Inverse Problems in Matching Markets and Signal Processing

by

Yifei Sun

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics

New York University

May 2018

Professor Alfred Galichon

Dedication

To my parents Jianshe Sun and Luhua Chen, with gratitude.

Acknowledgements

I would like to acknowledge the immense support and help of my advisor, Professor Alfred Galichon. Without his encouragement and enthusiasm this work would not have been possible. He has inspired me both intellectually and personally.

I would also like to thank Professor Arnaud Dupuy, who collaborated with me on the first and second chapters of this dissertation.

I am also grateful to Professor Eero Simoncelli, who originally introduced me to the fascinating world of numerical optimization. The third chapter is based on the joint work with him. My time spent in his LCV lab is memorable.

I would also like to thank Professor Esteban Tabak, Professor Sinan Güntürk, and Professor Robert Kohn for being my dissertation committee members.

Lastly, I want to thank all my friends at the Courant Institute and the LCV lab, including Kangping, Jiajun, Yuanxun, Zhiyuan, Zhilei, Monty, James, Zhe, Kede, Zhuo, Xingxin, Valero, Johannes, Chenyue, Nan, Fang, Chengcheng, as well as my roommates Onur, Wendy, Sylwia, and Flávio. Their support has made my Ph.D. journey enjoyable.

Abstract

Inverse problems arise in many fields such as economics, neuroscience, and engineering. In this dissertation we investigate several inverse problems.

In the first chapter of the dissertation, we address the problem of estimating transport surplus (a.k.a. matching affinity) in high dimensional optimal transport problems. Classical optimal transport theory specifies the matching affinity and determines the optimal joint distribution. In contrast, we study the inverse problem of estimating matching affinity based on the observation of the joint distribution, using an entropic regularization of the problem.

In the second chapter of the dissertation, we review gravity equations for bilateral trades and the corresponding Pseudo-Maximum-Likelihood (PML) estimation technique. We then establish the connections between PML and optimal transport.

In the third chapter of the dissertation, we consider the class of binary inverse problems, in which an observed signal is formed as a superposition of a subset of template signals drawn from a dictionary, and corrupted by additive noise. We assume a Bernoulli prior for the binary coefficients specifying the subset, with known mean taking any value between zero and one. We formulate a maximum a posteriori solution and provide an iterative algorithm for approximating it.

Contents

List of Figures

List of Tables

Chapter 1

Estimating matching affinity matrix under low-rank constraints

In this chapter, we address the problem of estimating transport surplus (a.k.a. matching affinity) in high dimensional optimal transport problems. Classical optimal transport theory specifies the matching affinity and determines the optimal joint distribution. In contrast, we study the inverse problem of estimating matching affinity based on the observation of the joint distribution, using an entropic regularization of the problem. To accommodate high dimensionality of the data, we propose a novel method that incorporates a nuclear norm regularization which effectively enforces a rank constraint on the affinity matrix. The low-rank matrix estimated in this way reveals the main factors which are relevant for matching.

1.1 Introduction

Optimal transport theory has attracted a lot of interest across a number of scientific disciplines, from pure mathematics [33] to various applications including machine learning [4] mathematical statistics [10] and economics [16]. The basic problem of optimal transport is how to form pairs of agents drawn from two populations in order to maximize the total utility, also called matching affinity. The resulting joint distribution of pairs is called an optimal matching, also called optimal transport plan.

Most of the theory of optimal transport has focused on the direct problem, namely solving for the optimal matching, taking the matching affinities as given. In contrast, we consider in this chapter the *inverse optimal transport problem*: given the observation of an optimal matching, what is the affinity function for which this matching is optimal[1]? This problem arises naturally in the study of two-sided matching markets, which appears in various fields of the social sciences. In sociology and economics, one instance of these markets is the "marriage market," following Becker [3]'s seminal analysis, where one observes the characteristics of both partners in married couples (such as education, height, personality traits, etc.), and one wants to infer (i) which characteristics attract or repel each other the most, and (ii) what combinations of characteristics are the most relevant for matching.

In models of matching markets, vectors of characteristics $x \in \mathbb{R}^d$ for one side of the market and $y \in \mathbb{R}^{d'}$ for the the other side are available, and the joint distribution $\hat{\pi}(x, y)$ across matched pairs is observed, and we are interested in

[1] In the theoretical computer science literature, this problem is known as an *inverse assignment problem*, see [5], Section 6.7 and references therein.

estimating the matching affinity function $\Phi(x, y)$. Broadly speaking, models of matching markets are divided into three categories: scalar index models, discrete models, and multivariate models, which we will now briefly survey.

Scalar index models. A number of papers use scalar index models: they assume that agents match on a pair of scalar indices $\tilde{x} = u^\top x$ and $\tilde{y} = v^\top y$, which are weighted sums of partners' characteristics. Following a suggestion by [3], a number of papers have used canonical correlation or linear regression techniques in order to estimate the weight vectors u and v; see for instance [19, 20], and a caution against the misuse of these techniques in [17]. A more robust ways to estimate the weight vectors has been suggested by [28] using rank correlation. See also [8].

Discrete models. Following a seminal paper by [9], a number of recent papers [15, 8, 18] have assumed that agents match based on discrete characteristics, either categorical variables like ethnicity, or binned, such as the income bracket. However, the binning of cardinal variables may be problematic as the results may depend heavily on the arbitrary choice of the thresholds. Therefore, these models suffer from limitations when dealing with non-categorical variables.

Continuous models. More recently, a continuous model has been proposed by [12], where the matching affinity is bilinear with respect to the matched pairs' characteristics, i.e. is given by $x^\top Ay$, where A, called the *affinity matrix* is a $d \times d'$ matrix is to be estimated. This model enables weighted interactions between any pair of characteristics. Of course, when the rank of A is one, $A = \lambda uv^\top$, and one recovers the scalar index models discussed above. But as soon as the rank of A is greater than one, a pair of scalar indices on each side of the market would not be sufficient to describe the matching affinity. The authors of [12] propose a moment matching procedure to estimate A, which can be computed via convex

optimization. However, as soon as the number of characteristics goes large, the number of parameters to be estimated grows quadratically, potentially leading to an overfit.

In this chapter, we propose a novel method for solving the inverse optimal transport problem in a high-dimensional setting, where we estimate the affinity matrix A under a rank constraint in order to capture the relevant dimensions of interaction on which matching occurs. An application to the marriage market is proposed, which uses the same data as in [12] and illustrates how our method allows one to identify the impact of narrowly defined personality traits without having to aggregate these into aggregate traits prior to the estimation of the affinity matrix as in that paper.

The rest of the chapter is organized as follows. Section 1.2 presents the matching equilibrium model and introduces the concept of affinity matrix. Section 1.3 though 1.6 describe the maximum likelihood estimation of the affinity matrix, including a low-rank regularized version. Section 1.7 presents the application to the Dutch marriage markets dataset. Section 1.8 concludes the chapter.

1.2 The model

We first briefly recall the optimal transport problem; see [33, 34] for more. Given two probability distributions μ_1 and μ_2 over $\mathbb{R}^d \times \mathbb{R}^{d'}$, the optimal transport problem is defined as

$$\max_{\pi \in \Pi(\mu_1, \mu_2)} \mathbb{E}_\pi \left[\Phi \left(X, Y \right) \right] \tag{1.1}$$

where $\Phi(x, y)$ is the measure of affinity between two agents $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^{d'}$ on each side of the market, and $\Pi(\mu_1, \mu_2)$ is the set of distributions $\pi(x, y)$ with

marginal distributions μ_1 and μ_2. Problem (1.1) is the *Monge-Kantorovich problem* of optimal transport.

1.2.1 Optimal solution vs equilibrium

The optimization problem (1.1) yields a centralized solution where a central planner would decide which pairs to form. However, most matching markets (including the marriage market which we study in this chapter) are *decentralized markets*, in which agents decide based on their own interest, leading to an equilibrium. It follows from the work of [3] and [26] that the centralized and the decentralized problems are equivalent. We sketch the argument as follows.

In decentralized problems, an *outcome* is the specification of a matching $\pi \in \Pi(\mu_1, \mu_2)$, and of individual payoffs $u(x)$ and $v(y)$, which are attained by agents of respective types x and y. The outcome is called *stable* when

$$u(x) + v(y) \geq \Phi(x, y) \ \ \forall x, y. \tag{1.2}$$

Stability is a required condition for equilibrium. Indeed, if (1.2) were not to hold, then $\varepsilon = \Phi(x, y) - u(x) - v(y)$ would be strictly positive, and thus by matching together, x and y could attain $u(x) + \varepsilon/2$ and $v(y) + \varepsilon/2$, which is strictly more than their equilibrium payoffs $u(x)$ and $v(y)$. At the same time if x and y are matched at equilibrium under π^{eq}, then feasibility imposes that $u(x) + v(y) = \Phi(x, y)$. Thus, taking expectations of both sides with respect to π^{eq} will get

$$\mathbb{E}_{\pi^{eq}}[\Phi(X, Y)] = \mathbb{E}_{\pi^{eq}}[u(X) + v(Y)] = \mathbb{E}_{\mu_1}[u(X)] + \mathbb{E}_{\mu_2}[v(Y)]. \tag{1.3}$$

Hence, $\pi^{eq} \in \Pi(\mu_1, \mu_2)$ is defined as an equilibrium matching whenever there exists functions u and v such that both conditions (1.2) and (1.3) hold.

Let us now show that if π^{eq} is an equilibrium, then it is a solution of (1.1). Consider π^{opt} a solution of problem (1.1). Taking expectations of both sides of (1.2) with respect to π^{opt} gets

$$\mathbb{E}_{\pi^{opt}}[\Phi(X,Y)] \leq \mathbb{E}_{\pi^{opt}}[u(X) + v(Y)] = \mathbb{E}_{\mu_1}[u(X)] + \mathbb{E}_{\mu_2}[v(Y)],$$

where the latter equality comes from the fact that $\pi^{opt} \in \Pi(\mu_1, \mu_2)$. Hence, $\mathbb{E}_{\pi^{opt}}[\Phi(X,Y)] \leq \mathbb{E}_{\pi^{eq}}[\Phi(X,Y)]$, but by definition of π^{opt}, these two quantities coincide and π^{eq} is optimal for the centralized problem (1.1). Hence, the decentralized solution (equilibrium matching) coincides with the centralized solution (optimal matching).

However, the analysis above assumes that the existence of a matching between two partners is purely deterministic given partners' observed characteristics, which is not realistic. In order to allow for some randomness arising from agent's unobserved heterogeneity in the matching process, we shall make use of a regularized version of the optimization formulation (1.1) in order to perform the estimation of Φ.

1.3 Modeling heterogeneity

It is a well-known result in optimal transport theory (see [34], Chapter 9) that, under suitable assumptions on Φ, the optimal matching will be *pure*, in the sense that any x is matched deterministically to a unique $y = T(x)$ for some bijective map T; in other words, the conditional distribution $\pi(y|x)$ of y given x, is reduced

to a single point mass. Clearly, in the presence of unobserved heterogeneity, this is no longer the case. Our approach to modeling uncertainty consists in adding an entropic regularization term in (1.1), leading to

$$\max_{\pi \in \Pi(\mu_1,\mu_2)} \mathbb{E}_\pi \left[\Phi\left(X,Y\right) - \sigma \ln \pi\left(X,Y\right)\right] \tag{1.4}$$

where $\sigma > 0$ is a temperature parameter, so that setting $\sigma = 0$ recovers program (1.1).

Recently a number of authors have studied such a regularized version of the Monge-Kantorovich problem (see for instance [4]; [18] and references therein). One notable feature of (1.4) is that the optimal matching $\pi\left(x,y\right)$ has form

$$\pi\left(x,y\right) = a\left(x\right) b\left(y\right) \exp\left(\Phi\left(x,y\right)/\sigma\right),$$

where $a\left(x\right)$ and $b\left(y\right)$ are set by imposing the constraint $\pi \in \Pi\left(\mu_1,\mu_2\right)$, that is

$$\begin{cases} \int a\left(x\right) b\left(y\right) \exp\left(\Phi\left(x,y\right)/\sigma\right) dy = \mu_1\left(x\right) \\ \int a\left(x\right) b\left(y\right) \exp\left(\Phi\left(x,y\right)/\sigma\right) dx = \mu_2\left(y\right) \end{cases}.$$

As a result, $a\left(x\right)$ and $b\left(y\right)$ can be obtained by the iterated proportional fitting procedure (IPFP), a.k.a. Sinkhorn's algorithm, which is presented in algorithm 1.

Algorithm 1 IPFP

Input: $b(y), \mu_1(x), \mu_2(y), \Phi(x,y), \sigma$
while not converged **do**
  $a\left(x\right) \leftarrow \mu_1\left(x\right) / \int b\left(y\right) \exp\left(\Phi\left(x,y\right)/\sigma\right) dy$
  $b\left(y\right) \leftarrow \mu_2\left(y\right) / \int a\left(x\right) \exp\left(\Phi\left(x,y\right)/\sigma\right) dx$
end while
Return: $a(x), b(y)$

1.4 Parameterization of the affinity function

We assume the simple parameterization of Φ as a bilinear form associated to some *affinity matrix* A, namely

$$\Phi_A(x, y) = x^\top A y. \tag{1.5}$$

This functional form will capture the interaction effects between the various dimensions of the characteristics. The sign of A_{ij} indicates that there is attractive (if positive $A_{ij} > 0$) or repulsive (if $A_{ij} < 0$) energy between coordinate i of x and coordinate j of y. On the contrary, $A_{ij} = 0$ means that there is no interaction between x_i and y_j.

By positive homogeneity, we can normalize the temperature parameter σ in front of the entropic term to $\sigma = 1$. Indeed, the solution of the problem with affinity function Φ and temperature σ coincides with the solution of the problem with affinity function Φ/σ and temperature one. Hence, we define

$$\mathcal{W}(A) = \max_{\pi \in \Pi(\mu_1, \mu_2)} \mathbb{E}_\pi \left[\Phi_A(X, Y) - \ln \pi(X, Y)\right]. \tag{1.6}$$

As before, the optimal matching π^A retains the form

$$\pi^A(x, y) = a(x)\, b(y) \exp\left(\Phi_A(x, y)\right), \tag{1.7}$$

where $a(x)$ and $b(y)$ are computed by the IPFP algorithm 1. It follows directly

from expression (1.7) that

$$\frac{\partial^2 \ln \pi^A(x,y)}{\partial x_i \partial y_j} = A_{ij},$$

which provides a nice interpretation of A as the matrix of cross-derivatives of the log-likelihood of a matched (x,y) pair. In the sequel, we shall focus on the estimation of the affinity matrix A.

1.5 Maximum likelihood estimation of the affinity matrix

We would like to estimate A based on an i.i.d. sample of matched pairs $\left(x^{(k)}, y^{(k)}\right)$, $k = 1, ..., N$, where $x^{(k)}$ and $y^{(k)}$ are respectively d and d'-dimensional vectors of characteristics, and the observed matching is defined as

$$\hat{\pi}(x,y) = \frac{1}{N}\sum_{k=1}^{N} \delta\left(x - x^{(k)}\right)\delta\left(y - y^{(k)}\right).$$

1.5.1 Unconstrained maximum likelihood

As implied by the next result, the likelihood function turns out to have a particularly tractable form and is globally concave.

Proposition 1. *(a) The log-likelihood $l(A;\hat{\pi})$ of observation $\hat{\pi}$ at parameter value A is given by*

$$l(A;\hat{\pi}) = N\mathbb{E}_{\hat{\pi}}\left[\log \pi^A(X,Y)\right] = N\left(\mathbb{E}_{\hat{\pi}}\left[\Phi_A(X,Y)\right] - \mathcal{W}(A)\right). \tag{1.8}$$

(b) It is a concave function of A, and its gradient is given by

$$\nabla l\left(A;\hat{\pi}\right)=N\left(\mathbb{E}_{\hat{\pi}}\left[X_iY_j\right]-\mathbb{E}_{\pi^A}\left[X_iY_j\right]\right). \tag{1.9}$$

Proof. (a) The log-likelihood of a pair $\left(x^{(k)},y^{(k)}\right)$ is given by $\log\pi^A\left(x^{(k)},y^{(k)}\right)$. As the pairs are independently sampled, the log-likelihood of the matching $\hat{\pi}$ is given by $l\left(A;\hat{\pi}\right)=\sum_{k=1}^{N}\log\pi^A\left(x^{(k)},y^{(k)}\right)=N\mathbb{E}_{\hat{\pi}}\left[\log\pi^A\left(X,Y\right)\right]$. It follows from (1.7) that $\mathbb{E}_{\hat{\pi}}\left[\log\pi^A\left(X,Y\right)\right]=\mathbb{E}_{\hat{\pi}}\left[\Phi_A\left(X,Y\right)\right]+\mathbb{E}_{\hat{\pi}}\left[\log a\left(X\right)+\log b\left(Y\right)\right]$, but as π^A and $\hat{\pi}$ both belong to $\Pi\left(\mu_1,\mu_2\right)$, it follows that

$$\begin{aligned}
&\mathbb{E}_{\hat{\pi}}\left[\log\pi^A\left(X,Y\right)\right]\\
=\;&\mathbb{E}_{\hat{\pi}}\left[\Phi_A\left(X,Y\right)\right]-\mathbb{E}_{\pi^A}\left[\Phi_A\left(X,Y\right)\right]+\mathbb{E}_{\pi^A}\left[\Phi_A\left(X,Y\right)\right]+\\
&\mathbb{E}_{\pi^A}\left[\log a\left(X\right)+\log b\left(Y\right)\right]\\
=\;&\mathbb{E}_{\hat{\pi}}\left[\Phi_A\left(X,Y\right)\right]-\mathbb{E}_{\pi^A}\left[\Phi_A\left(X,Y\right)\right]+\mathbb{E}_{\pi^A}\left[\log\pi^A\left(X,Y\right)\right]\\
=\;&\mathbb{E}_{\hat{\pi}}\left[\Phi_A\left(X,Y\right)\right]-\mathcal{W}\left(A\right),
\end{aligned}$$

hence $l\left(A;\hat{\pi}\right)=N\left\{\mathbb{E}_{\hat{\pi}}\left[\Phi_A\left(X,Y\right)\right]-\mathcal{W}\left(A\right)\right\}$.

(b) $\mathbb{E}_{\hat{\pi}}\left[\Phi_A\left(X,Y\right)\right]$ is linear in A, and $\mathcal{W}\left(A\right)$ is convex in A, so $l\left(A;\hat{\pi}\right)$ is concave. By the envelope theorem, $\nabla l\left(A;\hat{\pi}\right)=N\left\{\mathbb{E}_{\hat{\pi}}\left[X_iY_j\right]-\mathbb{E}_{\pi^A}\left[X_iY_j\right]\right\}$. □

Thus, conditions (1.9) imply that the maximum likelihood estimator $\hat{A}$ should solve

$$\mathbb{E}_{\pi^A}\left[X_iY_j\right]=\mathbb{E}_{\hat{\pi}}\left[X_iY_j\right] \tag{1.10}$$

for every pair $i\in\{1,\cdots,d\}$ and $j\in\{1,\cdots,d'\}$, which thus turns out to be equivalent to the moment matching procedure of Dupuy and Galichon (2014).

Hence, assuming w.l.o.g. that X and Y are centered at 0, this implies that $\hat{A}$ is the value of the parameter such that the predicted covariance matrix $cov_{\pi^A}(X, Y)$ will match the observed one $cov_{\hat{\pi}}(X, Y)$.

One important advantage of the concavity of the log-likelihood function $l(A; \hat{\pi})$ is that various additional regularizations can be incorporated into the estimation procedure. One could constrain A to be entry-wise nonnegative so that only attractive interactions are considered. One could also assume A is sparse, so that only a small number of pairs of characteristics interact. In this chapter, we are concerned with the case when only a small number of coordinates, which are linear combinations of the characteristics, interact. One shall then need to impose a requirement that the rank of A is small. The next sections propose an effective method for doing so which is implemented on two marriage market datasets.

1.6 Low-rank regularization

In some situations, two scalar dimensions $\tilde{x}$ and $\tilde{y}$, obtained linearly from x and y via $\tilde{x} = u^\top x$ and $\tilde{y} = v^\top y$, suffice to explain the solution $\hat{\pi}$, where u and v are two unit vectors of weights. In this case, A is simply a scalar multiple of rank one matrix $uv^\top$. More generally, when the rank of A is equal to r, the singular value decomposition (SVD) of A yields

$$A = USV^\top, \tag{1.11}$$

where S is a diagonal $r \times r$ matrix with strictly positive diagonal entries (called singular values) in the decreasing order, and U and V are two semi-orthogonal $d \times r$ matrices. In this case, the total interaction term is $x^\top Ay = \tilde{x}^\top S\tilde{y}$, where

$\tilde{x} = U^\top x$ and $\tilde{y} = V^\top y$ are the relevant dimensions of interaction. Note that $x^\top A y$ requires to sum over $d \times d'$ interaction terms whereas $\tilde{x}^\top S \tilde{y}$ only requires to sum over $r \leq \min\{d, d'\}$ interaction terms. Moreover, each singular value can be interpreted as the weight of the interaction between the corresponding relevant dimensions of $\tilde{x}$ and $\tilde{y}$ in the total interaction term.

One can incorporate the rank constraint into the maximization of the likelihood, whose expression is given in proposition 1, yielding

$$\begin{aligned} \max_A \quad & l(A; \hat{\pi}) \\ \text{s.t.} \quad & rk(A) \leq r. \end{aligned}$$

However, the general rank-constrained problem is non-convex and NP-hard, see [14]. A natural convex relaxation of the problem is done by replacing the rank of A by its nuclear norm (see e.g. [14, 25]), $\|A\|_*$, defined as the sum of the singular values of A. This yields a modified formulation of the problem as

$$\min_A \left\{ \mathcal{W}(A) - \mathbb{E}_{\hat{\pi}} \left[\Phi_A(X, Y) \right] + \lambda \|A\|_* \right\}, \tag{1.12}$$

where $\lambda \geq 0$ is the Lagrange multiplier of the nuclear norm constraint. Measuring the complexity of the model by the rank of the affinity matrix, equation (1.12) indicates that for $\lambda = 0$, one accepts the full complexity of the model and performs exact likelihood maximization whereas, for large values of λ, one simplifies the model and deviates from exact likelihood maximization. Hence, the parameter λ can be thought of as a parameter controlling the trade-off between exact likelihood maximization and the complexity of the model.

The computation for problems involving the nuclear norm can be efficiently

carried out using the proximal gradient descent method with guaranteed convergence (see e.g. [31]). As noted in the previous section, $l(A;\hat{\pi})$ is continuously differentiable with respect to A, and its gradient is given in expression (1.9). We now describe our complete procedure in algorithm 2.

Algorithm 2 Proximal gradient descent algorithm for problem (1.12)

Input: A, step size t, matched pairs $\left(x^{(k)}, y^{(k)}\right)$, $k = 1, ..., N$
while not converged **do**
  Using the IPFP algorithm 1 to compute the optimal matching π^A
  $A \leftarrow A - t\left(\sum_{i,j=1}^{N}(\pi_{ij}^A - \hat{\pi}_{ij})x^{(i)}(y^{(j)})^\top\right)$
  $[U, \mathsf{diag}(s_1, \cdots, s_d), V] = \mathsf{SVD}(A)$
  $A \leftarrow U\mathsf{diag}((s_1 - t\lambda)_+, \cdots, (s_d - t\lambda)_+)V^\top$
end while
Return: A

Additionally, we note that the nuclear norm regularization prevents overfitting the covariance mismatch $\|\mathbb{E}_{\pi^A}\left[XY^\top\right] - \mathbb{E}_{\hat{\pi}}\left[XY^\top\right]\|_F$, where $\|\cdot\|_F$ is the Frobenius norm of a matrix and which one recalls from expression (1.10) will be exactly equal to 0 without the nuclear norm regularization. Indeed, given U and V defined in expression (1.11), the first order optimality conditions [36] are

$$\mathbb{E}_{\pi^A}\left[XY^\top\right] - \mathbb{E}_{\hat{\pi}}\left[XY^\top\right] + \lambda UV^\top + N = 0, \tag{1.13}$$

where N satisfies $U^\top N = 0$, $NV = 0$, and $\|N\|_2 \leq \lambda$, with $\|N\|_2$ being the *spectral norm* of N. Equation (1.13) indicates that $\mathbb{E}_{\pi^A}\left[XY^\top\right] - \mathbb{E}_{\hat{\pi}}\left[XY^\top\right]$ and A have simultaneous SVDs. Moreover, the singular values of $\mathbb{E}_{\pi^A}\left[XY^\top\right] - \mathbb{E}_{\hat{\pi}}\left[XY^\top\right]$ corresponding to the strictly positive singular values of A will be exactly equal to λ, while the ones corresponding to the zero singular values of A will be less than or equal to λ. Thus, by varying λ, the covariance mismatch, which equals the l_2-norm of the singular values of $\mathbb{E}_{\pi^A}\left[XY^\top\right] - \mathbb{E}_{\hat{\pi}}\left[XY^\top\right]$, will change as well.

We select the best λ by repeating a five-fold cross-validation (CV) twice, resulting in ten different experiments. In each of the CV procedure, the whole dataset is randomly split into five parts with equal size. For each λ, we estimate A via (1.12) using 4 parts and record both $\mathcal{W}(A) - \mathbb{E}_{\hat{\pi}}[\Phi_A(X,Y)]$ and $\|\mathbb{E}_{\pi^A}\left[XY^\top\right] - \mathbb{E}_{\hat{\pi}}\left[XY^\top\right]\|_F$ evaluated on the remaining part. From this we obtain an estimated prediction error curve as a function of λ, and we select the λ value that minimizes both errors.

1.7 Application to marriage market data

We apply the low-rank optimal transport method to the case of bipartite matching in the marriage market. We revisit the data set used in [12]. This data confronts the analyst with the problem of selecting from a large set of observed characteristics of spouses, those that are important for matching affinities.

In this application, the analyst faces the difficult task of estimating an affinity matrix whose size is large, being the product of the number of observed characteristics of spouses, relative to the number of observations. The high ratio of parameters to observations creates overfitting concerns. A solution would be to construct combinations of the observed characteristics prior to the estimation, hence reducing the number of parameters of the associated affinity matrix. However, the construction of these combinations of characteristics requires the analyst to define weights based on prior information about matching affinity. In contrast, our low-rank optimal transport method allows the analyst to simultaneously estimate the affinity matrix while selecting the relevant combinations of characteristics using weights derived from the information contained in the affinity matrix itself.

1.7.1 Personality traits

We use the Dutch Household Survey (DHS) ran by the Dutch National Bank. In particular, a representative sample of 1,155 young couples observed in the period 1993-2002 in the Netherlands was constructed following the procedure outlined in [12]. In this sample, the analyst has access to detailed information about spouses' characteristics such as education, height, Body Mass Index (BMI)[2] and subjective health, but also about personality traits and attitude towards risk. Personality traits are herewith recovered by administrating the 16 Personality Factors test (16PF test) to respondents. This test consists in a 16-item questionnaire where each item corresponds to a primary factor describing a facet of one's personality. Attitude towards risk is recovered using a similar approach (see e.g. [11]). A 6-item questionnaire about risk attitude is administrated to the respondents, each item corresponding to a primary factor describing a facet of one's attitude towards risks.

In this application, the objective is to estimate matching affinities from the sample of 1,155 couples with characteristics (X, Y), where X and Y contain each 26 variables: education, height, BMI, subjective health and the 16 primary factors of personality traits and 6 primary factors of risk attitude. The associated affinity matrix has $26 \times 26 = 676$ parameters to be estimated, hence a ratio of 0.58 parameters per observation. In [12], the authors substantially reduced the dimensionality of the model by constructing 5 global factors of personality traits and 1 global factor of attitude towards risk. They relied on the psychology literature that shows that 5 global factors, often referred to as the "big 5," providing an overview of one's personality can be derived from the primary factors of the 16PF

[2]Weight in Kg divided by the square of height in meters.

using methods such as Factor Analysis. These 5 global factors are (orthogonal) linear combinations of the 16 primary factors. Similarly, as is standard in the economic literature (see e.g. [11]), a single global factor providing an overview of attitude towards risk can be derived as a linear combination of the underlying 6 primary factors. As a result, the authors in [12] were able to estimate a reduced affinity matrix of dimension $10 \times 10 = 100$, with a ratio of 0.09 parameters per observation. However, this requires to assume that i) either all or none of the primary factors belonging to a global factor matter and ii) their relative importance is proportional to their relative weight in the global factor. There are no reasons to expect this should hold universally since the weights used to create the global factors are chosen so as to provide an overview of an individual's personality or attitude towards risk and not to capture matching affinities. In contrast, our low-rank optimal transport approach, allows us to estimate the affinity matrix of size 676 associated with the primary factors while creating the relevant combinations of these factors that matter for matching affinities.

We use the low-rank optimal transport approach to estimate the affinity matrix A when considering characteristics including the primary factors. Inspection of Figure 1.1 indicates that $\lambda = 0.15$ gives slightly lower values of the CV errors of both $\mathcal{W}(A) - \mathbb{E}_{\hat{\pi}}[\Phi_A(X,Y)]$ and $\|\mathbb{E}_{\pi^A}\left[XY^\top\right] - \mathbb{E}_{\hat{\pi}}\left[XY^\top\right]\|_F$ than $\lambda = 0.1$ does. Since $\lambda = 0.15$ achieves this result with a lower rank of A, we use this value as the coefficient for the nuclear norm regularization. The left panel of Figure 1.2 reveals the rank of the affinity matrix is 12 hence indicating that only 12 relevant dimensions matter for matching affinities. Of those 12 relevant dimensions, the first three alone explain about 50% of the total matching affinity as indicated in the right panel of Figure 1.2.

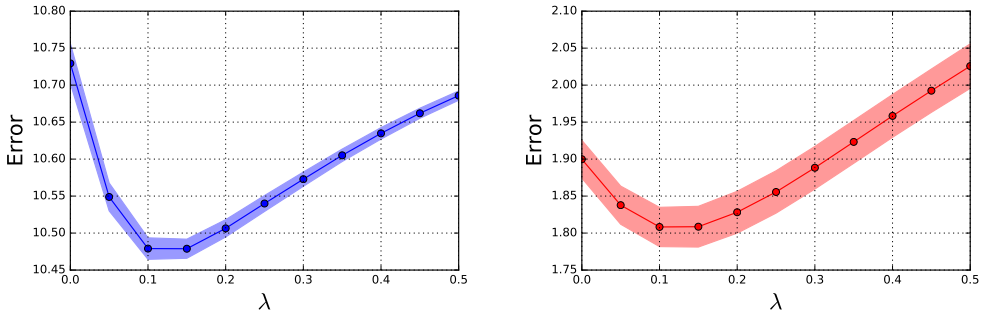

Figure 1.1: Errors of (left) the negative log-likelihood $\mathcal{W}(A) - \mathbb{E}_{\hat{\pi}}\left[\Phi_A(X,Y)\right]$ (right) the covariance mismatch $\|\mathbb{E}_{\pi^A}\left[XY^\top\right] - \mathbb{E}_{\hat{\pi}}\left[XY^\top\right]\|_F$.

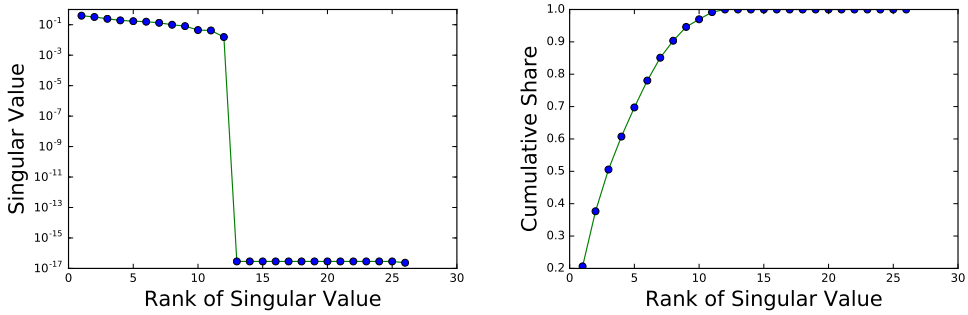

Figure 1.2: (Left) singular values (right) cumulative shares.

The loadings of the first three dimensions, reported in Table 1.1, reveal several important results. First, as in [12], we do find that the first relevant dimension loads principally on education, i.e. 0.85 for men and 0.83 for women respectively, whereas the second and third dimensions load principally on personality traits and attitude towards risk. However, using the primary factors rather than the global factors as in [12], we find that although conscientiousness matters for both men and women, the underlying primary factors at play differ across gender. For women, the primary factor "easily hurt, offended," belonging to the global factor conscientiousness, is the most important characteristic in the second dimension with a loading of 0.71. For men, the primary factor "easily hurt, offended" plays also an important role (loading of magnitude 0.42), but the primary factor "disciplined," also belonging to the global factor conscientiousness, is the most important one with a loading of 0.52. These results clearly illustrates that although conscientiousness matters, not all of its primary constituents do and different aspects matter differently for men and women.

A similar type of results holds for the third dimension, which loads on some but not all of the items measuring attitude towards risk. However, this dimension also loads on other variables such as height, BMI and subjective health, making its interpretation more difficult.

1.8 Conclusion and future research

In this chapter, we have demonstrated the effectiveness of rank-constrained estimation techniques when solving inverse optimal transport problems. Inverse optimal transport problems are often faced with large dimensionality of the data

Singular value	$s_1 = 0.39$		$s_2 = 0.32$		$s_3 = 0.24$	
Singular vector	U_1	V_1	U_2	V_2	U_3	V_3
Oriented toward people	-0.07	-0.04	-0.08	0.19	0.04	-0.08
Quick thinker	0.08	-0.01	0.25	-0.12	-0.04	0.05
Not easily worried	-0.10	0.00	0.29	0.11	0.08	-0.07
Stubborn, persistent	0.06	0.03	0.11	0.14	-0.11	0.02
Vivid, vivacious	0.00	0.02	0.04	0.19	-0.23	0.11
Meticulous	-0.12	-0.05	-0.01	0.04	0.16	0.17
Dominant	0.05	0.06	-0.09	-0.21	0.00	0.08
Easily hurt, offended	-0.06	-0.03	0.42	0.71	0.02	-0.02
Suspicious	0.08	0.01	0.16	0.14	0.11	0.04
Dreamer	-0.04	0.02	-0.08	0.23	0.00	0.04
Diplomatic, tactful	-0.06	0.07	0.07	-0.10	0.00	-0.03
Doubts about myself	0.06	0.13	-0.31	-0.35	0.17	0.09
Open to changes	0.10	0.03	0.12	0.04	-0.22	0.05
Independent, self-reliant	-0.10	-0.11	0.31	0.04	0.01	-0.09
Disciplined	0.01	0.01	0.52	0.17	0.01	-0.11
Irritable, quick tempered	0.00	-0.16	0.08	-0.17	-0.02	-0.03
Ready to take risk for high possible returns	0.17	0.24	-0.02	-0.07	0.27	0.29
Investments in shares are too risky	-0.31	-0.29	0.05	-0.07	0.48	0.53
Ready to borrow money for risky investment	0.12	0.13	-0.16	-0.03	-0.17	-0.08
Want to be certain my investments are safe	0.01	0.05	-0.12	0.02	0.42	0.24
Should take greater financial risks	-0.06	-0.07	-0.03	-0.05	-0.11	-0.18
Ready to risk losing money to gain money	0.10	0.09	0.06	-0.02	-0.38	-0.48
Educational level	0.85	0.83	0.12	0.12	0.27	0.21
Height	0.06	0.08	0.12	0.01	-0.12	-0.25
BMI	-0.20	-0.24	0.17	0.18	0.16	0.29
Subjective health	-0.01	0.01	-0.15	-0.06	-0.21	-0.17

Table 1.1: Loadings of the top three relevant dimensions of matching affinities, Dutch couples.

sets; hence it is crucial to develop dimensionality reduction techniques. We plan to investigate further applications of this methodology, including explaining the intensity of mercantile exchanges between countries by the similarities in their characteristics, predicting stable matches in online dating platforms, or understanding the determinants of workers' productivity on the labor market. We also plan to consider an extension of the present methodology to nonbipartite networks, which will allow to estimate the transport costs in minimum cost flow problems, with applications to analyzing urban transportation demand, as well as link formation in social networks.

Chapter 2

Optimal transport methods for gravity equations

In the previous chapter we formulate inverse optimal transport problems, in which we are given the observation of an optimal matching and seek the affinity function for which this matching is optimal. The application to the marriage market is then investigated. In this chapter, we discuss the application of optimal transport tools to another important set of inverse problems, the estimation of gravity models in bilateral trade, where given observed trade flows, we want to identify the most important factors that drive such activities. We introduce a common formulation of the gravity model for bilateral trade and discuss two most widely used estimation techniques in the literature, Ordinary Least Squares (OLS) and Pseudo Poisson Maximum Likelihood (PPML). The most innovative part of this chapter is the theoretical establishment of the connections between optimal transport methods and PPML by showing their equivalence in estimating the gravity model. Such equivalence will enable us to take advantage of the recent

advances in computational optimal transport to solve large-scale bilateral trade problems.

2.1 Introduction

Newton's gravity equation in Physics is known as

$$F_{ij} = G\frac{M_i M_j}{D_{ij}^2},$$

where F_{ij} is the gravitational force between objects i and j, G is the gravitational constant, M_i and M_j are object i's and j's masses respectively, and D_{ij} is the distance between them. Therefore, the force of gravity between two objects i and j are positively related to M and negatively related to D. The pioneering work of [30] initiated a vast theoretical and empirical literature on the gravity equation for trade. Theories based on different foundations for trade all predict a gravity relationship for trade flows analogous to Newton's law of universal gravitation. In its simplest form, the gravity equation for trade reads as follows

$$X_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3}, \tag{2.1}$$

where $\alpha_0, \alpha_1, \alpha_2$, and α_3 are unknown parameters to be estimated. Eq. (2.1) states that the trade flow from country i to country j, denoted by X_{ij}, is proportional to the product of the two countries' GDPs, denoted by Y_i and Y_j, and is inversely proportional to their distance, D_{ij}, broadly constructed to include all factors that might create trade resistance.

However, instead of Eq. (2.1) holding with certainty, we only expect it to hold

in expectation. Thus, to account for such deviations from theory, we require the relationship to hold in expectation

$$\mathbb{E}[X_{ij}|Y_i, Y_j, D_{ij}] = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3}. \tag{2.2}$$

The authors of [2] argue that the traditional gravity equation is not correctly specified, as it does not take into account the possibility of multilateral resistance. That is, the benchmark gravity model above does not account for the potential influence of, say, a third economy that trades with both country i and country j, on the trading flow between i and j. As a solution, they suggest to augment Eq. (2.2) with exporter and importer fixed effects, leading to

$$\mathbb{E}[X_{ij}|Y_i, Y_j, D_{ij}] = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} e^{\theta_1 \mathbf{1}_i + \theta_2 \mathbf{1}_j}, \tag{2.3}$$

where the new parameters θ_1 and θ_2 are to be estimated, and $\mathbf{1}_i$ and $\mathbf{1}_j$ are dummies identifying the exporter and importer.

Similar to Newton's gravity equation, we expect the signs of α_1 and α_2 to be positive (i.e., the trade flow between two countries with large GDP's is expected to be higher). We also expect the sign of α_3 to be negative, because the farther apart two countries are, the lower the trade flow between them. The majority of the gravity literature now follow this model specification albeit minor adjustments.

2.2 OLS

The multiplicative form of Eq. (2.3) is convenient. Indeed, there is a long tradition in the trade literature of taking a log-linearization step and estimating

the parameters of interest via least squares, using the equation

$$\ln X_{ij} = \ln \alpha_0 + \alpha_1 \ln Y_i + \alpha_2 \ln Y_j + \alpha_3 \ln D_{ij} + \theta_1 \mathbf{1}_i + \theta_2 \mathbf{1}_j + \varepsilon_{ij}, \tag{2.4}$$

where ε_{ij} are error terms satisfying $\mathbb{E}[\varepsilon_{ij}|Y_i, Y_j, D_{ij}] = 0$.

Since the whole model is now linear, one can perform OLS to estimate the parameters, i.e. by minimizing the objective function

$$\sum_{ij} \left[\ln X_{ij} - (\ln \alpha_0 + \alpha_1 \ln Y_i + \alpha_2 \ln Y_j + \alpha_3 \ln D_{ij} + \theta_1 \mathbf{1}_i + \theta_2 \mathbf{1}_j)\right]^2 .$$

However, the OLS approach suffers from two problems. Firstly, although the gravitational force in Newton's gravity model can be very small, it is never zero. This does not hold true for bilateral trade flows. For example, it should not come as a surprise that two countries with small GDPs that are geographically far away from each other did not trade in a certain year. The presence of zero trade flows invalidates the log-linearization step.

The second problem is caused by the violation of the assumptions of OLS, which requires the errors ε_{ij} to be uncorrelated with the regressors. However, the authors in [27] find overwhelming evidence that the error terms in the usual log-linear specification of the gravity equation are heteroskedastic, which leads to inconsistent estimates.

2.3 Generalized linear models

Before we delve into PPML, we provide a brief overview of the generalized linear model (GLM) (see e.g. [23, 37]). The GLM generalizes linear regression

by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. More specifically, we model each outcome y as to be generated from a particular distribution in the exponential family, which includes the normal, binomial, Poisson and Gamma distributions. The mean, μ, of the distribution of y depends on the independent variables, X, through the following relation

$$\mathbb{E}[y] = \mu = g^{-1}(X\beta),$$

where g is the link function.

In this framework, the variance is typically a function, V, of the mean

$$\text{Var}(y) = V(\mu) = V(g^{-1}(X\beta)). \tag{2.5}$$

For the standard normal, $V(\mu) = 1$, for the Bernoulli, $V(\mu) = \mu(1-\mu)$, and for the Poisson, $V(\mu) = \mu$.

The GLM literature realized early on that assumption (2.5) was too restrictive. As discussed in [37], no variance assumption is needed for the consistent estimation of β. Thus, the following relaxation of (2.5) is used

$$\text{Var}(y) = \sigma_o^2 V(g^{-1}(X\beta)), \tag{2.6}$$

where $\sigma_o^2 > 0$ is called the dispersion parameter. (2.6) is referred to as the GLM variance assumption.

2.4 PPML

To address the two issues associated with OLS, the authors in [27] propose to model each trade flow X_{ij} as a Poisson random variable with rate parameter λ_{ij}

$$X_{ij} \sim \text{Poisson}(\lambda_{ij}).$$

The motivation is to express X_{ij} as a linear combination of the regressors, followed by a transformation via the link function $g^{-1}(z) = \exp(z)$. To see the connection with Eq. (2.3), we group the country specific terms as

$$s_i = -(\alpha_1 \ln Y_i + \theta_1 \mathbf{1}_i) \quad \text{and} \quad m_j = -(\alpha_2 \ln Y_j + \theta_2 \mathbf{1}_j).$$

We also define

$$\phi_{ij} = \ln \alpha_0 + \alpha_3 \ln D_{ij}.$$

Eq. (2.3) then becomes

$$\mathbb{E}[X_{ij}|s_i, m_j, \phi_{ij}] = \exp(\phi_{ij} - s_i - m_j), \tag{2.7}$$

which indicates if one sets the rate parameter $\lambda_{ij} = \exp(\phi_{ij} - s_i - m_j)$, the Poisson distribution yields

$$\mathbb{E}[X_{ij}|s_i, m_j, \phi_{ij}] = \lambda_{ij} = \exp(\phi_{ij} - s_i - m_j).$$

The variables s_i and m_j are called the fixed effects because they are specific to each importer and exporter countries. Also, they are relaxed to be unobserved

and need to be estimated instead of being observed as in the original gravity equation (2.2). Additionally, ϕ_{ij} is often specified as a linear combination of pairwise measures of distance between the importer and exporter countries

$$\phi_{ij}^{\beta} = \sum_{k} \beta_k d_{ij}^k,$$

where the superscript β indicates the dependence of β. Standard choices of d_{ij}^k include

1. geographical bilateral distance between i and j.
2. indicator of contiguous borders; of common official language; of colonial ties.
3. trade policy variables such as the presence of a regional trade agreement or tariffs.

Therefore, the mean of X_{ij} is now linear in the parameters β_k, s_i, and m_j. For notational simplicity we will write

$$\mathbb{E}[X_{ij}|\boldsymbol{\beta}] = \exp(\mathbf{d}_{ij}^{\top}\boldsymbol{\beta}), \tag{2.8}$$

where $\boldsymbol{\beta}$ combines all the β_k, s_i, and m_j, and $\mathbf{d}_{ij}$ without the superscript combines all the measured regressors d_{ij}^k and dummy variables associated with s_i and m_j. We now present estimation methods for Eq. (2.8).

2.4.1 Estimation

The first naive model is non-linear least squares (NLS), which is defined via

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{ij} (\hat{X}_{ij} - \exp(\mathbf{d}_{ij}^{\top}\boldsymbol{\beta}))^2, \tag{2.9}$$

which implies the following set of first-order conditions

$$\sum_{ij} (\hat{X}_{ij} - \exp(\mathbf{d}_{ij}^{\top}\hat{\boldsymbol{\beta}})) \exp(\mathbf{d}_{ij}^{\top}\hat{\boldsymbol{\beta}})\mathbf{d}_{ij} = \mathbf{0}. \tag{2.10}$$

We can see from Eq. (2.10) that this estimator gives more weight to observations where $\exp(\mathbf{d}_{ij}^{\top}\hat{\boldsymbol{\beta}})$ is large. If the model is specified as

$$X_{ij} = \exp(\mathbf{d}_{ij}^{\top}\boldsymbol{\beta}) + \varepsilon_{ij},$$

then essentially Eq. (2.9) assumes the variance of ε_{ij} is constant, which is invalid because the errors are heteroscedastic. The Poisson distribution requires

$$\mathbb{E}[X_{ij}|\boldsymbol{\beta}] = \mathrm{Var}(X_{ij}|\boldsymbol{\beta}),$$

which we call the Poisson variance assumption. A weaker assumption, which we call the Poisson GLM variance assumption, is

$$\mathbb{E}[X_{ij}|\boldsymbol{\beta}] = \sigma_{\mathrm{o}}^2 \mathrm{Var}(X_{ij}|\boldsymbol{\beta}).$$

This kind of model specification is particularly suitable for accounting for the pattern of heteroskedasticity, since it tends to produce highly dispersed errors.

Then the objective function becomes

$$\sum_{ij}(\hat{X}_{ij} - \exp(\mathbf{d}_{ij}^{\top}\hat{\boldsymbol{\beta}}))\mathbf{d}_{ij} = \mathbf{0},$$

which means we give equal weights to all the observations. Without further information on the pattern of heteroskedasticity, it seems natural to give the same weight to all observations. It is shown in [37] that although the Poisson variance assumption is weakened, the estimator obtained is still consistent. In the context of the gravity model, while OLS tends to underweight smaller economies/observations with smaller values, PPML allows us to treat all observations more equally. Since the establishment of PPML as a more viable estimation method, PPML has been widely employed in the gravity literature.

2.5 Connection with optimal transport

In this section we want to show the equivalence between the regularized Monge-Kantorovich problem and PPML by showing they minimize the same objective function. We start with Eq. (1.6)

$$\mathcal{W}(A) = \max_{\pi\in\Pi(\mu_1,\mu_2)} \mathbb{E}_{\pi}\left[\Phi\left(X,Y\right) - \sigma\ln\pi\left(X,Y\right)\right]. \tag{2.11}$$

In the trade context, the above can be re-written as

$$\begin{aligned} \mathcal{W}(\beta) = \max_{X_{ij}\geqslant 0}&\left\{\sum_{ij} X_{ij}\phi_{ij}^{\beta} - \sigma\sum_{ij} X_{ij}\ln X_{ij}\right\} \\ \text{s.t.}\sum_{i} X_{ij} = Y_j \quad &\text{and} \quad \sum_{j} X_{ij} = X_i, \end{aligned} \tag{2.12}$$

where X_i and Y_j are the marginals representing the value of production and importers' expenditures, respectively. By duality, we can express $\mathcal{W}(\beta)$ as

$$\mathcal{W}(\beta) = \min_{s_i, m_j} \left\{ \sum_i s_i Y_i + \sum_j m_j X_j + \sigma \sum_{ij} \exp\left(\frac{\phi_{ij}^{\beta} - s_i - m_j}{\sigma} \right) \right\}. \tag{2.13}$$

Without loss of generality, we can set $\sigma = 1$.

Eq. (1.12) becomes

$$\min_{s_i, m_j} \min_{\beta} \left\{ \mathcal{W}(\beta) - \mathbb{E}_{\hat{X}_{ij}}[\phi_{ij}^{\beta}] \right\}, \tag{2.14}$$

with the nuclear norm regularization term omitted.

We show that this is exactly the same objective function as PPML is minimizing. In PPML we maximize the likelihood of the observations. Equivalently, this is to choose optimal parameters so that the negative log-likelihood is minimized. We recall the probability distribution of a Poisson random variable with rate parameter λ has the form $f_\lambda(z) = e^{-\lambda}\lambda^z / z!$. Thus, we can write the negative log likelihood of observing all the $\hat{X}_{ij}$'s as, up to a constant,

$$\begin{aligned}
-\ln L(\hat{X}; \boldsymbol{\beta}) &= -\sum_{ij} \ln \left(\frac{e^{-\lambda_{ij}} \lambda_{ij}^{\hat{X}_{ij}}}{\hat{X}_{ij}!} \right) \\
&= \sum_{ij} \left(\lambda_{ij} + \hat{X}_{ij} \ln \lambda_{ij} \right) \\
&= \sum_{ij} \exp\left(\phi_{ij}^{\beta} - s_i - m_j \right) - \sum_{ij} \hat{X}_{ij} \left(\phi_{ij}^{\beta} - s_i - m_j \right) \\
&= \sum_{ij} \exp\left(\phi_{ij}^{\beta} - s_i - m_j \right) + \sum_i s_i Y_i + \sum_j m_j X_j - \sum_{ij} \hat{X}_{ij} \phi_{ij}^{\beta},
\end{aligned}$$

which is exactly the objective function to be minimized in Eq. (2.14).

2.6 Discussion

In this chapter, we have extended the optimal transport methods introduced in the previous chapter to the estimation of the gravity equation in bilateral trades. We note that our optimal transport approach does not make any distributional assumptions of the trade flow, yet it coincides with the log-likelihood under a Poisson GLM. Theoretically, this justifies the choice of using GLM with the exponential link function in the estimation of the gravity equation. Computationally, this bridges the gap between recent computational investigations of optimal transport methods and the statistics literature.

Chapter 3

Binary inverse problems

We consider the class of binary inverse problems, in which an observed signal is formed as a superposition of a subset of template signals drawn from a dictionary, and corrupted by additive noise. For example, the time-varying voltage measured with an extracellular electrode may be described as a binary superposition of spike waveforms. We assume a Bernoulli prior for the binary coefficients specifying the subset, with known mean taking any value between zero and one (and thus not necessarily sparse). The inference task of estimating the binary coefficients that best explain the observed signal is an N-dimensional discrete optimization problem, with 2^N possible solutions, and thus NP-hard. We formulate an approximate MAP solution by (1) continuous relaxation of the problem, to a linear superposition of dictionary elements with continuous real-valued coefficients lying in the unit interval; (2) inclusion of a concave binarization term in the objective function; and (3) development of a schedule for automatic incremental increases of the multiplier on the concavity term, following a fixed sequence of values that lie between the eigenvalues of the quadratic expansion of the objective. We show

through simulations that this solution, which does not require selection of any regularization or other parameters, achieves the MAP solution for low dimensions ($N \leqslant 4$), and for high dimensions outperforms standard methods developed for sparse inverse problems, including simple thresholding, matching pursuit, LASSO, and iteratively-reweighted L1.

3.1 Introduction

Decomposing an observed signal as a sum of elementary template signals is a classic problem in statistical estimation, dating back more than 100 years. Recent methods have focused on sparse versions of these problems, in which only a small number of templates appear with non-zero coefficient amplitude in any given observed signal. Despite the tremendous success of these methods, they often rely on tuning parameters which can be difficult to choose in cases where there is no ground truth data. In addition, their asymmetric construction may not be appropriate for applications in which coefficients are binary (but not necessarily sparse). For example, in the neural spike sorting problem, one must decompose a time-varying voltage measurement into a sum of spike waveforms, whose amplitude is (approximately) constant.

Here, we develop a method for solving the binary inverse problem. Given a signal formed from a sum of a subset of basis functions drawn from a known dictionary of N such functions, and contaminated by additive Gaussian noise, we wish to recover the binary N-vector indicating which functions are present. Searching over the space of 2^N possible discrete solutions is not feasible for large N, so we construct a continuous relaxation of the problem, and gradually introduce a con-

cave regularization term using a pre-specified schedule, so as to force the solution toward a binary outcome. We show that this method achieves the statistically optimal (MAP) solution for low dimensions ($N \leqslant 4$), and outperforms several standard sparse inverse methods, even for sparse cases (i.e., when the prior probability of zero-valued coefficients is high).

3.2 Problem formulation

Consider the following generative model:

$$\mathbf{s} = \mathbf{D}\mathbf{b} + \mathbf{n}, \tag{3.1}$$

where $\mathbf{s} \in \mathbb{R}^M$ is the observed signal, $\mathbf{D}$ is an M by N matrix consisting of the basis functions (assumed to be known), and $\mathbf{b}$ is a binary coefficient vector in $\mathbb{R}^N$. The noise vector, $\mathbf{n}$, is assumed to be drawn from a standard normal $\mathcal{N}(0, \sigma^2\mathbf{I})$. If in addition, we assume a Bernoulli prior probability of the binary coefficients, with a mean (sparsity) level $p = \frac{1}{N}\mathbb{E}[\mathbf{1}^\top\mathbf{b}]$, where $\mathbf{1}$ is an N-vector of ones, then the posterior probability is

$$\begin{aligned} p(\mathbf{b}|\mathbf{s}) &\propto p(\mathbf{s}|\mathbf{b})p(\mathbf{b}) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^N} e^{-\|\mathbf{D}\mathbf{b}-\mathbf{s}\|^2/2\sigma^2}\ p^{\mathbf{1}^\top\mathbf{b}}(1-p)^{\mathbf{1}^\top(\mathbf{1}-\mathbf{b})}. \end{aligned}$$

Taking the negative log, multiplying by $2\sigma^2$, and discarding constant terms, we can express the *maximum a posteriori* (MAP) estimator as the following optimiza-

tion problem:

$$\hat{\mathbf{b}}_{\mathrm{MAP}}(\mathbf{s}) = \arg\min_{\mathbf{b}} \left\{ \|\mathbf{D}\mathbf{b} - \mathbf{s}\|^2 + 2\sigma^2 \log\left(\frac{1-p}{p}\right) \mathbf{1}^\top \mathbf{b} \right\} \quad \text{s.t.} \quad \mathbf{b} \in \{0,1\}^N. \tag{3.2}$$

3.2.1 Continuous relaxation and a concave term

The discrete optimization problem in (3.2) cannot generally be solved directly as there are 2^N possible solutions. Therefore, we relax the objective function into a continuous form, replacing binary coefficients $\mathbf{b}$ with a real-valued coefficient vector $\mathbf{c}$, constrained to lie within the unit interval:

$$f(\mathbf{c}) = \left\{ \|\mathbf{D}\mathbf{c} - \mathbf{s}\|^2 + 2\sigma^2 \log\left(\frac{1-p}{p}\right) \mathbf{1}^\top \mathbf{c} \right\} \quad \text{s.t.} \quad \mathbf{c} \in [0,1]^N. \tag{3.3}$$

In order to obtain binary solution from (3.3), we introduce a quadratic concave term, controlled by a multiplier λ, into the objective function:

$$f_\lambda(\mathbf{c}) = \left\{ \|\mathbf{D}\mathbf{c} - \mathbf{s}\|^2 + 2\sigma^2 \log\left(\frac{1-p}{p}\right) \mathbf{1}^\top \mathbf{c} - \lambda \|\mathbf{c} - \frac{1}{2}\mathbf{1}\|^2 \right\} \quad \text{s.t.} \quad \mathbf{c} \in [0,1]^N. \tag{3.4}$$

Unlike l_1 regularization, which is used to encourage sparse solutions by pushing coefficients toward zero [29, 7], this concave term is designed to push coefficients toward *either* zero or one (i.e., toward the corners of the unit hypercube). When $\lambda = 0$, we have the original objective function of Eq. (3.3). When $\lambda > 0$, the objective function is a sum of convex and concave functions. And when λ is greater than the largest eigenvalue of $\mathbf{D}^\top\mathbf{D}$, $f_\lambda(\mathbf{c})$ is fully concave, and the solution $\hat{\mathbf{c}}$ will be binary due to the unit interval constraint $\mathbf{c} \in [0,1]^N$.

3.2.2 One dimensional case: an analysis

To see the effect of the concave term in f_λ, we first consider the one-dimensional problem. In this case, Eq. (3.1) becomes

$$\mathbf{s} = b\mathbf{d} + \mathbf{n},$$

where $\mathbf{d}$ is a vector consisting of only one basis function. Eq. (3.2) reduces to

$$\hat{b}_{\text{MAP}}(\mathbf{s}) = \arg\min_b \left\{ b^2 - 2b\tilde{s} + 2\sigma^2 \log\left(\frac{1-p}{p}\right) b \right\} \quad \text{s.t.} \quad b \in \{0, 1\},$$

where we assume that $\mathbf{d}$ is a unit vector ($\mathbf{d}^\top\mathbf{d} = 1$) and define $\tilde{s} = \mathbf{d}^\top\mathbf{s}$, the projection of the signal onto the vector basis function.

Computing the objective at the two possible solutions and setting them to be equal give

$$\tilde{s}^* = \sigma^2 \log\left(\frac{1-p}{p}\right) + 0.5. \tag{3.5}$$

Thus, the MAP estimator corresponds to applying a threshold to the measurement $\tilde{s}$, at the location $\tilde{s}^*$ where these two functions cross, as illustrated in Fig. 3.1. In other words, $\hat{b}_{\text{MAP}} = 1$ when $\tilde{s} > \tilde{s}^*$, and 0 otherwise.

We can show that the continuous objective function in Eq. (3.4) gives exactly the same threshold rule, when the objective function is fully concave (i.e., when $\lambda > \mathbf{d}^\top\mathbf{d} = 1$). In this case,

$$\begin{aligned} f_\lambda(c) &= \|c\mathbf{d} - \mathbf{s}\|^2 + 2\sigma^2 \log\left(\frac{1-p}{p}\right) c - \lambda(c - \frac{1}{2})^2 \\ &= (1-\lambda)c^2 - (2\tilde{s} - 2\sigma^2 \log\left(\frac{1-p}{p}\right) - \lambda)c. \end{aligned}$$

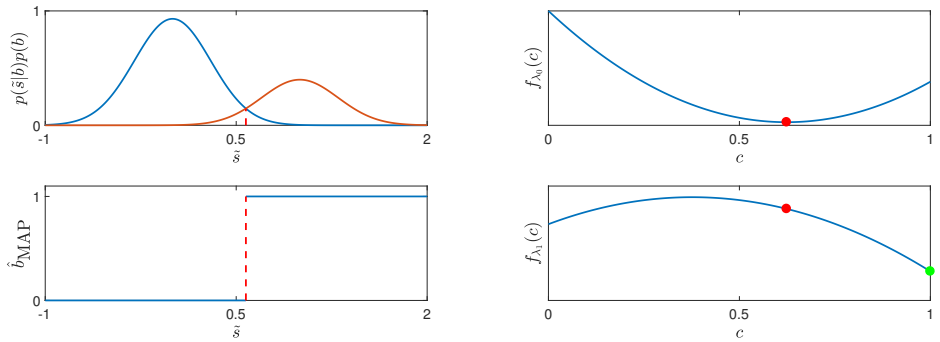

Figure 3.1: Illustration of the one-dimensional binary inverse problem. **Top left**: posterior distributions corresponding to the two options, $b \in \{0, 1\}$. **Bottom left**: The MAP estimator is computed by applying a threshold function to the measured value $\tilde{s}$, with a location specified by the point where the two distributions cross. **Right**: Iterated concavity solution. Top plot indicates the objective function $f_{\lambda_0}(c)$ over the unit interval, for a particular (arbitrarily chosen) measurement $\tilde{s}$. Red point indicates $\hat{c}_0(\tilde{s})$, the minimum. Bottom plot shows the objective $f_{\lambda_1}(c)$ for a value of $\lambda_1 = 2$, for which the function becomes concave. If one minimizes this function using constrained gradient descent, starting from $\hat{c}_0(\tilde{s})$ (red point), the descent will terminate at $\hat{c}_1(\tilde{s}) = 1$ (green point), which is equal to the MAP solution.

That is, $\hat{c}_\lambda(\tilde{s})$ is the same threshold solution as the MAP estimator in Eq. (3.5). Note that the threshold does not depend on λ, as long as $\lambda > 1$.

3.3 Iterative algorithm

The above analysis is only valid in the 1-dimensional case, and a threshold rule generally does not give the MAP solutions in multiple dimensions. In this section we propose an iterative algorithm for solving (3.2).

We denote the eigenvalues of $\mathbf{D}^\top\mathbf{D}$ as $e_1 \leqslant e_2 \leqslant \cdots \leqslant e_N$ and define a sequence of λ's to be $\lambda_0 = 0$, $\lambda_n = e_n$ $(0 < n < N)$, and $\lambda_N > e_N$. We start by solving the constrained *convex* objective function $f_{\lambda_0}(\mathbf{c})$ and denote the solution by $\hat{\mathbf{c}}_0$. We then iterate, minimizing $f_{\lambda_{n+1}}(\mathbf{c})$ starting from $\hat{\mathbf{c}}_n$ for each n. Each objective

function $f_{\lambda_n}(\mathbf{c})$ is the sum of a convex and a concave function, which we solve iteratively with sequential convex programming. Specifically, we initialize $\hat{\mathbf{c}}_n^{(0)} = \hat{\mathbf{c}}_{n-1}$. Then we iteratively perform a Taylor expansion around current estimate $\hat{\mathbf{c}}_n^{(k)}$:

$$f_{\lambda_n}(\mathbf{c}) = f_{\lambda_n}(\hat{\mathbf{c}}_n^{(k)}) + \left(\nabla f_{\lambda_n}(\hat{\mathbf{c}}_n^{(k)})\right)^\top \left(\mathbf{c} - \hat{\mathbf{c}}_n^{(k)}\right) + \frac{1}{2}\left(\mathbf{c} - \hat{\mathbf{c}}_n^{(k)}\right)^\top \nabla^2 f_{\lambda_n}(\hat{\mathbf{c}}_n^{(k)}) \left(\mathbf{c} - \hat{\mathbf{c}}_n^{(k)}\right).$$

Since $\frac{1}{2}\nabla^2 f_{\lambda_n}(\hat{\mathbf{c}}_n^{(k)}) = \mathbf{D}^\top\mathbf{D} - \lambda_n\mathbf{I}$ is indefinite, we keep only its positive semidefinite part and solve the following convex optimization problem to get the next iterate $\hat{\mathbf{c}}_n^{(k+1)}$:

$$\hat{\mathbf{c}}_n^{(k+1)} = \arg\min_{\mathbf{c}\in[0,1]^N} \left\{ \left(\nabla f_{\lambda_n}(\hat{\mathbf{c}}_n^{(k)})\right)^\top \mathbf{c} + \left(\mathbf{c} - \hat{\mathbf{c}}_n^{(k)}\right)^\top \left(\mathbf{D}^\top\mathbf{D} - \lambda_n\mathbf{I}\right)_+ \left(\mathbf{c} - \hat{\mathbf{c}}_n^{(k)}\right) \right\}. \tag{3.6}$$

This "majorization-minimization" approach has been referred to in various literatures as the concave-convex procedure, the difference of convex functions algorithm, and multi-stage convex relaxations. For convergence results, see e.g. [21, 1].

Once the sequence $\{\hat{\mathbf{c}}_n^{(k)}\}_{k=0,1,\ldots}$ converges, we obtain an approximation to the minimizer of f_{λ_n} and denote it as $\hat{\mathbf{c}}_n$. We then minimize $f_{\lambda_{n+1}}$ starting at $\hat{\mathbf{c}}_n$. We repeat the procedure until we reach f_{λ_N}, which is fully concave, and the final solution $\hat{\mathbf{c}}_N$ obtained by our algorithm will always be binary.

3.3.1 One dimensional case, revisited

We analyze the iterative algorithm in one dimension. When $\lambda_0 = 0$, we are solving the regular constrained least squares problem

$$\begin{aligned}&\arg\min_{c\in[0,1]}\left\{\|c\mathbf{d}-\mathbf{s}\|^2+2\sigma^2\log\left(\frac{1-p}{p}\right)c\right\}\\=&\arg\min_{c\in[0,1]}\left\{c^2+(2\sigma^2\log\left(\frac{1-p}{p}\right)-2\tilde{s})c\right\}.\end{aligned}$$

If $\tilde{s}-\sigma^2\log\left(\frac{1-p}{p}\right)\in[0,1]$, the solution is $\hat{c}_0=\tilde{s}-\sigma^2\log\left(\frac{1-p}{p}\right)$.

We then minimize $f_{\lambda_1}(c)$ where $\lambda_1 > 1$ using Eq. (3.6). Since the objective is fully concave, the quadratic term disappears and the minimization problem becomes

$$\hat{c}_1^{(1)}=\arg\min_{c\in[0,1]}\left\{\nabla f_{\lambda_1}(\hat{c}_0)c\right\}. \tag{3.7}$$

Since

$$\begin{aligned}\nabla f_{\lambda_1}(\hat{c}_0)&=2\hat{c}_0+2\sigma^2\log\left(\frac{1-p}{p}\right)-2\tilde{s}-2\lambda_1(\hat{c}_0-\frac{1}{2})\\&=-2\lambda_1(\hat{c}_0-\frac{1}{2})\end{aligned}$$

and (3.7) is a linear program with box constraints, we know if $\hat{c}_0 > \frac{1}{2}$ we will arrive at $\hat{c}_1^{(1)}=1$ and vice versa. Therefore, $\hat{c}_1^{(1)}$ is obtained by thresholding $\tilde{s}$ at $0.5+\sigma^2\log\left(\frac{1-p}{p}\right)$, identical to the MAP solution of Eq. (3.5).

We now consider the next iterate

$$\hat{c}_1^{(2)}=\arg\min_{c\in[0,1]}\left\{\nabla f_{\lambda_1}(\hat{c}_1^{(1)})c\right\},$$

where

$$\nabla f_{\lambda_1}(\hat{c}_1^{(1)}) = 2\hat{c}_1^{(1)} + 2\sigma^2 \log\left(\frac{1-p}{p}\right) - 2\tilde{s} - 2\lambda_1(\hat{c}_1^{(1)} - \frac{1}{2})$$
$$= 2(1-\lambda_1)\hat{c}_1^{(1)} + 2\sigma^2 \log\left(\frac{1-p}{p}\right) - 2\tilde{s} + \lambda_1.$$

We recall $\hat{c}_1^{(1)} \in \{0, 1\}$. If $\hat{c}_1^{(1)} = 1$, we know $\tilde{s} - \sigma^2 \log\left(\frac{1-p}{p}\right) = \hat{c}_0 > \frac{1}{2}$, so

$$\nabla f_{\lambda_1}(\hat{c}_1^{(1)}) < 2(1-\lambda_1)\hat{c}_1^{(1)} - (1-\lambda_1) = 1 - \lambda_1 < 0.$$

Thus, we again have $\hat{c}_1^{(2)} = 1$, and by induction all the subsequent iterates $\hat{c}_1^{(k)}$ will remain at 1. A similar analysis can be carried out for cases where $\hat{c}_1^{(1)} = 0$ and $\tilde{s} - \sigma^2 \log\left(\frac{1-p}{p}\right) < 0$ or $\tilde{s} - \sigma^2 \log\left(\frac{1-p}{p}\right) > 1$. Thus, in one dimension, our iterative algorithm gives exactly the same solution as the MAP solution of Eq. (3.5).

3.4 Two-dimensional case

In the previous section we proved that by sequentially minimizing the convex objective function $f_{\lambda_0}(\mathbf{c})$, and then the fully concave function $f_{\lambda_1}(\mathbf{c})$, one arrives at the true MAP solution. The second step corresponds to applying a threshold to the solution from the first step, $\hat{c}_0(\mathbf{s})$. In two dimensions, when the dictionary elements do not overlap (i.e., when they have disjoint supports), the matrix $\mathbf{D}^\top\mathbf{D}$ will be diagonal, and the objective function can be expressed as a sum of N independent scalar functions, each operating on a component of $\mathbf{c}$. Thus, the two-dimensional problem can be essentially treated as two independent one-dimensional problems. In this separable case, the MAP solution corresponds to applying the same thresholding rule to each component independently, and the two-step algorithm from the

previous section will again achieve the MAP solution.

However, when the two dictionary elements are correlated, with inner product ρ, the MAP decision boundaries become non-separable, as shown in the upper right panel of Fig. 3.2, for a prior (sparsity level) of $p = 0.5$. Specifically, we see that a boundary segment connecting the black and white regions, oriented at 45 degrees, is formed. The bottom right panel shows the MAP decision boundaries for $p = 0.45$, which are identical in shape, but shifted in the $(1, 1)$ direction.

Under these conditions, the two-step (thresholding) procedure (solving f_{λ_0} and then f_{λ_N}) results in non-separable decision boundaries that differ from the MAP solution (Fig. 3.2, upper middle panel). But we find that the full iterative concavity procedure (solving f_{λ_0}, f_{λ_1}, and f_{λ_2}) does achieve the MAP solution (Fig. 3.2, upper right panel). This also holds for different values of noise level, σ, and prior, p (Fig. 3.2, lower right panel). Moreover, we have verified through numerical simulation that the iterative procedure attains the true MAP solution for correlated dictionaries in three and four dimensions, whereas the two-step thresholding procedure, is suboptimal, producing errors near the non-horizontal and non-vertical segments of the MAP decision boundaries. This result cannot be generalized to the high-dimensional case, since the MAP solution cannot be computed. Instead, in the next section, we compare performance on high-dimensional simulated data with several widely used methods for sparse inverse problems.

3.5 High-dimensional simulations

In this section we compare our approach against four methods for solving binary inverse problems: simple thresholding, matching pursuit (a greedy method),

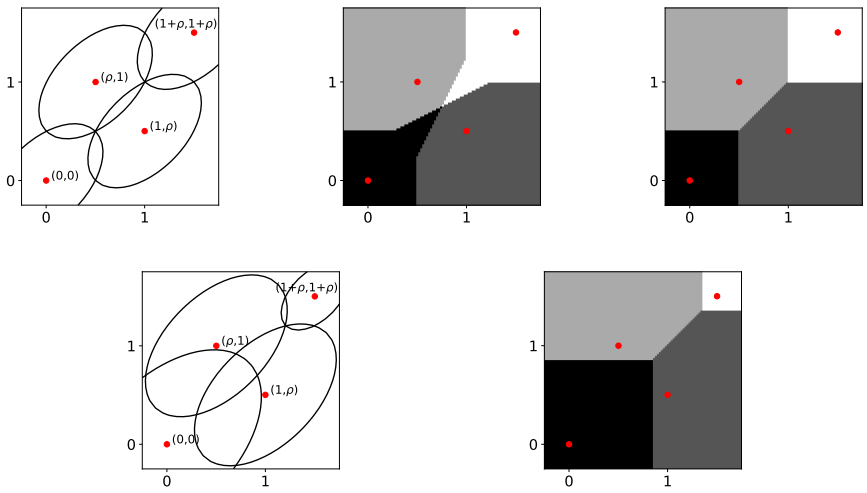

Figure 3.2: Illustrations of the two-dimensional binary inverse problem. Axes of all plots correspond to the components of $\mathbf{D}^\top\mathbf{s}$. **Top Left**: Posterior distribution for prior $p = 0.5$. Elliptical level curves of the Gaussian distributions corresponding to the four choices of $\mathbf{b}$, indicated by the four red points. **Top Middle**: The decision boundaries obtained by the two-step (thresholding) solution, solving f_{λ_0} and then f_{λ_2}. **Top Right**: The decision boundaries attained by the iterative concavity algorithm (i.e., by sequentially solving f_{λ_0}, f_{λ_1} and then f_{λ_2}), which are identical to the true MAP solution. **Bottom**: The same contour and MAP decision boundary plot, for a prior value $p = 0.45$. This solution is also attained by the iterative algorithm.

LASSO (l_1 regularization), and iteratively reweighed l_1 (IRL1). We performed tests on translational dictionaries containing two types of basis function, a univariate Gaussian, $g_1(x) \propto \exp(-x^2)$, and the derivative of a Gaussian, $g_2(x) \propto x \exp(-x^2)$. We consider two different prior values: $p = 0.5$ (non-sparse) and $p = 0.2$ (sparse). We examine performance under different noise levels (indicated by signal-to-noise ratio, SNR $= 1/\sigma$), as well as different spacing between basis functions, δ, expressed in units of the standard deviation used for g_1 and g_2. We chose $N = 32$ for all simulations described in this section. For these simulations, we find that the running time of our Iterative Concavity implementation is comparable to that of IRL1.

The greedy method (e.g. [22, 32]) starts with an all-zero solution vector, and iteratively flips one bit, chosen so as to most reduce the likelihood function (mean squared error). This procedure is repeated until the improvement in the mean squared error is sufficiently small.

The LASSO [29] solves an l_1-regularized version of the original objective function:

$$\hat{\mathbf{c}}_{\text{LASSO}} = \arg \min_{\mathbf{c} \in [0,1]^N} \left\{ \|\mathbf{D}\mathbf{c} - \mathbf{s}\|^2 + \lambda \|\mathbf{c}\|_1 \right\}. \tag{3.8}$$

Inclusion of the l_1 penalty encourages sparsity (in particular, it favors zero coefficients), while still retaining convexity.

The IRL1 method [6] uses a concave-convex procedure to solve the following optimization problem:

$$\hat{\mathbf{c}}_{\text{IRL1}} = \arg \min_{\mathbf{c} \in [0,1]^N} \left\{ \|\mathbf{D}\mathbf{c} - \mathbf{s}\|^2 + \sum_{i=1}^{N} \log(\varepsilon + |c_i|) \right\}.$$

It takes the LASSO solution as an initialization, i.e. $\hat{\mathbf{c}}^{(0)}_{\text{IRL1}} = \hat{\mathbf{c}}_{\text{LASSO}}$. At

each step k, a diagonal weight matrix $\mathbf{W}^{(k)}$ is constructed with $\text{diag}(\mathbf{W}^{(k)}) = 1/(\varepsilon + \hat{\mathbf{c}}_{\text{IRL1}}^{(k)})$, and the next iterate $\hat{\mathbf{c}}_{\text{IRL1}}^{(k+1)}$ is obtained via the following optimization problem

$$\hat{\mathbf{c}}_{\text{IRL1}}^{(k+1)} = \arg \min_{\mathbf{c} \in [0,1]^N} \left\{ \|\mathbf{D}\mathbf{c} - \mathbf{s}\|^2 + \lambda \|\mathbf{W}^{(k)}\mathbf{c}\|_1 \right\}. \tag{3.9}$$

For LASSO and IRL1, both the regularization parameter λ, and a threshold T to force the final solutions to be binary, need to be selected. In our simulations, we test values $\lambda \in \{0, 10^{-3}, 3 \cdot 10^{-3}, 0.01, , \cdots, 1, 3\}$ and $T \in \{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}, 0.01, 0.03, 0.1, 0.3, 0.5, 0.7, 0.9\}$.

3.5.1 Non-sparse case

With Bernoulli prior $p = 0.5$, we observe that both LASSO and IRL1 (which are designed for sparse problems) achieve their best performance at $\lambda = 0, T = 0.5$, which is essentially the same as solving $f_{\lambda_0}(\mathbf{c})$ to obtain $\hat{\mathbf{c}}_0$ and thresholding all components at 0.5. Therefore, for this case, we compare the iterative concavity algorithm against the greedy method and the thresholding solution, for a wide range of SNR and spacing values. The results are shown in Fig. 3.3. We compute the total error as the average of the sum of false alarms (identifying 0 as 1) and misses (identifying 1 and 0), averaged over 1000 randomly drawn examples.

We see that for both types of basis function, the greedy method performs poorly. When SNR is high and the separation between basis functions is large (i.e. the lower left region in the plots of Fig. 3.3), the greedy method performs as well as the other two. However, once the separation becomes sufficiently small, the performance of the greedy method drops rapidly. On the other hand, both the thresholding and our solutions are relatively robust to noise and basis function

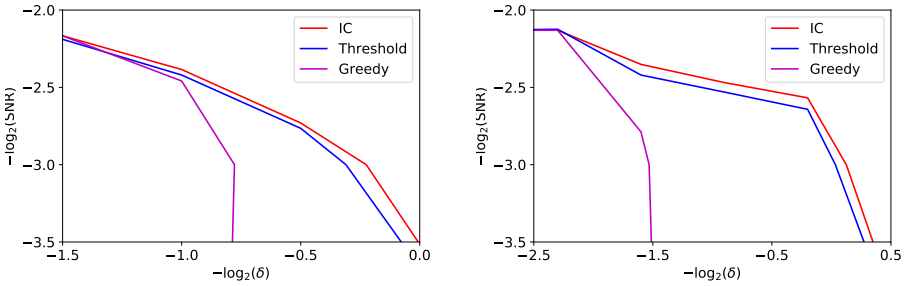

Figure 3.3: Performance comparison of the Iterative Concavity, Thresholding, and Greedy algorithms, as a function of noise level (SNR) and spacing (δ). Average performance, measured as the sum of misses and false alarms, was estimated by averaging over 1000 simulated trials. Level curves for each method are shown at 10% of the maximum error. **Left**: Dictionary of 32 Gaussian basis function. **Right**: Dictionary of 32 Gaussian derivative basis functions.

correlation, with ours outperforming thresholding at all spacing values.

3.5.2 Sparse case

We compared all five methods for a sparse Bernoulli prior of $p = 0.2$. The panels of Fig. 3.4 show the average number of misses and false alarms for all five methods, averaged over 1000 randomly drawn examples, for different combinations of SNR and spacing (δ), and for both the Gaussian and Gaussian derivative basis functions. The convex hulls for the LASSO and IRL1 solution points, obtained for a set of different tuning parameters (λ, T) are shown. For all (SNR, δ) pairs, the Iterative Concavity method always achieves the same or slightly better performance than the best achievable by LASSO and IRL1 (i.e., with the right values of λ and T), and is significantly better than the greedy and the thresholding solutions.

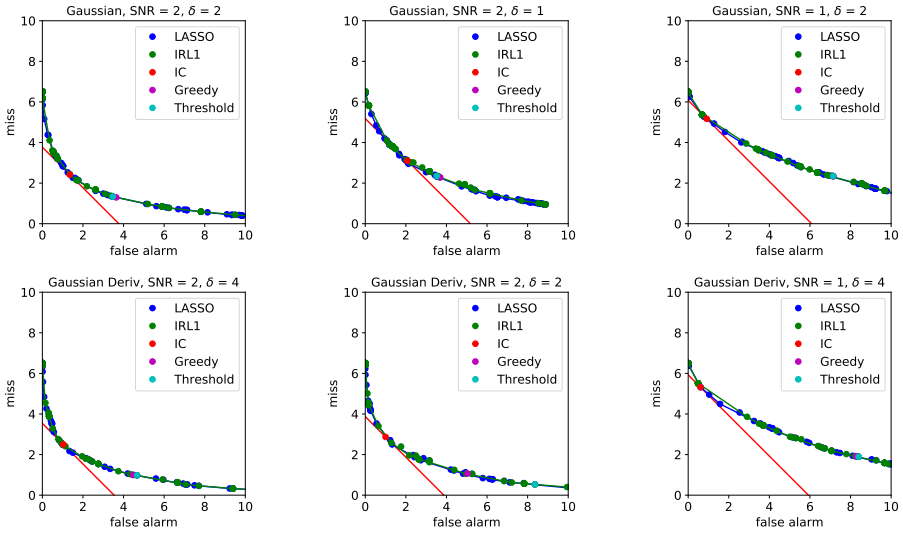

Figure 3.4: Comparison of the Iterative Concavity, LASSO, IRL1, Threshold, and Greedy algorithms. Each panel shows misses and false alarms in recovering a 32-dimensional binary coefficient vector, averaged over 1000 randomly drawn examples, for the indicated type of basis function, SNR and spacing (δ). The multiple points shown for LASSO and IRL1 correspond to different choices of their tuning parameters, (λ, T), and the green line indicates the convex hull of these points. The red line indicates the total errors (false alarms plus misses) for the Iterative Concavity solution.

Table 3.1: Optimal (λ, T) for LASSO and IRL1.

Parameters		Optimal (λ, threshold)	
Basis Function Type	(SNR, δ)	LASSO	IRL1
Gaussian	$(2, 2)$	(0.3, 0.5)	(0.1, 0.7)
Gaussian	$(2, 1)$	(0.1, 0.7)	(0, 0.9)
Gaussian	$(1, 2)$	(1, 0.9)	(1, 0.9)
Gaussian Derivative	$(2, 4)$	(0, 0.9)	(0, 0.9)
Gaussian Derivative	$(2, 2)$	(0.3, 0.5)	(0.1, 0.7)
Gaussian Derivative	$(1, 4)$	(1, 0.9)	(1, 0.9)

Table 3.1 indicates the (λ, T) pair that achieves the lowest error in each of the plots shown in Fig. 3.4. We see that the optimal combination for both LASSO and IRL1 vary significantly, and the threshold values are often quite large (e.g., 0.7-0.9). This is a significant disadvantage, since in applications such as spike sorting where no ground truth is available, it is not obvious how to select them.

3.6 Image dithering

In this section we apply out iterative concavity procedure to the task of image dithering. Given an original gray-scale image $\mathbf{I}_{\text{orig}}$ as a 2D matrix of pixel values between 0 and 1, we want to find a binarized image $\mathbf{I}_{\text{b}}$, whose pixel values can only be either 0 or 1, such that a pre-specified perceptual metric, $\|\cdot\|_{\text{perc}}$, is minimized. Mathematically, the problem can be formulated as

$$\min_{\mathbf{I}_{\text{b}}} \|\mathbf{I}_{\text{b}} - \mathbf{I}_{\text{orig}}\|_{\text{perc}}. \tag{3.10}$$

One most commonly used metric is the Mean Squared Error (MSE), which will simply yield the pixel-wise thresholded solution. Although there are standard

perceptual models developed in the human vision literature (see e.g. [35]), we will focus on the performance of our optimization procedure and adopt the weighted Fourier metric. More specifically, we let $\mathcal{F}$ denote the Fourier transform. We also denote $\mathbf{F}_{\text{orig}} = \mathcal{F}\mathbf{I}_{\text{orig}}$ and $\mathbf{F}_{\text{b}} = \mathcal{F}\mathbf{I}_{\text{b}}$ as the transformed images in the Fourier domain. The weighted Fourier metric is then defined as

$$\|\mathbf{I}_{\text{b}} - \mathbf{I}_{\text{orig}}\|_{\text{perc}} = (\text{vec}(\mathbf{F}_{\text{b}} - \mathbf{F}_{\text{orig}}))^{*}\mathbf{W}(\text{vec}(\mathbf{F}_{\text{b}} - \mathbf{F}_{\text{orig}})), \tag{3.11}$$

where $\mathbf{W}$ is a diagonal matrix with weights on the diagonal.

Following the approach developed in previous sections, we perform a continuous relaxation and introduce a concave term into the objective function, so Eq. (3.10) becomes

$$\min_{\mathbf{I}_{\text{c}}} \left\{ \|\mathbf{I}_{\text{c}} - \mathbf{I}_{\text{orig}}\|_{\text{perc}} - \lambda \|\mathbf{I}_{\text{c}} - \frac{1}{2}\mathbf{1}\mathbf{1}^{\top}\|^2 \right\}, \tag{3.12}$$

where the pixel values of $\mathbf{I}_{\text{c}}$ are allowed to lie within the interval $[0, 1]$. We note that the eigenvalues of the Hessian of $\|\mathbf{I}_{\text{c}} - \mathbf{I}_{\text{orig}}\|_{\text{perc}}$ are simply the diagonal entries of $\mathbf{W}$, so we can set our λ's to be such. We now compare the performance of our approach against the standard Floyd-Steinberg dithering procedure, as well as a greedy approach where one goes through the image from left to right, and then from top to bottom, each time greedily binarizing one pixel by minimizing the perceptual metric defined in (3.11). We apply these methods to an image of size 512 by 512. The weights in $\mathbf{W}$ of Eq. 3.11 are set by scaling $|f|^{0.25}$ such that the largest value is 1 and rounding to the nearest tenth. The results are presented in Fig. 3.5.

Comparing the bottom row against the output of the Floyd-Steinberg algorithm (the top right figure), we see both have sharper contrast. Indeed, this shows the

Figure 3.5: Comparison of the Iterative Concavity, Floyd-Steinberg, and Greedy algorithms. **Top Left**: The original gray-scale image. **Top Right**: The output of Floyd-Steinberg dithering, with a perceptual difference of 11007. **Bottom Left**: The output of Iterative Concavity, with a perceptual difference of 9977. **Bottom Right**: The output of Greedy, with a perceptual difference of 10599.

weighted Fourier metric is perceptually superior to MSE. The Floyd-Steinberg has a perceptual error of 11007, which is about 5 % higher than that achieved by the Greedy method, and is about 10 % higher than our Iterative Concavity method.

The bottom row shows the output of our Iterative Concavity method (bottom left) and the Greedy method (bottom right). Although both figures display significantly higher contrast when compared with the original image and the Floyd-Steinberg output, the output of our Iterative Concavity method retains much more detail. For instance, the details of the mesh next to the models right cheek can still be clearly seen, while in the right figure they melt into a block of dark pixels; the wisp of hair in front of the models left eye, visibly captured by our algorithm, is also lost by the Greedy algorithm. In addition, the color on the models face varies with more consistency in the left figure, whereas in the right figure the transitioning between different facial areas seems more rigid, especially around the nose.

3.7 Discussion

We have presented an Iterative Concavity algorithm for solving binary inverse problems, assuming a Bernoulli (but not necessarily sparse) coefficient prior. We show analytically in one dimension and numerically in two dimensions, that our method yields the optimal MAP decision boundaries. We also examined performance on simulated high-dimensional examples, comparing to a variety of sparse inverse methods. We find that Iterative Concavity substantially outperforms simple thresholding and greedy methods, and achieves or exceeds the best-case performance of LASSO and IRL1, taken over all possible values of their tuning parameters. This holds for different choices of SNR and basis function spacing, even

when the Bernoulli prior is low (i.e., sparse cases). Given that there is no generally agreed-upon method of selecting these parameters for problems lacking ground truth data, these results imply a significant advantage of our method.

We believe this approach can be generalized in a number of ways. The noise model should be generalizable to any additive noise with a log-convex distribution. We've only shown examples with convolutional dictionaries (i.e., formed from shifted copies of a single basis function), but the problem and solution formulation do not rely on this, and should apply equally well to more general dictionaries. Finally, if the basis functions are unknown, one could adopt a coordinate descent algorithm, such as in [24, 13], to iteratively alternate between optimizing the basis functions and the associated binary coefficients.

Bibliography

[1] Le Thi Hoai An and Pham Dinh Tao. The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems. *Annals of operations research*, 133(1):23–46, 2005.

[2] James E Anderson and Eric Van Wincoop. Gravity with gravitas: a solution to the border puzzle. *American economic review*, 93(1):170–192, 2003.

[3] Gary S Becker. A theory of marriage: Part i. *Journal of Political economy*, 81(4):813–846, 1973.

[4] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[5] Rainer E Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment problems, revised reprint*. SIAM, 2009.

[6] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier analysis and applications*, 14(5):877–905, 2008.

[7] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[8] Pierre-André Chiappori, Sonia Oreffice, and Climent Quintana-Domeque. Fatter attraction: anthropometric and socioeconomic matching on the marriage market. *Journal of Political economy*, 120(4):659–695, 2012.

[9] Eugene Choo and Aloysius Siow. Who marries whom and why. *Journal of political economy*, 114(1):175–201, 2006.

[10] Eustasio del Barrio, Evarist Giné, and Carlos Matrán. Central limit theorems for the wasserstein distance between the empirical and the true distributions. *Annals of Probability*, pages 1009–1071, 1999.

[11] Bas Donkers and Arthur Van Soest. Subjective measures of household preferences and financial decisions. *Journal of Economic Psychology*, 20(6):613–642, 1999.

[12] Arnaud Dupuy and Alfred Galichon. Personality traits and the marriage market. *Journal of Political Economy*, 122(6):1271–1319, 2014.

[13] Chaitanya Ekanadham, Daniel Tranchina, and Eero P Simoncelli. A unified framework and method for automatic neural spike identification. *Journal of neuroscience methods*, 222:47–55, 2014.

[14] Maryam Fazel. Matrix rank minimization with applications. *Doctoral dissertation, Stanford University*, 2002.

[15] Jeremy T Fox. Identification in matching games. *Quantitative Economics*, 1(2):203–254, 2010.

[16] Alfred Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

[17] Alfred Galichon and Arnaud Dupuy. Canonical correlation and assortative matching: A remark. *Annals of Economics and Statistics*, 119/120:375–383, 2015.

[18] Alfred Galichon and Bernard Salanié. Cupid's invisible hand: Social surplus and identification in matching models. 2012.

[19] Pieter A Gautier, Michael Svarer, and Coen N Teulings. Marriage and the city: Search frictions and sorting of singles. *Journal of Urban Economics*, 67(2):206–218, 2010.

[20] Lisa K Jepsen. The relationship between wife's education and husband's earnings: Evidence from 1960 to 2000. *Review of Economics of the Household*, 3(2):197–214, 2005.

[21] Gert R Lanckriet and Bharath K Sriperumbudur. On the convergence of the concave-convex procedure. In *Advances in neural information processing systems*, pages 1759–1767, 2009.

[22] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[23] Peter McCullagh and John A Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[24] Bruno A Olshausen and David J Field. Emergence of simple-cell recep-

tive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, Jun 1996.

[25] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[26] Lloyd S Shapley and Martin Shubik. The assignment game i: The core. *International Journal of game theory*, 1(1):111–130, 1971.

[27] JMC Santos Silva and Silvana Tenreyro. The log of gravity. *The review of economics and statistics*, 88(4):641–658, 2006.

[28] Marko Terviö. Studies of talent markets. *Doctoral dissertation, Massachusetts Institute of Technology*, 2003.

[29] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[30] Jan Tinbergen. The world economy. suggestions for an international economic policy. *Twentieth Century Fund*, 1962.

[31] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of optimization*, 6(15):615–640, 2010.

[32] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[33] Cédric Villani. *Topics in optimal transportation*. American Mathematical Society, 2003.

[34] Cédric Villani. *Optimal transport: old and new*. Springer, 2008.

[35] Andrew B Watson and Albert J Ahumada. A standard model for foveal detection of spatial contrast. *Journal of vision*, 5(9):717–740, 2005.

[36] G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170:33–45, 1992.

[37] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.