of the transition ratio. Hence, for the change step, only the likelihood ratio needs to be computed to determine the Metropolis-Hastings ratio r .

A.4.5 Bakeoff

We baked off nine regression data sets and assessed out-of-sample RMSE using 10-fold cross-validation. The results are displayed in Table A.2.

We conclude that the implementation outlined in this paper performs approximately the same as the previous implementation with regards to predictive accuracy.

Dataset Name	bartMachine	BayesTree	randomForest
boston	4.66	4.59	4.58
triazine	0.04*	0.05	0.04
ozone	48.02	47.75	48.77
baseball	0.02	0.02*	0.02
wine.red	0.73	0.69*	0.76
ankara	2.62	2.61	3.46
wine.white	0.49	0.49	0.54
pole	4.00	3.65	11.56
compactiv	46.17*	47.92	40.74

Table A.2: Average of 20 replicates of out-of-sample RMSE values on 9 datasets using three machine learning algorithms. Asterisks indicate a significant difference between `bartMachine` and `BayesTree` at a significance level of 5% with a Bonferroni correction. Comparisons with `randomForest`'s performance were not conducted.

A.5 BART for Panel Data

Sometimes data is collected in a panel, meaning the n observations are collected from w people or “workers” (the usual notation is m but this is being used for number of trees in the BART setup). Many times all observations for a given worker have things in common, *i.e.* they are correlated. The higher the correlation, the closer the effective sample size is to w and not n . The lower the correlation, the closer the sample size is to n and not w . The goal of this project is to create BART models that can account for these correlations.

A.5.1 The Updated Model

Denote n_1, n_2, \dots, n_w as the number of observations collected from all w workers ($n = \sum_{q=1}^w n_q$). We further assume that each worker’s submissions are exchangeable meaning there is no time dependency. And further, we assume each worker *shares* the same correlation. The BART model now becomes:

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon} \approx \mathfrak{F}_1^{\otimes}(\mathbf{X}) + \mathfrak{F}_2^{\otimes}(\mathbf{X}) + \dots + \mathfrak{F}_m^{\otimes}(\mathbf{X}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{B}) \quad (\text{A.5})$$

where \mathbf{X} is the $n \times p$ design matrix *sorted* by from the first worker’s contributions to the last worker’s contribution and \mathbf{B} is a block diagonal matrix with the blocks being intraclass correlation matrices of dimensions n_1, n_2, \dots, n_w :

$$\mathbf{B} := \begin{bmatrix} \mathbf{D}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D}_{n_w} \end{bmatrix} \quad \text{where}$$

$$\mathbf{D}_{n_q} := \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}_{n_q} = \rho(\mathbf{J}_{n_q} - \mathbf{I}_{n_q}) + \mathbf{I}_{n_q} \quad (\text{A.6})$$

Thus, we need to just update the Gibbs sampler to sample for ρ and update the likelihood calculations.

A.5.2 The Update Metropolis-within-Gibbs Sampler

The new Gibbs sampler is below and we spend the rest of the section going over each piece that has changed.

$$\begin{aligned} 1: & \quad \mathfrak{F}_1 \mid \mathbf{R}_{-1}, \sigma^2, \rho & (\text{A.7}) \\ 2: & \quad \mathcal{E}_1 \mid \mathfrak{F}_1, \mathbf{R}_{-1}, \sigma^2, \rho \\ 3: & \quad \mathfrak{F}_2 \mid \mathbf{R}_{-2}, \sigma^2, \rho \\ 4: & \quad \mathcal{E}_2 \mid \mathfrak{F}_2, \mathbf{R}_{-2}, \sigma^2, \rho \\ & \quad \vdots \\ 2m-1: & \quad \mathfrak{F}_m \mid \mathbf{R}_{-m}, \sigma^2, \rho \\ 2m: & \quad \mathcal{E}_m \mid \mathfrak{F}_m, \mathbf{R}_{-m}, \sigma^2, \rho \\ 2m+1: & \quad \sigma^2 \mid \mathfrak{F}_1, \mathcal{E}_1, \dots, \mathfrak{F}_m, \mathcal{E}_m, \mathbf{E}, \rho \\ 2m+2: & \quad \rho \mid \mathfrak{F}_1, \mathcal{E}_1, \dots, \mathfrak{F}_m, \mathcal{E}_m, \mathbf{E}, \sigma^2 \end{aligned}$$

We now tackle each posterior sampling in its own subsection.

A.5.2.1 Posterior Sampling of ϑ_t

The posterior $\vartheta_t \mid \mathfrak{F}_t, \mathbf{R}_{-t}, \sigma^2, \rho$ is actually a compound step sampling μ_ℓ for all $\ell = 1, \dots, b$ terminal nodes in the tree, $\mu_\ell \mid \mathbf{R}_{-t,\ell}, \sigma^2, \rho$. In the standard Bayesian setup, we have:

$$\mathbb{P}(\mu_\ell \mid \mathbf{R}_{-t,\ell}, \sigma^2, \rho) \propto \underbrace{\mathbb{P}(\mathbf{R}_{-t,\ell} \mid \mu_\ell, \sigma^2, \rho)}_{\text{likelihood}} \underbrace{\mathbb{P}(\mu_\ell; \sigma_\mu^2)}_{\text{prior}}$$

The likelihood is multivariate normal sharing the same mean and the prior is univariate normal. Let $\boldsymbol{\mu}_\ell := \mu_\ell \mathbf{1}_{n_\ell}$

$$\begin{aligned} \mathbb{P}(\mu_\ell \mid \mathbf{R}_{-t,\ell}, \sigma^2, \rho) &\propto \mathcal{N}_{n_\ell}(\boldsymbol{\mu}_\ell, \sigma^2 \mathbf{B}_\ell) \mathcal{N}(0, \sigma_\mu^2) \\ &= \frac{1}{(2\pi)^{n_\ell/2} |\sigma^2 \mathbf{B}_\ell|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{R}_{-t,\ell} - \boldsymbol{\mu}_\ell)^\top (\sigma^2 \mathbf{B}_\ell)^{-1} (\mathbf{R}_{-t,\ell} - \boldsymbol{\mu}_\ell)\right) \times \\ &\quad \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left(-\frac{1}{2\sigma_\mu^2} \mu_\ell^2\right) \\ &\propto \exp\left(-\frac{1}{2} \left(\frac{1}{\sigma^2} \underbrace{(\mathbf{R}_{-t,\ell} - \boldsymbol{\mu}_\ell)^\top \mathbf{B}_\ell^{-1} (\mathbf{R}_{-t,\ell} - \boldsymbol{\mu}_\ell)} + \frac{1}{\sigma_\mu^2} \mu_\ell^2 \right)\right) \end{aligned}$$

The underbraced above is a quadratic form which can be written as:

$$(\mathbf{R}_{-t,\ell} - \boldsymbol{\mu}_\ell)^\top \mathbf{B}_\ell^{-1} (\mathbf{R}_{-t,\ell} - \boldsymbol{\mu}_\ell) = \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} (r_i - \mu_\ell)(r_j - \mu_\ell)$$

where $b_{\ell,ij}^{-1}$ is the i, j th entry of \mathbf{B}_ℓ^{-1} and r_i is the i th entry of $\mathbf{R}_{-t,\ell}$. Rewriting everything inside the overbraced terms, we find:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} r_i r_j - \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} r_i \mu_\ell - \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} r_j \mu_\ell + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} \mu_\ell^2 + \frac{1}{\sigma_\mu^2} \mu_\ell^2$$

We now collect terms to arrive at the coefficients for a quadratic equation:

$$\begin{aligned} & \underbrace{\left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} \right)}_a \mu_\ell^2 + \underbrace{\left(-\frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} (r_i + r_j) \right)}_b \mu_\ell \\ & + \underbrace{\frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} r_i r_j}_c \end{aligned} \quad (\text{A.8})$$

We can safely ignore c since within the proportionality of the posterior, this term does not depend on our parameter of interest μ_ℓ . In order to complete the square, we need the above form to look like $d(\mu_\ell - e)^2$ which means $d = a$ and $e = -\frac{b}{2a}$ (the d^2 term can be safely ignored since it is not a function of μ_ℓ and disappears as a proportionality constant). Thus, we arrive at:

$$\begin{aligned} \mathbb{P}(\mu_\ell \mid \mathbf{R}_{-t,\ell}, \sigma^2, \rho) & \propto \exp\left(-\frac{1}{2}a \left(\mu_\ell - \left(-\frac{b}{2a}\right)\right)^2\right) \propto \mathcal{N}\left(-\frac{b}{2a}, \frac{1}{a}\right) \\ & = \mathcal{N}\left(\frac{\frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} (r_i + r_j)}{2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1}\right)}, \frac{1}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1}}}\right) \end{aligned}$$

We can further simplify this expression. If we organize the elements of $\mathbf{R}_{-t,\ell}$ by worker, then since all workers are independent, we can add their contributions together because the elements of \mathbf{B}^{-1} are zeroes outside of a common worker's rows and columns.

$$\mu_\ell \mid \mathbf{R}_{-t,\ell}, \sigma^2, \rho \sim \mathcal{N} \left(\frac{\frac{1}{\sigma^2} \sum_{q=1}^w \sum_{i=1}^{n_{\ell q}} \sum_{j=1}^{n_{\ell q}} d_{\ell q,ij}^{-1} (r_i + r_j)}{2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^w \sum_{i=1}^{n_{\ell q}} \sum_{j=1}^{n_{\ell q}} d_{\ell q,ij}^{-1} \right)}, \frac{1}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^w \sum_{i=1}^{n_{\ell q}} \sum_{j=1}^{n_{\ell q}} d_{\ell q,ij}^{-1}} \right) \quad (\text{A.9})$$

The notation $n_{\ell q}$ refers to the number of the q th worker in the ℓ th node. The notation $d_{\ell q,ij}^{-1}$ is the i, j th term of the q th diagonal matrix in \mathbf{B} (see Equation A.6). Each of the matrices \mathbf{D} is an ‘‘equicorrelation matrix’’ whose inverse is:

$$\begin{aligned} \mathbf{D}_{n_{\ell q}}^{-1} &= \frac{1}{1-\rho} \mathbf{I}_{n_{\ell q}} + \frac{\rho}{(1-\rho)(1+(n_{\ell q}-1)\rho)} \mathbf{J}_{n_{\ell q}} \\ \Rightarrow d_{\ell q,ij}^{-1} &= \begin{cases} -\frac{\rho}{(1-\rho)(1+(n_{\ell q}-1)\rho)} + \frac{1}{1-\rho} & \text{when } i=j \\ -\frac{\rho}{(1-\rho)(1+(n_{\ell q}-1)\rho)} & \text{when } i \neq j \end{cases} \end{aligned}$$

Now we can simplify some of the terms in the normal notation:

$$\begin{aligned} \sum_{i=1}^{n_{\ell q}} \sum_{j=1}^{n_{\ell q}} d_{\ell q,ij}^{-1} &= n_{\ell q} \left(\frac{1}{1-\rho} \right) + n_{\ell q}^2 \left(-\frac{\rho}{(1-\rho)(1+(n_{\ell q}-1)\rho)} \right) \\ &= \frac{n_{\ell q}}{1-\rho} \left(1 - \frac{n_{\ell q}\rho}{1+(n_{\ell q}-1)\rho} \right) \\ &= \frac{n_{\ell q}}{1-\rho} \left(\frac{1+(n_{\ell q}-1)\rho}{1+(n_{\ell q}-1)\rho} - \frac{n_{\ell q}\rho}{1+(n_{\ell q}-1)\rho} \right) \\ &= \frac{n_{\ell q}}{1-\rho} \frac{1-\rho}{1+(n_{\ell q}-1)\rho} \\ &= \frac{n_{\ell q}}{1+(n_{\ell q}-1)\rho} \end{aligned} \quad (\text{A.10})$$

Now we consider the term $\sum_{i=1}^{n_{\ell q}} \sum_{j=1}^{n_{\ell q}} d_{\ell q,ij}^{-1} (r_i + r_j)$. On the diagonal we have:

$$\begin{aligned}
& \sum_{i=1}^n \left(\frac{1}{1-\rho} - \frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) (r_i + r_i) \\
&= 2n_{\ell_q} \left(\frac{1}{1-\rho} - \frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) \bar{r}_q
\end{aligned}$$

On the off-diagonal we have:

$$\sum_{i \neq j} -\frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} (r_i + r_j) = -2 \left(\frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) (n_{\ell_q}-1)n_{\ell_q}\bar{r}_q$$

Adding these two together we have:

$$\begin{aligned}
& n_{\ell_q} \left(\frac{1}{1-\rho} - \frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) \bar{r}_q - \\
& 2 \left(\frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) (n_{\ell_q}-1)n_{\ell_q}\bar{r}_q \\
&= 2n_{\ell_q}\bar{r}_q \left(\frac{1}{1-\rho} - \frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} - (n_{\ell_q}-1) \frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) \\
&= 2n_{\ell_q}\bar{r}_q \left(\frac{1}{1-\rho} - n_{\ell_q} \frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) \\
&= 2n_{\ell_q}\bar{r}_q \frac{1}{1-\rho} \left(1 - n_{\ell_q} \frac{\rho}{1+(n_{\ell_q}-1)\rho} \right) \\
&= 2n_{\ell_q}\bar{r}_q \frac{1}{1-\rho} \left(\frac{1-\rho}{1+(n_{\ell_q}-1)\rho} \right) \\
&= \frac{2n_{\ell_q}\bar{r}_q}{1+(n_{\ell_q}-1)\rho} = \frac{2 \sum_{i=1}^{n_{\ell_q}} r_i}{1+(n_{\ell_q}-1)\rho} \tag{A.11}
\end{aligned}$$

Thus, our posterior sampling becomes:

$$\mathbb{P}(\mu_\ell | \mathbf{R}_{-t,\ell}, \sigma^2, \rho) = \mathcal{N} \left(\frac{\frac{1}{\sigma^2} \sum_{q=1}^w \frac{\sum_{i=1}^{n_{\ell_q}} r_i}{1 + (n_{\ell_q} - 1)\rho}}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^w \frac{n_{\ell_q}}{1 + (n_{\ell_q} - 1)\rho}}, \frac{1}{\frac{1}{\sigma_\mu^2} + \sum_{q=1}^w \frac{n_{\ell_q}}{1 + (n_{\ell_q} - 1)\rho}} \right)$$

A.5.2.2 Posterior Sampling of σ^2

The standard Bayes Rule gives us:

$$\begin{aligned} \mathbb{P}(\sigma^2 | \mathbf{E}, \rho) &\propto \underbrace{\mathbb{P}(\mathbf{E} | \sigma^2, \rho)}_{\text{likelihood}} \underbrace{\mathbb{P}(\sigma^2; \nu, \lambda)}_{\text{prior}} \\ &= \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{B}) \text{InvGamma} \left(\frac{\nu}{2}, \frac{\nu \lambda}{2} \right) \\ &= \frac{1}{(2\pi)^{n/2} |\sigma^2 \mathbf{B}|^{1/2}} \exp \left(-\frac{1}{2} \mathbf{E}^\top (\sigma^2 \mathbf{B})^{-1} \mathbf{E} \right) \times \\ &\quad \frac{\left(\frac{\nu \lambda}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp \left(-\frac{\nu \lambda}{2\sigma^2} \right) \\ &\propto \frac{1}{(\sigma^2)^{n/2}} (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp \left(-\frac{1}{2\sigma^2} (\mathbf{E}^\top \mathbf{B}^{-1} \mathbf{E} + \nu \lambda) \right) \\ &= (\sigma^2)^{-(\frac{\nu+n}{2}+1)} \exp \left(-\frac{\frac{1}{2} (\mathbf{E}^\top \mathbf{B}^{-1} \mathbf{E} + \nu \lambda)}{\sigma^2} \right) \\ &= \text{InvGamma} \left(\frac{\nu + n}{2}, \frac{\mathbf{E}^\top \mathbf{B}^{-1} \mathbf{E} + \nu \lambda}{2} \right) \end{aligned}$$

Now, we can simplify the quadratic form above. Ordering \mathbf{E} by worker and using the fact that \mathbf{B} is diagonal,

$$\mathbf{E}^\top \mathbf{B}^{-1} \mathbf{E} = \sum_{q=1}^w \mathbf{E}_q^\top \mathbf{D}_q^{-1} \mathbf{E}_q = \sum_{q=1}^w \underbrace{\sum_{i=1}^{n_{\ell_q}} \sum_{j=1}^{n_{\ell_q}} d_{\ell_q, ij}^{-1} e_{q,i} e_{q,j}}_{}$$

where \mathbf{E}_q is the error vector for worker q . The underbraced above becomes:

$$\begin{aligned}
& \left(\frac{1}{1-\rho} - \frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) \sum_{i=1}^{n_{\ell_q}} e_i^2 - \\
& \frac{\rho}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \sum_{i=1}^{n_{\ell_q}} \sum_{j=1}^{n_{\ell_q}} e_{q,i} e_{q,j} \mathbb{1}_{i \neq j} \\
= & \frac{1}{1-\rho} \left(1 - \frac{\rho}{1+(n_{\ell_q}-1)\rho} \sum_{i=1}^{n_{\ell_q}} e_i^2 - \frac{\rho}{1+(n_{\ell_q}-1)\rho} \sum_{i=1}^{n_{\ell_q}} \sum_{j=1}^{n_{\ell_q}} e_{q,i} e_{q,j} \mathbb{1}_{i \neq j} \right) \\
= & \frac{1}{1-\rho} \left(\frac{1+n_{\ell_q}\rho-2\rho}{1+(n_{\ell_q}-1)\rho} \sum_{i=1}^{n_{\ell_q}} e_i^2 - \frac{\rho}{1+(n_{\ell_q}-1)\rho} \sum_{i=1}^{n_{\ell_q}} \sum_{j=1}^{n_{\ell_q}} e_{q,i} e_{q,j} \mathbb{1}_{i \neq j} \right) \\
= & \frac{1}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \left(\underbrace{(1+n_{\ell_q}\rho-2\rho)}_{\text{SSE}_q} \sum_{i=1}^{n_{\ell_q}} e_i^2 - \rho \underbrace{\sum_{i=1}^{n_{\ell_q}} \sum_{j=1}^{n_{\ell_q}} e_{q,i} e_{q,j} \mathbb{1}_{i \neq j}}_{\text{CPE}_q} \right) \\
= & \frac{1}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \left((1+n_{\ell_q}\rho-2\rho)SSE_q - \rho CPE_q \right) \quad (\text{A.12})
\end{aligned}$$

where SSE_q is the sum of squared error in worker q and CPE_q is the sum of the cross products in worker q save the diagonal elements.

Thus, the sampling for σ^2 becomes:

$$\mathbb{P}(\sigma^2 | \mathbf{E}, \rho) = \text{InvGamma} \left(\frac{\nu+n}{2}, \frac{\nu\lambda}{2} + \left(\frac{1}{2(1-\rho)} \sum_{q=1}^w \frac{(1+n_{\ell_q}\rho-2\rho)SSE_q - \rho CPE_q}{1+(n_{\ell_q}-1)\rho} \right) \right)$$

A.5.2.3 Posterior Sampling of ρ

The standard Bayes Rule gives us:

$$\mathbb{P}(\rho | \mathbf{E}, \sigma^2) \propto \underbrace{\mathbb{P}(\mathbf{E} | \sigma^2, \rho)}_{\text{likelihood}} \underbrace{\mathbb{P}(\rho)}_{\text{prior}}$$

We now assume the prior is:

$$\mathbb{P}(\rho) = U(-1, 1)$$

This leaves us with:

$$\begin{aligned} \mathbb{P}(\rho \mid \mathbf{E}, \sigma^2) &\propto \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{B}) U(-1, 1) \\ &\propto \frac{1}{(2\pi)^{n/2} |\sigma^2 \mathbf{B}|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{E}^\top (\sigma^2 \mathbf{B})^{-1} \mathbf{E}\right) \quad (1) \\ &\propto \frac{1}{|\mathbf{B}|^{1/2}} \exp\left(-\frac{1}{2\sigma^2} \mathbf{E}^\top \mathbf{B}^{-1} \mathbf{E}\right) \end{aligned}$$

The determinant term above becomes:

$$\begin{aligned} |\mathbf{B}_\ell|^{1/2} &= (|\mathbf{D}_1| \cdot \dots \cdot |\mathbf{D}_{w_\ell}|)^{1/2} \\ &= \left(((n_1 - 1)\rho + 1)(1 - \rho)^{n_1 - 1} \cdot \dots \cdot ((n_{w_\ell} - 1)\rho + 1)(1 - \rho)^{n_{w_\ell} - 1} \right)^{1/2} \\ &= \left(\prod_{q=1}^{w_\ell} ((n_q - 1)\rho + 1)(1 - \rho)^{n_q - 1} \right)^{\frac{1}{2}} \\ &= \left((1 - \rho)^{n_\ell - w_\ell} \prod_{q=1}^{w_\ell} ((n_q - 1)\rho + 1) \right)^{\frac{1}{2}} \quad (\text{A.13}) \end{aligned}$$

Using the formula we found before, we have:

$$\begin{aligned} \mathbb{P}(\rho \mid \mathbf{E}, \sigma^2) &\propto (1 - \rho)^{-\frac{n-w}{2}} \left(\prod_{q=1}^w (n_q - 1)\rho + 1 \right)^{-\frac{1}{2}} \times \\ &\exp\left(-\frac{1}{2\sigma^2(1 - \rho)(1 + (n_{\ell q} - 1)\rho)} \sum_{q=1}^w ((1 + n_{\ell q}\rho - 2\rho)SSE_q - \rho CPE_q)\right) \end{aligned}$$

This is clearly not a distribution we can directly sample from. Thus, we now use a Metropolis-Hastings step which can be accepted or rejected. Denote ρ_* as the proposed value of ρ .

$$r = \frac{\mathbb{P}(\rho_* \rightarrow \rho) \mathbb{P}(\rho_* | \mathbf{E}, \sigma^2)}{\mathbb{P}(\rho \rightarrow \rho_*) \mathbb{P}(\rho | \mathbf{E}, \sigma^2)} \quad (\text{A.14})$$

We will propose a very simple jump distribution as $\mathbb{P}(\rho \rightarrow \rho_*) = U(-1, 1)$. The risk is a sampler that converges slowly. We will revisit this jump distribution later if the proportion of acceptances is too low. Since the jump ratio cancels, we are left with:

$$\begin{aligned} r &= \frac{\mathbb{P}(\rho_* | \mathbf{E}, \sigma^2)}{\mathbb{P}(\rho | \mathbf{E}, \sigma^2)} \\ &= \sqrt{\left(\frac{1-\rho}{1-\rho_*} \right)^{n-w} \prod_{q=1}^w \frac{(n_q-1)\rho+1}{(n_q-1)\rho_*+1} \exp\left(-\frac{1}{2\sigma^2} \left(\left(\frac{1}{(1-\rho_*)(1+(n_{\ell_q}-1)\rho_*)} \right) \sum_{q=1}^w ((1+n_{\ell_q}\rho_*-2\rho_*)SSE_q - \rho_*CPE_q) + \left(\frac{1}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) \sum_{q=1}^w ((1+n_{\ell_q}\rho-2\rho)SSE_q - \rho CPE_q) \right) \right)} \end{aligned}$$

This in log form becomes

$$\begin{aligned} \ln(r) &= \frac{1}{2} \left((n-w) \ln\left(\frac{1-\rho}{1-\rho_*} \right) + \sum_{q=1}^w \ln\left(\frac{(n_q-1)\rho+1}{(n_q-1)\rho_*+1} \right) \right) - \frac{1}{2\sigma^2} \left(\left(\frac{1}{(1-\rho_*)(1+(n_{\ell_q}-1)\rho_*)} \right) \sum_{q=1}^w ((1+n_{\ell_q}\rho_*-2\rho_*)SSE_q - \rho_*CPE_q) + \left(\frac{1}{(1-\rho)(1+(n_{\ell_q}-1)\rho)} \right) \sum_{q=1}^w ((1+n_{\ell_q}\rho-2\rho)SSE_q - \rho CPE_q) \right) \end{aligned}$$

A.5.3 Posterior Sampling of \mathfrak{T}_t

We now handle the Metropolis steps $1, 3, \dots, 2m - 1$ for all three tree proposal steps.

Below is the Metropolis ratio where the parameter sampled is the tree and the data is the responses unexplained by other trees denoted by \mathbf{R} . We denote the new, proposal tree with an asterisk and the original tree without the asterisk.

$$r = \frac{\mathbb{P}(\mathfrak{T}_* \rightarrow \mathfrak{T}) \mathbb{P}(\mathfrak{T}_* | \mathbf{R}, \sigma^2, \rho)}{\mathbb{P}(\mathfrak{T} \rightarrow \mathfrak{T}_*) \mathbb{P}(\mathfrak{T} | \mathbf{R}, \sigma^2, \rho)} \quad (\text{A.15})$$

We accept a draw from the posterior distribution of trees if a draw from a standard uniform distribution is less than the value of r . Immediately we note that it is difficult (if not impossible) to calculate the posterior probabilities for the trees themselves. Instead, we employ Bayes' Rule,

$$\mathbb{P}(\mathfrak{T} | \mathbf{R}, \sigma^2) = \frac{\mathbb{P}(\mathbf{R} | \mathfrak{T}, \sigma^2, \rho) \mathbb{P}(\mathfrak{T} | \sigma^2, \rho)}{\mathbb{P}(\mathbf{R} | \sigma^2, \rho)},$$

and plug the result into Equation A.15 to obtain:

$$r = \underbrace{\frac{\mathbb{P}(\mathfrak{T}_* \rightarrow \mathfrak{T})}{\mathbb{P}(\mathfrak{T} \rightarrow \mathfrak{T}_*)}}_{\text{transition ratio}} \times \underbrace{\frac{\mathbb{P}(\mathbf{R} | \mathfrak{T}_*, \sigma^2, \rho)}{\mathbb{P}(\mathbf{R} | \mathfrak{T}, \sigma^2, \rho)}}_{\text{likelihood ratio}} \times \underbrace{\frac{\mathbb{P}(\mathfrak{T}_*)}{\mathbb{P}(\mathfrak{T})}}_{\text{tree structure ratio}}.$$

Note that the probability of the tree structure is independent of σ^2 and ρ .

The goal of this section is to explicitly calculate r for all possible tree proposals — GROW, PRUNE and CHANGE. Note that the transition ratio and tree structure ratios are the same between iidBART and panel data BART and these expressions are explained in the arxiv document. Thus, we only talk derive the likelihood ratios

here.

Note that our actual implementation uses the following expressions in log form for numerical accuracy.

A.5.3.1 Grow Proposal

To calculate the likelihood, the tree structure determines which responses fall into which of the b terminal nodes. Thus,

$$\mathbb{P}\left(R_1, \dots, R_n \mid \mathbb{F}, \sigma^2, \rho\right) = \prod_{\ell=1}^b \mathbb{P}\left(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} \mid \sigma^2, \rho\right)$$

where each term on the right hand side is the probability of responses in one of the b terminal nodes, which are independent by assumption. The R_ℓ 's denote the data in the ℓ th terminal node and where n_ℓ denotes how many observations are in each terminal node and $n = \sum_{\ell=1}^b n_\ell$.

We now find an analytic expression for the node likelihood term. Remember, if the mean in each terminal node, which we denote μ_ℓ , was known, then we would have $R_{\ell_1}, \dots, R_{\ell_{n_\ell}} \mid \mu_\ell, \sigma^2 \stackrel{iid}{\sim} \mathcal{N}(\mu_\ell, \sigma^2)$. BART requires μ_ℓ to be margined out, allowing the Gibbs sampler in Equation A.7 to avoid dealing with jumping between continuous spaces of varying dimensions (Chipman et al., 2010, page 275). Recall that one of the BART model assumptions is a prior on the average value of $\mu \sim \mathcal{N}(0, \sigma_\mu^2)$ and thus,

$$\mathbb{P}\left(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} \mid \sigma^2, \rho\right) = \int_{\mathbb{R}} \mathbb{P}\left(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} \mid \mu_\ell, \sigma^2, \rho\right) \mathbb{P}\left(\mu_\ell; \sigma_\mu^2\right) d\mu_\ell$$

Substituting, we find

$$\frac{1}{((2\pi)^{n_\ell+1} (\sigma^2)^{n_\ell} |\mathbf{B}| \sigma_\mu^2)^{\frac{1}{2}}} \times \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} (\mathbf{R}_{-t,\ell} - \boldsymbol{\mu}_\ell)^\top \mathbf{B}_\ell^{-1} (\mathbf{R}_{-t,\ell} - \boldsymbol{\mu}_\ell) + \frac{1}{\sigma_\mu^2} \mu_\ell^2\right)\right) d\mu_\ell$$

we need to complete the square and account for the constant additive term:

$$a\mu^2 + b\mu + c = d(\mu - e)^2 + f \quad \Rightarrow \quad d = a \quad \text{and} \quad e = -\frac{b}{2a} \quad \text{and} \quad f = c - \frac{b^2}{4a}$$

Thus the integral above using using Equation A.8 should become

$$\exp(f) \int_{\mathbb{R}} \exp(d(\mu - e)^2) d\mu_\ell$$

Substituting, we find:

$$\exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} r_i r_j - \frac{\left(\sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} (r_i + r_j) \right)^2}{4\sigma^2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} \right)} \right)\right) \times \int_{\mathbb{R}} \exp\left(-\frac{1}{2} \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} \right) \left(\mu_\ell - \frac{\frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} (r_i + r_j)}{2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} \right)} \right)^2\right) d\mu_\ell$$

The above integral is a Gaussian integral which resolves to $\sqrt{2\pi\text{var}}$ where “var” is the variance parameter. Thus the margined likelihood becomes:

$$\begin{aligned}
& \mathbb{P}(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} \mid \sigma^2, \rho) \\
&= \frac{1}{((2\pi)^{n_\ell+1} (\sigma^2)^{n_\ell} |\mathbf{B}_\ell| \sigma_\mu^2)^{\frac{1}{2}}} \sqrt{\frac{2\pi}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1}}} \times \\
& \exp \left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} r_i r_j - \frac{\left(\sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} (r_i + r_j) \right)^2}{4\sigma^2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} \right)} \right) \right) \\
&= \left((2\pi)^{n_\ell} (\sigma^2)^{n_\ell} |\mathbf{B}_\ell| \sigma_\mu^2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} \right) \right)^{-\frac{1}{2}} \times \\
& \exp \left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} r_i r_j - \frac{\left(\sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} (r_i + r_j) \right)^2}{4\sigma^2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} \right)} \right) \right) \quad (\text{A.16})
\end{aligned}$$

Since the likelihoods are solely determined by the terminal nodes, the proposal tree differs from the original tree by only the selected node to be grown, denoted by ℓ , which becomes two daughters after the GROW step denoted by ℓ_L and ℓ_R . Hence, the likelihood ratio becomes:

$$\frac{\mathbb{P}(\mathbf{R} \mid \mathfrak{F}_*, \sigma^2)}{\mathbb{P}(\mathbf{R} \mid \mathfrak{F}, \sigma^2)} = \frac{\mathbb{P}(R_{\ell_{L,1}}, \dots, R_{\ell_{L,n_{\ell,L}}} \mid \sigma^2) \mathbb{P}(R_{\ell_{R,1}}, \dots, R_{\ell_{R,n_{\ell,R}}} \mid \sigma^2)}{\mathbb{P}(R_{\ell_1}, \dots, R_{\ell_{n_\ell}} \mid \sigma^2)} \quad (\text{A.17})$$

Plugging Equation A.16 into Equation A.17 three times yields the ratio for the GROW step. We begin with the non-exponential term. We can reduce it to

$$\sqrt{\frac{|B_\ell|}{\underbrace{|B_{\ell L}| |B_{\ell R}|}} \frac{\left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1}\right)}{\left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} b_{\ell L,ij}^{-1}\right) \left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{i=1}^{n_{\ell R}} \sum_{j=1}^{n_{\ell R}} b_{\ell R,ij}^{-1}\right)}} \quad (\text{A.18})$$

where $n_{\ell L}$ and $n_{\ell R}$ denote the number of data points in the newly grown left and right daughter nodes.

We can now use the determinant formula of Equation A.13 to simplify the underbraced term above:

$$(1 - \rho)^{w_{\ell L} + w_{\ell R} - w_\ell} \prod_{q=1}^w \frac{((n_{\ell q} - 1) \rho + 1)}{\left((n_{\ell L q} - 1) \rho + 1\right) \left((n_{\ell R q} - 1) \rho + 1\right)} \quad (\text{A.19})$$

Since the B_ℓ 's are diagonal, the second term can be rewritten as sums over the workers from Equation A.9:

$$\frac{\left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{q=1}^w \sum_{i=1}^{n_{\ell q}} \sum_{j=1}^{n_{\ell q}} d_{\ell q,ij}^{-1}\right)}{\left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{q=1}^w \sum_{i=1}^{n_{\ell L q}} \sum_{j=1}^{n_{\ell L q}} d_{\ell L q,ij}^{-1}\right) \left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{q=1}^w \sum_{i=1}^{n_{\ell R q}} \sum_{j=1}^{n_{\ell R q}} d_{\ell R q,ij}^{-1}\right)}$$

Now we have expressions for the $d_{\ell q,ij}^{-1}$ terms from Equation A.10 which we substitute in to become:

$$\begin{aligned}
& \left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{q=1}^{w_\ell} \frac{n_{\ell q}}{1 + (n_{\ell q} - 1)\rho} \right) \\
& \frac{\left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{q=1}^{w_{L}} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho} \right) \left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{q=1}^{w_{R}} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho} \right)}{\sigma^2 \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_\ell} \frac{n_{\ell q}}{1 + (n_{\ell q} - 1)\rho} \right)} \\
= & \frac{\left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{L}} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho} \right) \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{R}} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho} \right)}{\left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{L}} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho} \right) \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{R}} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho} \right)} \quad (\text{A.20})
\end{aligned}$$

Thus, the whole term becomes:

$$\begin{aligned}
& \sqrt{\sigma^2 (1 - \rho)^{w_L + w_R - w_\ell} \left(\prod_{q=1}^w \frac{(n_{\ell q} - 1)\rho + 1}{\left((n_{Lq} - 1)\rho + 1 \right) \left((n_{Rq} - 1)\rho + 1 \right)} \right)} \times \\
& \sqrt{\frac{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_\ell} \frac{n_{\ell q}}{1 + (n_{\ell q} - 1)\rho}}{\left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_L} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho} \right) \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_R} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho} \right)}}}
\end{aligned}$$

Now we examine the portion of the exponential term and we find:

$$\begin{aligned}
& \exp \left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_{L}} \sum_{j=1}^{n_{L}} b_{\ell_L, ij}^{-1} r_i r_j - \frac{\left(\sum_{i=1}^{n_{L}} \sum_{j=1}^{n_{L}} b_{\ell_L, ij}^{-1} (r_i + r_j) \right)^2}{4\sigma^2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_{L}} \sum_{j=1}^{n_{L}} b_{\ell_L, ij}^{-1} \right)} \right) \right) \times \\
& \exp \left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_{R}} \sum_{j=1}^{n_{R}} b_{\ell_R, ij}^{-1} r_i r_j - \frac{\left(\sum_{i=1}^{n_{R}} \sum_{j=1}^{n_{R}} b_{\ell_R, ij}^{-1} (r_i + r_j) \right)^2}{4\sigma^2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_{R}} \sum_{j=1}^{n_{R}} b_{\ell_R, ij}^{-1} \right)} \right) \right) \times
\end{aligned}$$

$$\exp \left(\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} r_i r_j - \frac{\left(\sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} (r_i + r_j) \right)^2}{4\sigma^2 \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} \right)} \right) \right)$$

Collecting terms inside the exponential and dropping the exponential notation and the $-\frac{1}{2\sigma^2}$ term, we arrive at:

$$\begin{aligned} & \sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} b_{\ell L,ij}^{-1} r_i r_j + \sum_{i=1}^{n_{\ell R}} \sum_{j=1}^{n_{\ell R}} b_{\ell R,ij}^{-1} r_i r_j - \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{\ell,ij}^{-1} r_i r_j + \\ & \frac{1}{4\sigma^2} \left(\frac{\left(\sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} b_{\ell L,ij}^{-1} (r_i + r_j) \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} b_{\ell L,ij}^{-1}} - \frac{\left(\sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} b_{\ell L,ij}^{-1} (r_i + r_j) \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} b_{\ell L,ij}^{-1}} - \frac{\left(\sum_{i=1}^{n_{\ell R}} \sum_{j=1}^{n_{\ell R}} b_{\ell R,ij}^{-1} (r_i + r_j) \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_{\ell R}} \sum_{j=1}^{n_{\ell R}} b_{\ell R,ij}^{-1}} \right) \end{aligned}$$

Since the \mathbf{B}_ℓ 's are diagonal, all terms can be rewritten as sums over the workers from Equation A.9:

$$\begin{aligned} & \sum_{q=1}^{w_\ell} \sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} d_{\ell Lq,ij}^{-1} r_i r_j + \sum_{q=1}^{w_\ell} \sum_{i=1}^{n_{\ell R}} \sum_{j=1}^{n_{\ell R}} d_{\ell Rq,ij}^{-1} r_i r_j - \sum_{q=1}^{w_\ell} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} d_{\ell q,ij}^{-1} r_i r_j + \tag{A.21} \\ & \frac{1}{4\sigma^2} \left(\frac{\left(\sum_{q=1}^{w_\ell} \sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} d_{\ell Lq,ij}^{-1} (r_i + r_j) \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_\ell} \sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} d_{\ell Lq,ij}^{-1}} - \frac{\left(\sum_{q=1}^{w_\ell} \sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} d_{\ell Lq,ij}^{-1} (r_i + r_j) \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_\ell} \sum_{i=1}^{n_{\ell L}} \sum_{j=1}^{n_{\ell L}} d_{\ell Lq,ij}^{-1}} - \frac{\left(\sum_{q=1}^{w_\ell} \sum_{i=1}^{n_{\ell R}} \sum_{j=1}^{n_{\ell R}} d_{\ell Rq,ij}^{-1} (r_i + r_j) \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_\ell} \sum_{i=1}^{n_{\ell R}} \sum_{j=1}^{n_{\ell R}} d_{\ell Rq,ij}^{-1}} \right) \end{aligned}$$

Now we can use the simplification found in Equation A.12 to reduce the first line in the above expression to:

$$\begin{aligned} & \frac{1}{1-\rho} \left(\sum_{q=1}^{w_{\ell L}} \frac{1}{1+(n_{\ell Lq}-1)\rho} \left((1+n_{\ell Lq}\rho-2\rho)SSR_{Lq} - \rho CPR_{Lq} \right) \right. \\ & \left. + \sum_{q=1}^{w_{\ell R}} \frac{1}{1+(n_{\ell Rq}-1)\rho} \left((1+n_{\ell Rq}\rho-2\rho)SSR_{Rq} - \rho CPR_{Rq} \right) \right) \end{aligned}$$

$$- \sum_{q=1}^{w_\ell} \frac{1}{1 + (n_{\ell q} - 1)\rho} \left((1 + n_{\ell q}\rho - 2\rho)SSR_q - \rho CPR_q \right) \quad (\text{A.22})$$

Now we can use simplifications found in Equations A.11 and A.10 to reduce the second line of the above expression to (and then we further simplify it):

$$\begin{aligned} & \frac{1}{4\sigma^2} \left(\frac{\left(\sum_{q=1}^{w_\ell} \frac{2 \sum_{i=1}^{n_{\ell q}} r_i}{1 + (n_{\ell q} - 1)\rho} \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_\ell} \frac{n_{\ell q}}{1 + (n_{\ell q} - 1)\rho}} - \frac{\left(\sum_{q=1}^{w_{Lq}} \frac{2 \sum_{i=1}^{n_{Lq}} r_i}{1 + (n_{Lq} - 1)\rho} \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_{Lq}} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho}} - \frac{\left(\sum_{q=1}^{w_{Rq}} \frac{2 \sum_{i=1}^{n_{Rq}} r_i}{1 + (n_{Rq} - 1)\rho} \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_{Rq}} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho}} \right) \\ &= \frac{1}{4} \left(\frac{\left(\sum_{q=1}^{w_\ell} \frac{2 \sum_{i=1}^{n_{\ell q}} r_i}{1 + (n_{\ell q} - 1)\rho} \right)^2}{\frac{\sigma^2}{\sigma_\mu^2} + \sum_{q=1}^{w_\ell} \frac{n_{\ell q}}{1 + (n_{\ell q} - 1)\rho}} - \frac{\left(\sum_{q=1}^{w_{Lq}} \frac{2 \sum_{i=1}^{n_{Lq}} r_i}{1 + (n_{Lq} - 1)\rho} \right)^2}{\frac{\sigma^2}{\sigma_\mu^2} + \sum_{q=1}^{w_{Lq}} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho}} - \frac{\left(\sum_{q=1}^{w_{Rq}} \frac{2 \sum_{i=1}^{n_{Rq}} r_i}{1 + (n_{Rq} - 1)\rho} \right)^2}{\frac{\sigma^2}{\sigma_\mu^2} + \sum_{q=1}^{w_{Rq}} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho}} \right) \\ &= \sigma_\mu^2 \left(\frac{\left(\sum_{q=1}^{w_\ell} \frac{\sum_{i=1}^{n_{\ell q}} r_i}{1 + (n_{\ell q} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_\ell} \frac{n_{\ell q}}{1 + (n_{\ell q} - 1)\rho}} - \frac{\left(\sum_{q=1}^{w_{Lq}} \frac{\sum_{i=1}^{n_{Lq}} r_i}{1 + (n_{Lq} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{Lq}} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho}} - \frac{\left(\sum_{q=1}^{w_{Rq}} \frac{\sum_{i=1}^{n_{Rq}} r_i}{1 + (n_{Rq} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{Rq}} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho}} \right) \quad (\text{A.23}) \end{aligned}$$

Putting it all together, the log ratio is:

$$\begin{aligned} & \frac{1}{2} \left(\ln(\sigma^2) + (w_{Lq} + w_{Rq} - w_\ell) \ln(1 - \rho) + \sum_{q=1}^{w_\ell} \left(\ln \left(\frac{(n_{\ell q} - 1)\rho + 1}{((n_{Lq} - 1)\rho + 1)((n_{Rq} - 1)\rho + 1)} \right) \right) + \right. \\ & \ln \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_\ell} \frac{n_{\ell q}}{1 + (n_{\ell q} - 1)\rho} \right) - \ln \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{Lq}} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho} \right) - \ln \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{Rq}} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho} \right) \Big) \\ & + \frac{\sigma_\mu^2}{2\sigma^2} \left(\frac{\left(\sum_{q=1}^{w_{Lq}} \frac{\sum_{i=1}^{n_{Lq}} r_i}{1 + (n_{Lq} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{Lq}} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho}} + \frac{\left(\sum_{q=1}^{w_{Rq}} \frac{\sum_{i=1}^{n_{Rq}} r_i}{1 + (n_{Rq} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{Rq}} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho}} - \frac{\left(\sum_{q=1}^{w_\ell} \frac{\sum_{i=1}^{n_{\ell q}} r_i}{1 + (n_{\ell q} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_\ell} \frac{n_{\ell q}}{1 + (n_{\ell q} - 1)\rho}} \right) \\ & - \frac{1}{2\sigma^2(1 - \rho)} \left(\sum_{q=1}^{w_{Lq}} \frac{1}{1 + (n_{Lq} - 1)\rho} \left((1 + n_{Lq}\rho - 2\rho)SSR_{Lq} - \rho CPR_{Lq} \right) + \right. \\ & \sum_{q=1}^{w_{Rq}} \frac{1}{1 + (n_{Rq} - 1)\rho} \left((1 + n_{Rq}\rho - 2\rho)SSR_{Rq} - \rho CPR_{Rq} \right) - \\ & \left. \sum_{q=1}^{w_\ell} \frac{1}{1 + (n_{\ell q} - 1)\rho} \left((1 + n_{\ell q}\rho - 2\rho)SSR_q - \rho CPR_q \right) \right) \quad (\text{A.24}) \end{aligned}$$

A good way to check if this is correct is to set $\rho = 0$ and it should simplify to the

log form of the iidBART GROW ratio.

A.5.3.2 Prune Proposal

This is simply the inverse of the likelihood ratio for the grow proposal. Thus the log-likelihood ratio is just the negative of Equation A.24.

A.5.3.3 Change

The proposal tree differs from the original tree only in the two daughter nodes of the selected change node. These two terminal nodes have the unexplained responses apportioned differently. Denote $R_{1.}$ as the residuals of the first daughter node and $R_{2.}$ as the unexplained responses in the second daughter node. Thus we begin with:

$$\frac{\mathbb{P}\left(\mathbf{R} \mid \mathfrak{F}_*, \sigma^2\right)}{\mathbb{P}\left(\mathbf{R} \mid \mathfrak{F}, \sigma^2\right)} = \frac{\mathbb{P}\left(R_{L^*1}, \dots, R_{L^*n_{L^*}} \mid \sigma^2\right) \mathbb{P}\left(R_{R^*1}, \dots, R_{R^*n_{R^*}} \mid \sigma^2\right)}{\mathbb{P}\left(R_{L1}, \dots, R_{L_{N_L}} \mid \sigma^2\right) \mathbb{P}\left(R_{R1}, \dots, R_{R_{N_R}} \mid \sigma^2\right)} \quad (\text{A.25})$$

Substitution Equation A.16 four times into the above Equation A.25, we find a constant term and an exponential term. We begin with the constant term which is similar to Equation A.18:

$$\sqrt{\frac{\frac{|B_{L^*}| |B_{R^*}|}{|B_L| |B_R|} \left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{L^*,ij}^{-1}\right) \left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} b_{R^*,ij}^{-1}\right)}{\left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{i=1}^{n_{L_L}} \sum_{j=1}^{n_{L_L}} b_{L,ij}^{-1}\right) \left(1 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{i=1}^{n_{L_R}} \sum_{j=1}^{n_{L_R}} b_{R,ij}^{-1}\right)}}$$

We can simplify the above underbraced term similar to Equation A.19 to find:

$$(1 - \rho)^{(w_{L^*} + w_{R^*}) - (w_L + w_R)} \prod_{q=1}^w \frac{((n_{L_q} - 1) \rho + 1) ((n_{R_q} - 1) \rho + 1)}{((n_{L^*q} - 1) \rho + 1) ((n_{R^*q} - 1) \rho + 1)}$$

We can then simplify the other term and it is similar to what we found in Equation A.20:

$$\frac{\left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{L^*}} \frac{n_{L^*q}}{1 + (n_{L^*q} - 1)\rho} \right) \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{R^*}} \frac{n_{R^*q}}{1 + (n_{R^*q} - 1)\rho} \right)}{\left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_L} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho} \right) \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_R} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho} \right)}$$

Now for the exponential component of Equation A.25. Collecting terms inside the exponential and dropping the exponential notation and the $-\frac{1}{2\sigma^2}$ term, we arrive at an expression similar to Equation A.21:

$$\begin{aligned} & \sum_{q=1}^{w_{L^*}} \sum_{i=1}^{n_{L^*}} \sum_{j=1}^{n_{L^*}} d_{L^*q,ij}^{-1} r_i r_j + \sum_{q=1}^{w_{R^*}} \sum_{i=1}^{n_{R^*}} \sum_{j=1}^{n_{R^*}} d_{R^*q,ij}^{-1} r_i r_j - \\ & \sum_{q=1}^{w_L} \sum_{i=1}^{n_L} \sum_{j=1}^{n_L} d_{Lq,ij}^{-1} r_i r_j - \sum_{q=1}^{w_R} \sum_{i=1}^{n_R} \sum_{j=1}^{n_R} d_{Rq,ij}^{-1} r_i r_j \end{aligned}$$

plus $\frac{1}{4\sigma^2}$ times the following

$$\begin{aligned} & \frac{\left(\sum_{q=1}^{w_L} \sum_{i=1}^{n_L} \sum_{j=1}^{n_L} d_{Lq,ij}^{-1} (r_i + r_j) \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_L} \sum_{i=1}^{n_L} \sum_{j=1}^{n_L} d_{Lq,ij}^{-1}} + \frac{\left(\sum_{q=1}^{w_R} \sum_{i=1}^{n_R} \sum_{j=1}^{n_R} d_{Rq,ij}^{-1} (r_i + r_j) \right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_R} \sum_{i=1}^{n_R} \sum_{j=1}^{n_R} d_{Rq,ij}^{-1}} - \end{aligned}$$

$$\frac{\left(\sum_{q=1}^{w_{L^*}} \sum_{i=1}^{n_{L^*}} \sum_{j=1}^{n_{L^*}} d_{L^*q,ij}^{-1}(r_i + r_j)\right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_{L^*}} \sum_{i=1}^{n_{L^*}} \sum_{j=1}^{n_{L^*}} d_{L^*q,ij}^{-1}} - \frac{\left(\sum_{q=1}^{w_{R^*}} \sum_{i=1}^{n_{R^*}} \sum_{j=1}^{n_{R^*}} d_{R^*q,ij}^{-1}(r_i + r_j)\right)^2}{\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma^2} \sum_{q=1}^{w_{R^*}} \sum_{i=1}^{n_{R^*}} \sum_{j=1}^{n_{R^*}} d_{R^*q,ij}^{-1}}$$

The first line of the tiny formula above can be reduced similarly to Equation A.22:

$$\begin{aligned} & \frac{1}{1-\rho} \left(\sum_{q=1}^{w_{L^*}} \frac{1}{1+(n_{L^*q}-1)\rho} ((1+n_{L^*q}\rho-2\rho)SSR_{L^*q} - \rho CPR_{L^*q}) \right. \\ & \quad + \sum_{q=1}^{w_{R^*}} \frac{1}{1+(n_{R^*q}-1)\rho} ((1+n_{R^*q}\rho-2\rho)SSR_{R^*q} - \rho CPR_{R^*q}) \\ & \quad - \sum_{q=1}^{w_L} \frac{1}{1+(n_{Lq}-1)\rho} ((1+n_{Lq}\rho-2\rho)SSR_{Lq} - \rho CPR_{Lq}) \\ & \quad \left. - \sum_{q=1}^{w_R} \frac{1}{1+(n_{Rq}-1)\rho} ((1+n_{Rq}\rho-2\rho)SSR_{Rq} - \rho CPR_{Rq}) \right) \end{aligned}$$

And the second line of the tiny formula can be reduced similarly to Equation A.23

as σ_μ^2 times:

$$\begin{aligned} & \frac{\left(\sum_{q=1}^{w_L} \frac{\sum_{i=1}^{n_{Lq}} r_i}{1+(n_{Lq}-1)\rho}\right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_L} \frac{n_{Lq}}{1+(n_{Lq}-1)\rho}} + \frac{\left(\sum_{q=1}^{w_R} \frac{\sum_{i=1}^{n_{Rq}} r_i}{1+(n_{Rq}-1)\rho}\right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_R} \frac{n_{Rq}}{1+(n_{Rq}-1)\rho}} - \\ & \frac{\left(\sum_{q=1}^{w_{L^*}} \frac{\sum_{i=1}^{n_{L^*q}} r_i}{1+(n_{L^*q}-1)\rho}\right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{L^*}} \frac{n_{L^*q}}{1+(n_{L^*q}-1)\rho}} - \frac{\left(\sum_{q=1}^{w_{R^*}} \frac{\sum_{i=1}^{n_{R^*q}} r_i}{1+(n_{R^*q}-1)\rho}\right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{R^*}} \frac{n_{R^*q}}{1+(n_{R^*q}-1)\rho}} \end{aligned}$$

Thus, the log ratio is below and similar to Equation A.24:

$$\begin{aligned}
& \frac{1}{2} \left(((w_{L^*} + w_{R^*}) - (w_L + w_R)) \ln(1 - \rho) \right. \\
& + \sum_{q=1}^{w_L} \left(\ln \left(\frac{((n_{L_q} - 1)\rho + 1)((n_{R_q} - 1)\rho + 1)}{((n_{L^*q} - 1)\rho + 1)((n_{R^*q} - 1)\rho + 1)} \right) \right) \\
& + \ln \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{L^*}} \frac{n_{L^*q}}{1 + (n_{L^*q} - 1)\rho} \right) + \ln \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{R^*}} \frac{n_{R^*q}}{1 + (n_{R^*q} - 1)\rho} \right) \\
& - \ln \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_L} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho} \right) \\
& \left. - \ln \left(\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_R} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho} \right) \right) \tag{A.26}
\end{aligned}$$

plus $\frac{\sigma_\mu^2}{2\sigma^2}$ times:

$$\begin{aligned}
& \frac{\left(\sum_{q=1}^{w_{L^*}} \frac{\sum_{i=1}^{n_{L^*q}} r_i}{1 + (n_{L^*q} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{L^*}} \frac{n_{L^*q}}{1 + (n_{L^*q} - 1)\rho}} + \frac{\left(\sum_{q=1}^{w_{R^*}} \frac{\sum_{i=1}^{n_{R^*q}} r_i}{1 + (n_{R^*q} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_{R^*}} \frac{n_{R^*q}}{1 + (n_{R^*q} - 1)\rho}} - \\
& \frac{\left(\sum_{q=1}^{w_L} \frac{\sum_{i=1}^{n_{Lq}} r_i}{1 + (n_{Lq} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_L} \frac{n_{Lq}}{1 + (n_{Lq} - 1)\rho}} - \frac{\left(\sum_{q=1}^{w_R} \frac{\sum_{i=1}^{n_{Rq}} r_i}{1 + (n_{Rq} - 1)\rho} \right)^2}{\sigma^2 + \sigma_\mu^2 \sum_{q=1}^{w_R} \frac{n_{Rq}}{1 + (n_{Rq} - 1)\rho}}
\end{aligned}$$

minus $\frac{1}{2\sigma^2(1-\rho)}$ times:

$$\begin{aligned}
& \sum_{q=1}^{w_{L^*}} \frac{1}{1 + (n_{L^*q} - 1)\rho} \left((1 + n_{L^*q}\rho - 2\rho)SSR_{L^*q} - \rho CPR_{L^*q} \right) + \\
& \sum_{q=1}^{w_{R^*}} \frac{1}{1 + (n_{R^*q} - 1)\rho} \left((1 + n_{R^*q}\rho - 2\rho)SSR_{R^*q} - \rho CPR_{R^*q} \right) -
\end{aligned}$$

$$\sum_{q=1}^{w_L} \frac{1}{1 + (n_{Lq} - 1)\rho} \left((1 + n_{Lq}\rho - 2\rho)SSR_{Lq} - \rho CPR_{Lq} \right) - \sum_{q=1}^{w_R} \frac{1}{1 + (n_{Rq} - 1)\rho} \left((1 + n_{Rq}\rho - 2\rho)SSR_{Rq} - \rho CPR_{Rq} \right) \quad (A.27)$$

Bibliography

- Abramovitz, M., Scitovsky, T., and Inkeles, A. (1973). Economic growth and its discontents. *Bulletin of the American Academy of Arts and Sciences*, pages 11–27.
- Akkaya, C., Conrad, A., Wiebe, J., and Mihalcea, R. (2010). Amazon mechanical turk for subjectivity word sense disambiguation. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazons Mechanical Turk*, pages 195–203.
- Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.
- Alm, C., Roth, D., and Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics.
- Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, pages 196–205. Springer.
- Angrist, J. D. (2001). Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors. *Journal of Business & Economic Statistics*, 19(1):2–28.
- Appleton, J. J., Christenson, S. L., Kim, D., and Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of School Psychology*, 44(5):427–445.
- Appleton, J. J., Christenson, S. L., Kim, D., and Reschly, A. L. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, 45(5):369–386.
- Ariely, D., Kamenica, E., and Prelec, D. (2008). Man’s Search for Meaning: The Case of Legos. *Journal of Economic Behavior & Organization*, 67(3-4):671 – 677.
- Atkinson, A. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, 69(1):61–67.
- Bache, K. and Lichman, M. (2013). UCI machine learning repository.
- Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D. (2012). Personality and patterns of facebook usage. *Web Science*.

- Barankay, I. (2010). Rankings and social tournaments: Evidence from a field experiment. *University of Pennsylvania mimeo*.
- Begg, C. and Iglewicz, B. (1980). A treatment allocation procedure for sequential clinical trials. *Biometrics*, 36(1):81–90.
- Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com’s mechanical turk. *Political Analysis*, 20:351–368.
- Blattenberger, G. and Fowles, R. (2014). Avalanche forecasting: Using bayesian additive regression trees (BART). In *Demand for Communications Services—Insights and Perspectives*, pages 211–227. Springer.
- Bleich, J. and Kapelner, A. (2014). Bayesian additive regression trees with parametric models of heteroskedasticity. *arXiv*.
- Bleich, J., Kapelner, A., George, E., and Jensen, S. (2014). Variable selection for BART: An application to gene regulation. *arXiv preprint*.
- Bleich, J., Kapelner, A., Jensen, S., and George, E. (2013). Using bart for variable selection. *ArXiv e-prints*.
- Breiman, L. (2001a). Random forests. *Machine learning*, 45(1):5–32.
- Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–231.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. wadsworth & brooks. *Monterey, CA*.
- Brown, S. W., Rood, T., and Palmer, M. (2010). Number or nuance: Which factors restrict reliable word sense annotation? In *LREC: The International Conference on Language Resources and Evaluation*, pages 3237–3243.
- Bush, A. and Parasuraman, A. (1984). Assessing response quality. A self-disclosure approach to assessing response quality in mall intercept and telephone interviews. *Psychology and Marketing*, 1(3-4):57–71.
- Cacioppo, J. and Petty, R. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1):116–131.
- Cacioppo, J., Petty, R., and Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48(3):306–307.
- Callison-Burch, C. (2010). Creating speech and language data with Amazon’s Mechanical Turk. *NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 1–12.
- Cartwright, N. (2007). Are rcts the gold standard? *BioSocieties*, 2:11–20.
- Chall, J. and Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.
- Chandler, D. and Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowd-sourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133.
- Chilton, L., Horton, J., Miller, R., and Azenkot, S. (2010). Task search in a human computation market. *Proceedings of the ACM SIGKDD workshop on human computation*, pages 1–9.

- Chipman, H., George, E., and McCulloch, R. (2010). BART: Bayesian Additive Regressive Trees. *The Annals of Applied Statistics*, 4(1):266–298.
- Chklovski, T. and Mihalcea, R. (2003). Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Proceedings of the Conference on Recent Advances on Natural Language Processing*.
- Chow, S. and Chang, M. (2008). Adaptive design methods in clinical trials - a review. *Orphanet journal of rare diseases*, 3:11.
- Cook, T. and Campbell, D. (1979). *Quasi-Experimentation Design & Analysis Issues for Field Settings*. Houghton Mifflin Company.
- Couper, M. P., Tourangeau, R., and Marvin, T. (2009). Taking the Audio Out of Audio-CASI. *Public Opinion Quarterly*, 73(2):281–303.
- Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474.
- Csikszentmihalyi, M. (1997). *Creativity: Flow and the psychology of discovery and invention*. Harper Perennial.
- Damien, P., Dellaportas, P., Polson, N., and Stephens, D. (2013). *Bayesian Theory and Applications*, pages 455–464. Oxford University Press, first edition.
- Davidov, D., Tsur, O., and Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *Coling 2010: Posters*, pages 241–249, Beijing, China. Coling 2010 Organizing Committee.
- Davies, M. (2008). *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available online at <http://corpus.byu.edu/coca/>.
- DeMaio, T. (1984). Social desirability and survey measurement: A review. *Surveying Subjective Phenomena*, 2:257–282.
- Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75.
- Diener, E., Inglehart, R., and Tay, L. (2012). Theory and validity of life satisfaction scales. *Social Indicators Research*, pages 1–31.
- Diener, E., Ng, W., Harter, J., and Arora, R. (2010). Wealth and happiness across the world: material prosperity predicts life evaluation, whereas psychosocial prosperity predicts positive feeling. *Journal of Personality and Social Psychology*, 99(1):52.
- Diener, E. and Seligman, M. E. (2002). Very happy people. *Psychological Science*, 13(1):81–84.
- Ding, Y. and Simonoff, J. (2010). An investigation of missing data methods for classification trees applied to binary response data. *The Journal of Machine Learning Research*, 11:131–170.
- Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *Diversity*, page 26.
- Druckman, J. N. and Kam, C. D. (2011). *Students as experimental participants: A defense of the 'narrow data base' in Handbook of experimental political science*, pages 41–57. Cambridge University Press.

- Easton, M. (2006). Britain's happiness in decline. *BBC News*, 2.
- Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417.
- Efron, B. and Tibshirani, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447.
- Eliashberg, J. (2010). *Green-lighting Movie Scripts: Revenue Forecasting and Risk Management*. PhD thesis, University of Pennsylvania.
- Ericsson, K. A. (2002). Attaining excellence through deliberate practice: Insights from the study of expert performance. *The pursuit of excellence through education*, pages 21–55.
- Eriksson, K. and Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, 5(3):159–163.
- Fan, R., Chang, K., Hsieh, C., Wang, X., and Lin, C. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.
- Fellbaum, C., Grabowski, J., Landes, S., and L, S. (1997). Analysis of a hand-tagging task. In *Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics*, pages 34–40.
- Forgeard, M. J., Jayawickreme, E., Kern, M. L., and Seligman, M. E. (2011). Doing the right thing: Measuring wellbeing for public policy. *International Journal of Wellbeing*, 1(1).
- Foster, H., Hanno, P., Nickel, J., Payne, C., Mayer, R., Burks, D., Yang, C., Chai, T., Kreder, K., Peters, K., Lukacz, E., FitzGerald, M., Cen, L., Landis, J., Propert, K., Yang, W., Kusek, J., and Nyberg, L. (2010). Effect of amitriptyline on symptoms in treatment naïve patients with interstitial cystitis/painful bladder syndrome. *The Journal of urology*, 183(5):1853–8.
- Frankl, V. E. (1997). *Man's search for ultimate meaning*. Insight Books/Plenum Press.
- Fredricks, J., McColskey, W., Meli, J., Montrose, B., Mordica, J., and Mooney, K. (2011). Measuring student engagement in upper elementary through high school: A description of 21 instruments. *Issues & Answers Report, REL*, 98(098).
- Freedman, D. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193.
- Friedman, J. (1991). Multivariate adaptive regression splines (with discussion and a rejoinder by the author). *The annals of statistics*, 19:1–67.
- Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232.
- Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.
- Gacs, P. and Lovász, L. (1981). *Khachiyan's algorithm for linear programming*. Springer.
- Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall / CRC, second edition.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6:721–741.

- George, E. I. and Robert, C. P. (1992). Capture-Recapture Estimation Via Gibbs Sampling. *Biometrika*, 79(4):677.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*, volume 2. CRC press.
- Gill, A., Nowson, S., and Oberlander, J. (2009). What are they blogging about? personality, topic and motivation in blogs. *Proc. of AAAI ICWSM*.
- Gneezy, U. and List, J. A. (2006). Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments. *Econometrica*, 74(5):1365–1384.
- Golder, S. and Macy, M. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881.
- Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2014). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics (in press)*.
- Greevy, R., Lu, B., Silber, J., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics (Oxford, England)*, 5(2):263–75.
- Halko, N., Martinsson, P., and Tropp, J. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.
- Han, B., Enas, N., and McEntegart, D. (2009). Randomization by minimization for unbalanced treatment allocation. *Statistics in medicine*, 28:3329–3346.
- Hansen, B. and Klopfer, S. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627.
- Harrison, G. W. and List, J. A. (2004). Field Experiments. *Journal of Economic Literature*, 42(4):1009 – 1055.
- Hastie, T. and Tibshirani, R. (2000). Bayesian Backfitting. *Statistical Science*, 15(3):196–213.
- Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.
- Henrich, J., Heine, S., and Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, 33(2-3):61–83.
- Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.
- Hoffman, J. and Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception and Psychophysics*.
- Holmes, S. and Kapelner, A. (2010). Quality assessment of feature counting using a distributed workforce: Crowd counting a crowd. working paper.
- Holmes, S., Kapelner, A., and Lee, P. (2009). An interactive java statistical image segmentation system: Gemident. *Journal of Statistical Software*.
- Horton, J. and Chilton, L. (2010). The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 209–218. ACM.

- Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425.
- Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90 percent solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 57–60. Association for Computational Linguistics.
- Hu, F. and Rosenberger, W. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. John Wiley & Sons, Inc.
- Ide, N. and Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):2–40.
- Ipeirotis, P. (2010). Demographics of Mechanical Turk. CeDER working paper CeDER-10-01, New York University, Stern School of Business.
- Jennison, C. and Turnbull, B. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall / CRC.
- Kalish, L. and Begg, C. (1987). The impact of treatment allocation procedures on nominal significance levels and bias. *Controlled clinical trials*, 8(2):121–35.
- Kapelner, A. and Bleich, J. (2013a). Bartmachine: A powerful tool for machine learning. *ArXiv e-prints*.
- Kapelner, A. and Bleich, J. (2013b). Prediction with missing data via bayesian additive regression trees. *arXiv*.
- Kapelner, A. and Chandler, D. (2010). Preventing Satisficing in Online Surveys : A “Kapcha” to Ensure Higher Quality Data. In *CrowdConf ACM Proceedings*.
- Kapelner, A., Holmes, S., and Lee, P. (2007a). www.GemIdent.com.
- Kapelner, A., Kaliannan, K., Schwartz, H., Ungar, L. H., and Foster, D. P. (2012). New insights from coarse word sense disambiguation in the crowd. In *CoLING*.
- Kapelner, A. and Krieger, A. (2014). Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*.
- Kapelner, A., Lee, P., and Holmes, S. (2007b). An interactive statistical image segmentation and visualization system. *Medical Information Visualisation- Medevis -IEEE*.
- Kapelner, A. and Ungar, L. (2013). Crowdsourcing for statisticians designing applications and analyzing data on mturk. https://dl.dropboxusercontent.com/u/93146/jsm_tutorial.pdf. [JSM 2013 Tutorial Slides].
- Kibler, D., Aha, D., and Albert, M. (1989). Instance-based prediction of real-valued attributes. *Computational Intelligence*, 5:51.
- Kindo, B., Wang, H., and Pe, E. (2013). MBACT - Multiclass Bayesian Additive Classification Trees. *arXiv*, pages 1–29.
- Kosinski, M. and Stillwell, D. (2012). mypersonality project. In <http://www.mypersonality.org/wiki/>.
- Kramer, A. (2010). An unobtrusive behavioral model of gross national happiness. In *Proc of the 28th int conf on Human factors in comp sys*, pages 287–290. ACM.

- Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(1):61–70.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50:537–67.
- Krosnick, J. A. (2000). The threat of satisficing in surveys: the shortcuts respondents take in answering questions. *Survey Methods Centre Newsletter*.
- Krosnick, J. A., Narayan, S., and Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 1996(70):29–44.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.
- Lawless, N. M. and Lucas, R. E. (2011). Predictors of regional well-being: a county level analysis. *Social Indicators Research*, 101(3):341–357.
- Layard, R. (2005). Rethinking public economics: The implications of rivalry and habit. *Economics and Happiness*, pages 147–170.
- Levitt, S. and List, J. (2009). Field experiments in economics: the past, the present, and the future. *European Economic Review*, 53(1):1–18.
- Levitt, S. D. and List, J. A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives*, 21(2):153–174.
- Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R news*, 2:18–22.
- Little, R. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134.
- Little, R. and Rubin, D. (2002). *Statistical analysis with missing data*. Wiley, second edition.
- Mairesse, F., Walker, M., Mehl, M., and Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500.
- Mason, W. and Suri, S. (2011). Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior research methods*, *Forthcoming*.
- McCrae, R. and Costa, P. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81–90.
- McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.
- McDonald, R., Hannan, K., Neylon, T., Wells, M., and Reynar, J. (2007). Structured models for fine-to-coarse sentiment analysis. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 432.
- McEntegart, D. (2003). The pursuit of balance using stratified and dynamic randomization techniques: an overview. *Drug Information Journal*, 37:293–308.

- Meyer, G., Finn, S., Eyde, L., Kay, G., Moreland, K., Dies, R., Eisman, E., Kubiszyn, T., and Read, G. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, 56(2):128.
- Mihalcea, R. and Liu, H. (2006). A corpus-based approach to finding happiness. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs*, page 19.
- Morgan, K. and Rubin, D. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282.
- Morris, M. D. (2011). *Design of Experiments: An Introduction Based on Linear Models*. Chapman & Hall / CRC Press.
- Navigli, R. (2009). Word sense disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69.
- Nowson, S. (2007). Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *In Proceedings of the International Conference on Weblogs and Social*.
- Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872.
- Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). *Statistical inference from capture data on closed animal populations*, volume 62. wildlife society.
- Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.
- Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.
- Paolacci, G., Chandler, J., and Ipeirotis, P. (2010a). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419.
- Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010b). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.
- Parent, G. (2010). Clustering dictionary definitions using Amazon Mechanical Turk. *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 21–29.
- Passonneau, R. J., Bhardwaj, V., Salieb-Aouissi, A., and Ide, N. (2011). Multiplicity and word sense: Evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*.
- Pavot, W. and Diener, E. (1993). Review of the satisfaction with life scale. *Psychological Assessment*, 5(2):164.
- Pearlin, L. I. and Schooler, C. (1978). The structure of coping. *Journal of health and social behavior*, pages 2–21.
- Pennebaker, J. W., Chung, C., Ireland, M., Gonzales, A., and Booth, R. (2007). The development and psychometric properties of liwc2007. *Austin, TX, LIWC. Net*.

- Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Word J. Of The Int Ling Assoc.*
- Peterson, C. and Park, N. (2003). Positive psychology as the evenhanded positive psychologist views it. *Psychological Inquiry*, 14(2):143–147.
- Pinter-Wollman, N., Wollman, R., Guetz, A., Holmes, S., and Gordon, D. M. (2011). The effect of individual variation on the structure and function of interaction networks in harvester ants. *Journal of The Royal Society Interface*, 8(64):1562–1573.
- Pocock, S. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31(1):103–115.
- Pradhan, S., Loper, E., and Dligach, D. (2007). Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92.
- Pratola, M., Chipman, H., Higdon, D., McCulloch, R., and Rust, W. (2013). Parallel bayesian additive regression trees. Technical report, University of Chicago.
- Preston, A. E. (1989). The Nonprofit Worker in a For-Profit World. *Journal of Labor Economics*, 7(4):438–463.
- R Development Core Team (2005). R: A language and environment for statistical computing. ISBN 3-900051-07-0.
- Raghavarao, D. (1980). Use of distance function in sequential treatment assignment for prognostic factors in the controlled clinical trial. *Calcutta Statist. Assoc. Bulletin*, 29:99–102.
- Raudenbush, S., Martinez, A., and Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis*, 29(1):5–29.
- Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48, Ann Arbor, Michigan. Association for Computational Linguistics.
- Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer.
- Rosen, S. (1986). The theory of equalizing differences. *Handbook of labor economics*, 1:641–692.
- Rosenbaum, P. (2010). *Design of observational studies*. Springer, New York.
- Rosenberger, W. and Sverdlov, O. (2008). Handling Covariates in the Design of Clinical Trials. *Statistical Science*, 23(3):404–419.
- Rosso, B. D., Dekas, K. H., and Wrzesniewski, A. (2010). On the meaning of work: A theoretical integration and review. *Research in Organizational Behavior*, 30:91–127.
- Rothwell, D. (2005). External validity of randomised controlled trials: “to whom do the results of this trial apply?”. *Lancet*, 365:82–93.
- Rowe, J. W. and Kahn, R. L. (1987). Human aging: usual and successful. *Science*.
- Rubin, D. (1978). Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. Technical report, Proceedings of the Survey Research Methods Section, American Statistical Association.

- Rubin, D. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328.
- Russell, B., Torralba, A., and Murphy, K. (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, pages 157–173.
- Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68.
- Ryff, C. D. and Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of personality and social psychology*, 69(4):719.
- Saleem, S., Prasad, R., Vitaladevuni, S., Pacula, M., Crystal, M., Marx, B., Sloan, D., Vasterling, J., and Speroff, T. (2012). Automatic detection of psychological distress indicators and severity assessment from online forum posts. In *Proceedings of COLING 2012*, pages 2375–2388, Mumbai, India. The COLING 2012 Organizing Committee.
- Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall / CRC.
- Schaufeli, W. B., Bakker, A. B., and Salanova, M. (2006). The measurement of work engagement with a short questionnaire a cross-national study. *Educational and psychological Measurement*, 66(4):701–716.
- Schueler, S. M. and Seligman, M. E. (2010). Pursuit of pleasure, engagement, and meaning: Relationships to subjective and objective measures of well-being. *The Journal of Positive Psychology*, 5(4):253–263.
- Scott, N., McPherson, G., Ramsay, C., and Campbell, M. (2002). The method of minimization for allocation to clinical trials. a review. *Controlled clinical trials*, 23(6):662–74.
- Seligman, M. E. P. (2011). *Flourish: A Visionary New Understand of Happiness and Well-being*. Free Press.
- Senn, S. (2000). Consensus and controversy in pharmaceutical statistics. *Journal of the Royal Statistical Society: Series D*, 49(2):135–176.
- Setiadi, A. F., Ray, N. C., Kohrt, H. E., Kapelner, A., Carcamo-Cavazos, V., Levic, E. B., Yadegarynia, S., van der Loos, C. M., Schwartz, E. J., and Holmes, S. (2010). Quantitative, architectural analysis of immune cell subsets in tumor-draining lymph nodes from breast cancer patients and healthy lymph nodes. *PloS one*, 5(8):e12420.
- Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.
- Simon, H. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*.
- Simon, R. (1979). Restricted randomization designs in clinical trials. *Biometrics*, 35(2):503–512.
- Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., and Li Zhu, W. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237.
- Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

- Sorokin, A. and Forsyth, D. (2008). Utility Data Annotation with Amazon Mechanical Turk. *First IEEE Workshop on Internet Vision, CVPR 08*.
- Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 43(1):155–167.
- Steger, M. F., Kashdan, T. B., Sullivan, B. A., and Lorentz, D. (2008). Understanding the search for meaning in life: Personality, cognitive style, and the dynamic between seeking and experiencing meaning. *Journal of Personality*, 76(2):199–228.
- Stekhoven, D. and Bühlmann, P. (2012). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28:112–8.
- Stern, S. (2004). Do scientists pay to be scientists? *Management Science*, 50(6):835–853.
- Stiglitz, J. E., Sen, A., and Fitoussi, J. (2009). Report by the commission on the measurement of economic performance and social progress. *The Commission on the Measurement of Economic Performance and Social Progress website*. Available online at: <http://www.stiglitz-sen-fitoussi>.
- Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM.
- Student (1931). The Lanarkshire milk experiment. *Biometrika*, 23(3):398–406.
- Sumner, C., Byers, A., Boochever, R., and Park, G. (2012). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. www.onlineprivacyfoundation.org.
- Taddy, M., Gramacy, R., and Polson, N. (2011). Dynamic Trees for Learning and Design. *Journal of the American Statistical Association*, 106(493):109–123.
- Tang, Y. and Chen, H. (2011). Emotion modeling from writer/reader perspectives using a microblog dataset. *Sentiment Analysis where AI meets Psychology (SAAIP)*, page 11.
- Taves, D. (1974). Minimization: a new method of assigning patients to treatment and control groups. *Clinical pharmacology and therapeutics*, 15(5):443.
- Tay, L., Tan, K., Diener, E., and Gonzalez, E. (2012). Social relations, health behaviors, and health outcomes: A survey and synthesis. *Applied Psychology: Health and Well-Being*.
- Taylor, S. (1965). Eye movements in reading: Facts and fallacies. *American Educational Research Journal*, 2(4):187.
- Taylor, S. E. (2011). Social support: A review. *The Oxford Handbook of Health Psychology*, pages 189–214.
- Thaler, R. (1985). Mental Accounting and Consumer Choice. *Marketing Science*, 4(3):199 – 214.
- Therneau, T. and Atkinson, E. (1997). An introduction to recursive partitioning using the rpart routines. Technical report, Mayo Foundation.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
- Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

- Troyanskaya, O., Cantor, M., and Sherlock, G. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.
- Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 23(5):1–35.
- Twala, B., Jones, M., and Hand, D. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956.
- Urbanek, S. (2011). *rJava: Low-level R to Java interface*. R package version 0.9-3 available on CRAN.
- Von Ahn, L., Blum, M., Hopper, N., and Langford, J. (2003). CAPTCHA: Using hard AI problems for security. *Advances in Cryptology EUROCRYPT 2003*.
- Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. John Wiley & Sons, Inc., 2nd edition.
- Wrzesniewski, A., McCauley, C., Rozin, P., and Schwartz, B. (1997). Jobs, careers, and callings: People’s relations to their work. *Journal of research in personality*, 31(1):21–33.
- Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373.
- Yip, P., Xi, L., Arnold, R., and Hayakawa, Y. (2005). A beta-binomial model for estimating the size of a heterogeneous population. *Australian & New Zealand Journal of Statistics*, 47(3):299–308.
- Zhou, Q. and Liu, J. S. (2008). Extracting sequence features to predict protein–DNA interactions: a comparative study. *Nucleic acids research*, 36(12):4137–4148.