Entropy and Information Recovery in Linear Economic Models

by

Douglas James Miller

B.S. (Iowa State University) 1987
M.S. (Cornell University) 1991
M.A. (University of California at Berkeley) 1994

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Agricultural and Resource Economics

in the

GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor George G. Judge, Chair
Professor Larry S. Karp
Professor Rudolph J. Beran

1994

The dissertation of Douglas James Miller is approved:

George N Judge 15 November 94
Chair Date

Larry A Karp 15 Nov 94
Date

RJ Beran 15 Nov 94
Date

University of California at Berkeley

1994

Entropy and Information Recovery in Linear Economic Models

by

Douglas James Miller

Abstract

Entropy and Information Recovery in Linear Economics Models

by

Douglas James Miller

Doctor of Philosophy in Agricultural and Resource Economics

University of California at Berkeley

Professor George G. Judge, Chair

The purpose of the dissertation research is to examine the the properties and performance of the generalized maximum entropy (GME) and generalized minimum cross-entropy (GCE) methods of information recovery. The GME-GCE framework was devised by Judge and Golan as a feasible means for solving linear inverse problems. Given the limitations of economic data, such problems are frequently ill-posed or ill-conditioned, and the associated solutions are either non-unique or unstable. Using limited prior knowledge (from theory or experience), the unknown system parameters and disturbances of the general linear model are reparameterized in terms of probabilities. The GME-GCE problem is to recover the set of probability distributions on the unknowns that satisfy the observed (sample) information and are 'closest' to the prior information.

Although the GME-GCE solution does not take a closed form, the dual formulation of the problem may be used to compute the solution using unconstrained techniques, and a computer algorithm is presented. By treating the dual approach as an M-estimation problem, the solution to the GME-GCE problem is shown to be consistent and asymptotically normal under modest regularity conditions. In finite samples, the GME-GCE solution also exhibits shrinkage properties, including reduced precision loss (i.e. mean squared error) relative to the traditional estimators.

The performance of the GME-GCE solutions for common economic inverse problems are examined with Monte Carlo sampling experiments. First, a bounded mean is recovered from a single observation, and the GME solution is compared to restricted ML and Bayes methods. Second, an ill-conditioned design matrix is devised, and the

unknown parameters are recovered by LS, RLS, ridge, and GME-GCE. Finally, a linear model subject to an AR(1) error process is used to demonstrate the performance of GME under various error specifications. In general, the GME–GCE solutions risk–dominate the traditional methods, are robust under alternate model specifications, and are able to avoid the ill effects of poor prior information. The generalized entropy framework may be extended to a variety of familiar economic inverse problems, including qualitative choice models, inverse control problems, and Markov chains.

George G. Judge

Dedicated

to

Jen

Contents

List of Figures

List of Tables

Mathematical Notation

$\imath_n$	an $(n \times 1)$ vector of ones
I_n	an $(n \times n)$ identity matrix
J_n	an $(n \times n)$ matrix of ones
$\emptyset$	empty set
$\text{int}(\mathcal{A})$	interior of set $\mathcal{A}$
$\partial(\mathcal{A})$	boundary of set $\mathcal{A}$
$\otimes$	Kronecker product
$\odot$	Hadamard product
$\rightarrow$	point-wise convergence
$\xrightarrow{p}$	convergence in probability (weak)
$\Rightarrow$	convergence in distribution (law)
$O(a_T)$	at most of order a_T
$O_p(a_T)$	at most of stochastic order a_T
$\cosh(\cdot)$	hyperbolic cosine
$\tanh(\cdot)$	hyperbolic tangent

Abbreviations

CLT	Central Limit Theorem
FOC	First Order Conditions
GCE	Generalized Cross Entropy
GLM	General Linear Model
GME	Generalized Maximum Entropy
GMM	Generalized Method of Moments
IV	Instrumental Variables
LS	Least Squares
ME	Jaynes' Maximum Entropy
ML	Maximum Likelihood
MOM	Method of Moments
MOR	Method of Regularization
SLLN	Strong Law of Large Numbers
SOC	Second Order Conditions
WLLN	Weak Law of Large Numbers

Acknowledgements

As stated within, George Judge and Amos Golan are entirely responsible for the GME–GCE framework. The dissertation reports my own original research contributions, but these efforts are one part of a much larger research line initiated by George and Amos. Departing their company is the only real reservation I have about leaving Berkeley for my alma mater. I am very grateful that they have allowed me to join in the fun. I also grateful for the support of a very patient and helpful committee: Rudy Beran, Larry Karp, Jeff Perloff, and Brian Wright. Finally, thanks to the many seminar participants at Berkeley, Davis, Iowa State, Maryland, and Purdue.

Chapter 1

Inverse Problems and Information Recovery in Economics

1.1 Linear Inverse Problems in Economics

One objective of economic research is to predict the behavior of agents, $y \in \mathcal{Y}$, from information about their economic environment, $X \in \mathcal{X}$. Unfortunately, the underlying economic system

$$m: \mathcal{X} \to \mathcal{Y}$$

is rarely known or observable, and the research task becomes a two-stage process. An image of the system, $\hat{m} \in \mathcal{M}$, must be recovered from available information before inferences about choices or actions may be formed. The set of available information, $\mathcal{I}$, typically includes the implications of theory, experience, and other prior knowledge as well as *indirect* measurements of the system (e.g. previous observations of y and X). The two stages of the research problem, prediction and information recovery, are commonly known as the *direct* problem,

(DP) $$\hat{m}: \mathcal{X} \to \mathcal{Y}$$

and the *inverse* problem,

(IP) $$n: \mathcal{I} \to \mathcal{M}$$

respectively.

To illustrate the direct-inverse problem framework, consider the (direct) problem of predicting aggregate retail demand for some good in a future period. Individual and aggregate demand correspondences are rarely known or observed, so inferences about future behavior must be based on some image of the underlying system. Consumer theory provides one basis for analyzing individual and aggregate demands, and a typical system expresses quantity demanded as a function of relevant prices, income, and demographic or other exogenous variables. The information set then includes past observations on these variables as well as the implications of consumer theory (e.g. homogeneity) and previous research on demand for the good. The inverse problem is to recover an image of the aggregate demand system that satisfies the assembled information in some reasonable fashion.

In general, inverse problems with unique solutions are said to be *well-posed*. The resulting image of the economic system, $\hat{m}$, may be used to solve the direct problem

$$(\widehat{\mathrm{DP}}) \qquad \hat{m}\colon X_0 \to \hat{y}$$

where X_0 is the conditioning information and $\hat{y}$ is the associated prediction. If the inverse problem does not have a unique solution, the problem is said to be *ill-posed* (Tikhonov and Arsenin, 1977). Sabatier (1987) discusses the various degrees of ill-posedness (e.g. multiple, set-valued, or infinitely-many solutions), and O'Sullivan (1986) provides a summary of recent research on ill-posed problems. Fundamentally well-posed problems are said to be *ill-conditioned* if there are multiple solutions or if the solution changes sharply given modest shifts in the available information (unstable). Limitations of the indirect observations are the most common source of ill-conditioning, and the potential problems include actual or incidental dependencies among the observations as well as measurement errors.

By assuming additional information or imposing regularity conditions, researchers may be able to devise a refined class of admissible systems, $\mathcal{M}^* \subset \mathcal{M}$, which admits a unique solution. A common approach is to restrict $\mathcal{M}^*$ to some family of finite-dimensional parametric systems

$$(1.1) \qquad y = m(X, \beta)$$

where $\beta \in \mathcal{B} \subset \Re^K$. Among the class of parametric forms, the general linear model

$$(\mathrm{GLM}) \qquad y = X\beta + e$$

is a convenient approximation of the true system. Here, y is a T-vector of indirect observations composed of additive *signal* and *noise* components. The signal, $X\beta$, is a linear combination of K non-stochastic explanatory variables, X, with response weights β. The *noise* term, e, is a T-vector of unobserved disturbances that may represent the random aspects of human behavior as well as approximation, specification, or measurement errors. The problem of recovering information about β is known as a *linear* inverse problem.

If the system is non-stochastic *and* is observed without noise, IP is said to be a *pure* inverse problem. The task of solving pure inverse problems is largely an exercise in applied mathematics. In the case of the pure GLM, $y = X\beta$, the linear inverse problem may be solved as a system of equations. The standard solution is a linear inversion operator, A, such that

$$\widehat{\beta} = A(X)y$$

If X is a square matrix (i.e. $T = K$) with full column rank, $r(X) = K$, $A = X^{-1}$. If $T > K$, the rank of X is still K, and any $K \times K$ partition of X may be inverted to form A. Searle (1982, p. 234) shows that $A = (X'X)^{-1}X'$ is an equivalent approach. Clearly, X^{-1} is not unique and the problem is ill-posed if $T < K$. Further, the problem is ill-conditioned if $T \geq K$ but $r(X) < K$.

Given the numerous sources of noise in the underlying system or the indirect observations, pure inverse problems are rare in practice. In the presence of disturbances, IP is said to be an inverse problem *with noise*. The information set for the linear inverve problem, $\mathcal{I}$, may now include the distribution, bounds, moments, or other properties of the disturbances. For the GLM, $\mathcal{I}$ should include any known properties of e relevant to the task of recovering information about β. The principle methods used to recover information in the noise case are discussed in the next section.

Returning to the demand example, the available information is rarely enough to identify the aggregate demand structure, and the model is often restricted to a finite-dimensional linear model with some theoretical basis (e.g. Rotterdam, AIDS, or linear expenditure system). Given the considerable potential for measurement and specification errors, as well as randomness in consumer choice, the task of recovering the demand system is almost certainly a linear inverse problem with noise. Further, the demand problem may be ill-posed if the number of observations used to recover the demand system is insufficient. Limitations of the observations such as collinearity or measurement errors may result in inadmissible or unstable estimates of β, even if the problem is well-posed. The effects of ill-conditioning may be treated by imposing additional prior information on the unknown parameters. For example, the unknown parameters in log-linear demand systems take the form of price, income, or other

elasticities, which may be bounded or signed *a priori*.

Economic and other data are often aggregated observations of unreplicated, non–experimental events. Further, the indirect observations may be noisy, limited, partial, or incomplete. For example, demand statistics are often based on preliminary surveys subject to sampling errors or revisions (noisy). The observations may represent quarterly or annual (limited) measures of average per capita (partial) disappearance rather than true demand (incomplete). The properties of the demand example are characteristic of, but not peculiar to, economic inverse problems. Consequently, ill–posed or ill–conditioned inverse problems are rather common in economics and other disciplines, and solution techniques designed for well–posed orthogonal experimental designs may be infeasible or inappropriate.

Conversely, economic theory or empirical knowledge may provide information that may be used to regularize the inverse problem. For example, consumer theory may impose restrictions on the signs or magnitudes of unknown demand parameters (e.g. elasticities, flexibilities, and multipliers). The prior information may take the form of subjective probability distributions specified on the relevant parameter space, $\mathcal{B}$. Therefore, many inverse problems in economics may be augmented to form a refined solution space, $\mathcal{M}^*$. The critical aspect of the information recovery task is employing substantive prior knowledge rather than creative assumptions to regularize the inverse problem.

The set of solution methods is very large, and it includes many variations based on the properties of the available sample and prior information as well as the underlying estimation and inferential philosophy. Before introducing the entropy–based methods of information recovery, a brief review of the traditional methods is provided. In this way, the entropy methods may be compared with the existing techniques as the discussion proceeds.

1.2 Traditional Methods of Information Recovery

A variety of estimation and inferential philosophies and associated methods have been devised to recover information from inverse problems with noise. Throughout

much of the following discussion, the noise term, e, is assumed to be a random vector with distribution $F(e)$. Given that X is fixed, the associated density or mass function for y given X and β is given by the Radon–Nikodyn derivative

$$\frac{dF(y;X,\beta)}{dv} = f(y;X,\beta)$$

with respect to some measure, v (e.g. counting or Lebesgue). For convenience, the probability functions are abbreviated $F(y)$ and $f(y)$.

If the distribution is unknown, it is common to assume that $F(e)$ is centered about 0 and has a finite, positive definite variance–covariance matrix, Σ_e. Alternately, suppose e is a vector of disturbances such that

$$e \in \mathcal{E}_c = \{z \in \Re^T : z'\Sigma_e^{-1}z \leq c\}$$

for some bound, $c > 0$. The disturbances may be non–random but unknown, or they may be random on a bounded support. Although Σ_e or c are rarely known in practice, the properties of many estimators are largely unaffected if the scale parameter is estimated in a consistent fashion. Nonetheless, the assumption will be weakened later.

1.2.1 Likelihood Methods

The density or mass function, $f(y)$, is perhaps the most common tool used to specify and recover information about the underlying system. By specifying a parametric functional form, researchers employ very specific information about the stochastic properties of y. When viewed as a function of β, $f(y)$ is known as the *likelihood function* of β given y and X, $L_y(\beta)$. The Likelihood Principle (LP) (Berger and Wolpert, 1988) asserts that all of the information about β for a given sample is expressed in $L_y(\beta)$. The frequentist and Bayesian approaches to information recovery employ $L_y(\beta)$, but the associated methods and their interpretation are fundamentally different.

Maximum Likelihood Estimation

The Maximum Likelihood (ML) approach has a long history in statistics and may be traced back to Gauss (Bickel and Doksum, 1977, p. 99). The modern ML

theory is based on efforts by R. A. Fisher (Fisher, 1950) to devise an efficient means of recovering information about the system parameters. Given the parametric form $L_y(\beta)$, ML recovers the image of the system parameter, β, that is 'most likely' for the observed sample, y and X. Formally, the ML estimation problem solves

$$\text{(ML)} \qquad \max_{\beta \in \mathcal{B}} L_y(\beta)$$

A considerable portion of the statistical estimation literature is devoted to the properties and limitations of ML techniques. In general, the ML solutions are admissible because $L_y(\beta)$ is defined on $\mathcal{B}$. Although the finite–sample properties may be unknown, ML estimates are typically consistent as well as asymptotically unbiased, efficient, and normal under suitable sets of regularity conditions (Spanos, 1986; Lehmann, 1983). Conversely, ML solutions may not exist for small samples (ill–posed problems) or in special cases of certain parametric families (e.g. χ^2_1, contaminated normal, three–parameter log–normal).

In ill–conditioned problems, $L_y(\beta)$ may be very 'flat'. Consequently, the global maximum of the likelihood function may be difficult to identify, and the resulting ML estimates will be unstable. The relevant parameter space for the ML problem, $\mathcal{B}$, may be restricted to half–spaces or other subsets of $\Re^K$ consistent with available prior information about β.

Bayesian and Minimax Inference

The Bayesian or subjective approach takes the likelihood function as the sample information used to form inferences about β. By treating the unknown parameters as random variables, prior information about the unknowns may be expressed as a subjective probability distribution, $g(\beta)$. The sample and non–sample information is combined through Bayes Rule to derive the *posterior* distribution of β given the observed sample

$$\begin{aligned} g(\beta|y) &= \frac{f(y|\beta) \cdot g(\beta)}{f(y)} \\ &\propto f(y|\beta) \cdot g(\beta) \end{aligned}$$

The posterior distribution, $g(\beta|y)$, is the principle product of Bayesian inference, but point estimates may be recovered under some loss function, $\mathrm{L}(\beta, \hat{\beta})$. The point estimate, $\hat{\beta}$, minimizes posterior risk (expected loss)

$$\rho(\hat{\beta}) = \int \mathrm{L}(\beta, \hat{\beta}) \cdot g(\beta|y) d\beta$$

If $L(\beta, \hat{\beta})$ is a squared-error, absolute-error, or 0-1 loss function, $\hat{\beta}$ is the mean, median, or mode of $g(\beta|y)$, respectively. Finally, the predictive density, $f(\hat{y}|y)$, may be used to solve the DP.

Advocates of Bayesian inference cite the conceptual appeal of prior probability measures and the implications of the LP as evidence in favor of the subjective approach. In addition, Bayesian analysts may remove nuisance parameters by averaging over their prior distribution, discuss the unknown parameters in probabilistic terms, and examine the robustness of the results under alternate sets of prior information. The principle drawbacks of the Bayesian approach are the inevitable controversy surrounding subjective probability and the analytical and computational burdens of solving most problems. For example, the posterior risk functions must be numerically evaluated if the posterior is not an amenable form or is of high dimension.

Bayesians argue that ML estimates are often special cases of the subjective approach. For example, Bayes Rule implies that the posterior distribution is proportional to the likelihood function under a diffuse prior. Then, the Bayes and ML point estimates will be equal for any posterior risk function that is minimized at the posterior mode. Under an informative prior distribution, the likelihood is weighted to reflect our information about β. Consequently, some authors view the Bayesian approach as a weighted, penalized, or generalized ML approach.

Minimax estimation is related to the Bayes approach in that it considers the global risk-consequences of choosing $\widehat{\beta} \in \mathcal{B}$. The minimax objective guards against very large losses by choosing a point estimate that minimizes the 'maximum risk' function

$$\min_{\widehat{\beta}} \max_{\beta \in \mathcal{B}} \int \mathrm{L}(\beta, \widehat{\beta}) dF(y)$$

The minimax approach has been justified by decision theoretic arguments which show that a minimax estimator is Bayes for the worst possible prior distribution (Lehmann,

1983, Theorem 4.2.1). Critics of the minimax approach argue that guarding against the 'maximum loss' is a very conservative means of information recovery. Further, closed-form minimax estimators rarely exist in practice due to the difficult task of evaluating the maximum risk function.

An alternate formulation of the minimax problem is used in the *optimal recovery* (OR) literature, and Donoho (1994, p. 255) provides an interesting example to illustrate the principles involved. For $K = 1$, suppose we are trying to recover an image of β from a single observation, $y = \beta + e$. Here, $\beta \in [-d, d]$ is unknown, and $e \in [-c, c]$ is selected by an antagonistic opponent. The minimax rule used for optimal recovery is

$$\text{(OR)} \qquad \min_{\widehat{\beta}} \sup_{e} |\widehat{\beta} - \beta| \qquad \widehat{\beta} \leq d \quad e \leq c$$

The solution to the problem is

$$\widehat{\beta}_{OR} = \begin{cases} 0 & \text{if } c < d \\ y & \text{if } c > d \\ \phi y & \text{if } c = d \end{cases} \tag{1.2}$$

for some $\phi \in [0, 1]$. Consequently, games such as the OR problem may be treated as inverse problems. Although most economists are familiar with the minimax criterion in statistical and game-theoretic exercises, these are rarely viewed as cases of the same solution framework.

Other Model-based Methods

The properties of the assumed statistical model, $F(y)$, may be used to derive estimators under alternate criteria. One of the most common approaches is to restrict the class of admissible estimators and choose an optimal member of the class. Common restrictions include unbiasedness or invariance to transformations, and examples of optimality rules are the Best Linear Unbiased (BLU), Uniform Minimum Variance Unbiased (UMVU), and Minimum Risk Equivariant (MRE) criteria.

The James–Stein estimator and other shrinkage techniques (Lehmann, 1983, Section 4.6) are members of a very important class of estimation tools. In general, shrink-

age techniques improve mean squared error (MSE) by multiplying the traditional sample estimators by some shrinkage factor. The resulting estimation rule may be biased but has smaller variance than the original estimator. For an appropriate shrinkage factor, the variance of the shrinkage rule will be small enough to offset the associated bias and reduce MSE. In some cases, the shrinkage rules may be equivalent to Bayesian or MOR estimators. Note that the estimator derived for the optimal recovery example, Equation (1.2), is a shrinkage rule — the observation is used in proportion to the relative strength of the underlying signal–noise ratio.

1.2.2 Methods of Moments

Pearson's Method of Moments (MOM) is one of the oldest techniques for recovering information about unknown parameters. Suppose x is a random variable with distribution $F(x, \theta)$ where θ is a K-vector of unknown parameters. Given T observations of x, the MOM approach attempts to identify the unknowns by using K functions of the sample. Typically, the first K noncentral moments of the population

$$\mu_k = \int x^k dF(x, \theta)$$

are equated with their sample analogs

$$\hat{\mu}_k = T^{-1} \sum_{t=1}^{T} X_t^k$$

The K equations are then solved for the unknown parameters

$$\hat{\theta}_k = g_k(\hat{\mu}_1, \ldots, \hat{\mu}_K)$$

A classic MOM example considers a random variable, x, with unknown mean μ and variance σ^2. Assuming we know the first two sample moments of the data, we can write the moment relations as

$$T^{-1} \sum_{t=1}^{T} X_t = \mu = E[x] \tag{1.3}$$

$$T^{-1} \sum_{t=1}^{T} X_t^2 = \mu^2 + \sigma^2 = E[x^2] \tag{1.4}$$

Solving for μ and σ^2, the MOM estimates are

$$\hat{\mu} = \bar{X}_T \tag{1.5}$$

$$\hat{\sigma}^2 = T^{-1}\sum_{t=1}^{T}(X_t - \bar{X}_T)^2 \tag{1.6}$$

which are identical to the ML estimates for a $N(\mu, \sigma^2)$ model. Davidson and Solomon (1974) note that the MOM and ML rules may be related for certain exponential families.

In general, MOM provides consistent estimators for the unknown parameters. If the sample moments are not one-to-one functions of the population moments and each $g_k(\cdot)$ is continuous, then the estimators, $\{\hat{\theta}\}$, converge in probability to θ if the sample moments are also consistent (Spanos, 1986, p. 256). Fisher and other early critics of MOM point out that more efficient estimators of θ may be developed. In particular, Casella and Berger (1990, p. 342) note that MOM estimates are not generally functions of sufficient statistics, and these may be improved by the implications of the Rao–Blackwell Theorem (Lehmann, 1983, Thm. 1.4.6). However, the efficiency arguments require knowledge of the distribution, and MOM does not use such specific information.

The Generalized Method of Moments (GMM) was developed by Hansen (1982). Given statistical functionals of the sample and the unknown parameters such that

$$E_\theta\{h(y, \theta)\} = 0$$

the sample analogs of the expectations are solved for $\hat{\theta}$. For a given sample, the system of analog equations may not have a solution, and $\hat{\theta}$ is selected to minimize the norm of the residual vector, $\|h(y, \hat{\theta})\|$. The GMM approach is one member of the family of M–estimators devised by Huber (1981).

The class of GMM estimators includes many familiar 'minimum distance' techniques as special cases (e.g. classic MOM and LS). Recent advances in estimation theory and in computing power have renewed interest in the concept of 'moment matching', especially among economists who use models characterized by 'orthogonality' conditions (e.g. FOC or IV relations). If the moment or orthogonality equations are

complex functions of the sample and the unknown parameters, standard distribution theory may not provide a likelihood basis for solving the problem. GMM estimators are typically consistent and asymptotically normal, and they may be asymptotically efficient under an appropriately weighted objective norm.

Although GMM does not require distributional assumptions, efficiency and other properties of the estimator may be improved if such information is available. The GMM estimates are not guaranteed to be theoretically plausible (admissible), and this is often cited as a principle drawback. As in the ML case, prior information about β may be used to restrict the relevant parameter space, but the moment-based approach may be difficult to employ if prior information is more complex. Extensions of the analog principles are explored by Manski (1988), and recent attempts to use moments in a Bayesian framework have been explored by Zellner (1994).

1.2.3 Regularization Methods

As stated at the beginning of the chapter, regularity conditions may be used to derive unique solutions for otherwise ill-posed problems. The conditions may reflect subjective or other prior information, dual objectives for the recovered information, or merely convenient assumptions. Tikhonov (Tikhonov and Arsenin, 1977) formalized the regularization concept for a family of techniques known as the Method of Regularization (MOR).

In general, the MOR objective reflects the fidelity of the recovered system to the indirect observations and to the regularity conditions. Formally, the general MOR objective may written as

$$\text{(MOR)} \qquad \mathcal{L} = \|m(y,\beta)\| + \eta\phi(\beta)$$

where $m(y,\beta)$ follows from Equation (1.1), $\|\cdot\|$ is some norm on $\mathcal{Y}$ and $\phi(\cdot)$ is a penalty function that reflects information about the plausible values of β. The trade-off between the components of the regularization objective is provided by η.

For the GLM, a familiar example is the ridge regressor (Judge, Hill, Griffiths,

Lütkepohl and Lee, 1988, Section 21.4.3) which chooses β to minimize

$$\mathcal{L} = (y - X\beta)'(y - X\beta) + \eta\beta' C\beta$$

Here, the quadratic regularization (QR) problem penalizes parameter vectors that have a large weighted Euclidean norm, $\beta' C\beta$, where C is a positive semi-definite matrix of weights. The ridge or smoothing parameter, η, establishes the trade-off between the squared-error objective and the penalty function. As $\eta \to 0$, $\widehat{\beta} \to \widehat{\beta}_{LS}$, and $\widehat{\beta} \to 0$ as $\eta \to \infty$. Another familiar technique in the MOR family is the cubic spline smoother used in non-parametric regression (Härdle, 1990). The cubic spline rule minimizes the MOR objective

$$\mathcal{L} = (y - g(X))'(y - g(X)) + \eta \int [g''(x)]^2 \, dx$$

among the set of twice continuously differentiable functions, $g \in C^2$.

The regularization techniques are very flexible and may be used to extend a variety of existing methods. In effect, Bayesian inference is a form of regularization that uses Bayes Rule to form the trade-off between sample and prior information. If the penalty function is an alternate loss function on the parameter space (e.g. squared-error loss), the MOR objective provides a 'dual-loss' criterion for information recovery. For example, Zellner (1991) derives a family of estimators under dual quadratic loss functions.

1.2.4 Summary of Traditional Methods

The traditional methods of information recovery provide a variety of plausible approaches in cases where the sample information is well-defined and well-behaved. However, difficulties typically arise if the available information is limited, partial, or incomplete, and such cases are frequently encountered in practice. Although regularization methods may employ prior information to solve the inverse problem, the techniques are often motivated by convenience rather than fidelity to the prior knowledge. For these reasons, it is useful to consider alternate methods designed to solve inverse problems given limited information.

1.3 Information Theoretic Alternatives

After WWII, information theory evolved from wartime research on the increasingly important problems of sending and receiving coded messages over noisy communcation channels. Information theory is one basis for research in the fields of computer science and electrical and electronic engineering. It has also been used in a variety of disciplines to recover information about an unobserved signal from noisy, indirect observations.

1.3.1 Shannon's Entropy

Shannon's *entropy* measures the degree of uncertainty expressed in a probability distribution for a random event. If there are K possible outcomes for the event and p_i is the probability of observing outcome i, the entropy of the distribution is

$$H(p) = -\sum_{i=1}^{K} p_i \log(p_i) \tag{1.7}$$

Assuming $0 \cdot \log(0) = 0$, $H(p) = 0$ for a degenerate probability distribution ($p_i = 1$ for some i), and $H(p) = \log(K)$ when p is discrete uniform.[1] Thus, entropy measures uncertainty by taking a value of zero when the outcome is certain (least uncertain) and achieving a maximum for the most uncertain distribution.

Conceptually, information reduces uncertainty, and entropy measures uncertainty by accounting for expected or missing information about the event. In fact, Theil (1971, p. 25) notes that uncertainty and expected information should be viewed as dual concepts. To demonstrate the notion of expected information, consider a single Bernoulli trial with success probability $p > 0$. Let the trial be a single baseball season, and suppose we observe a success if the Chicago Cubs win the World Series. If p is very small, we would be 'infinitely shocked' *a postieri* to observe a success (i.e. the Cubs won the World Series). Alternately, we would not be very surprised by a success if p is close to 1.

[1] Although the logarithm may take an arbitrary base, it is customary to use base-K logs to scale the range of $H(p)$ to $[0, 1]$. For convenience, all logarithms used in this research will be natural logarithms.

Note that $-\log(\cdot)$ is one function that maps $[0,1]$ to $[0,\infty]$ in this fashion. In the baseball example, a message that the Cubs won the World Series would yield an 'informational score' of

$$-1 \cdot \log(p) - 0 \cdot \log(1-p) = -\log(p)$$

whereas a loss would provide

$$-0 \cdot \log(p) - 1 \cdot \log(1-p) = -\log(1-p)$$

A priori, our expected 'information' is simply

$$-p \cdot \log(p) - (1-p) \cdot \log(1-p) = H(p)$$

Although the $-\log(\cdot)$ function seems rather arbitrary, it is justified by certain axioms of information theory (for a rough proof, see Theil (1967, p. 6)).

To show that entropy also relates to *missing* information, note that there is no missing information if $H(p) = 0$; we know the outcome of the season with certainty (win or lose). If the Cubs' prospects are equally likely ($p = 0.5$), $H(p)$ peaks at $\log(2)$. In this case, we do not have any real information about the outcome of the season. Thus, the concepts of uncertainty, missing information, and expected information may be related through $H(p)$.

A more general form of entropy was later introduced by Kullback (1959)

$$I(p,q) = \sum_{i=1}^{K} p_i \cdot \log\left(\frac{p_i}{q_i}\right) \tag{1.8}$$

which is measure of the distance between distributions p and q.[2] Kullback's measure is called *cross-entropy* (Good, 1963), Kullback–Liebler directed divergence, or I-divergence in the statistics and information theory literature. Shannon's entropy is a special case of $I(p,q)$ because

$$I(p,q) = \sum_{i} p_i \log(p_i K) = -H(p) + \log(K) \tag{1.9}$$

when q is a discrete uniform distribution.

[2]Assuming the support of p is a subset of the support of q

Although $I(p,q) \geq 0 \ \forall \ p,q$ and $I(p,q) = 0$ *iff* $p = q$, $I(p,q)$ is not a true distance function. The commutative property, $I(p,q) = I(q,p)$, does not hold if $p \neq q$. However, $I(p,q)$ is known as a 'directed' divergence because it is a useful measure of uncertainty as we move with the flow of information. For example, suppose p is an observed frequency distribution for a set of independent trials with distribution q. Then, $I(p,q)$ is the 'average' information gathered from the observations. Alternately, q may be a prior distribution with associated posterior distribution, p. In this case, $I(p,q)$ measures the additional sample information reflected in the posterior relative to the prior.

The entropy measures may be extended to distributions defined with respect to more general probability measures. By using the Riemann–Stieltjes integral

$$\int \log\left(p(x)/q(x)\right) dP(x)$$

we can compute the cross–entropy of discrete, continuous, or mixture distributions. However, Georgescu-Roegen (1971, p.395) proves that Shannon's entropy does not extend to continuous distributions. There are also more general forms of entropy that include $H(p)$ and $I(p,q)$ as special cases — refer to Maasoumi (1993) for details.

Finally, it is important to note that informational entropy and physical entropy are distinct concepts. Reportedly, John von Neumann urged Shannon to name his measure entropy, and he cited two appealing reasons. First, the functional form of $H(p)$ is similar to the entropy measures used in physics. Second, von Neumann argued that no one really understands the concept of entropy, and Shannon would enjoy a considerable advantage in future debates. Regardless of the truth in the reports, the damage is done and there has been a great deal of confusion about the relationship between physical and informational entropy. For example, Georgescu-Roegen (1971) focuses on the physical nature of economic systems, but uses both versions of entropy throughout the text. Although conceptual links may be constructed, the two entropies are only related in cases for which the probability distribution is defined over macrostates in a thermodynamic system. Denbigh and Denbigh (1985) provide a good discussion of the conceptual differences and similarities.

1.3.2 Jaynes' Maximum Entropy Formalism

Some direct problems involve predicting discrete actions or choices based on a more complex, but unobserved system. In many cases, an image of the full system may not be required, and a probability distribution over the discrete set of outcomes may be sufficient to solve DP. If the indirect observations are sample moments for the discrete outcomes, these may be expressed as linear functions of the unknown probabilities. Then, the distribution may be recovered by solving the associated linear inverse problem. If the observations are limited and the inverse problem is ill–posed, the traditional methods of inference do not provide a basis for solving the problem.

Jaynes developed the Maximum Entropy (ME) formalism as a feasible means for solving ill–posed *pure* linear inverse problems for unknown probability distributions (Jaynes, 1957a; Jaynes, 1957b). The ME method selects the probability distribution that satisfies the observed information and uses as little extra information as possible. Given that Shannon's entropy may be viewed as a measure of missing information, Jaynes suggested choosing the candidate distribution with maximum entropy. In his own words, the ME solution "agrees with what is known, but expresses 'maximum uncertainty' with respect to all other matters" (Jaynes, 1985, p. 231). Thus, Jaynes uses the information criterion to regularize the ill–posed linear inverse problem and derive a unique solution.

Formally, let y be the vector of T observed moments with associated supports, X. Then, the T moments may be written as a function of the K unknown probabilities, $y = Xp$, which a member of the GLM family. For example, if y_1 is the average outcome, the first row of X is the set of possible outcomes. If y_2 is the average squared outcome, the second row of X contains the set of squared outcomes. In any case, Jaynes' ME solution selects $p \gg 0$ to maximize

$$\text{(ME)} \qquad H(p) = -\sum_{i=1}^{K} p_i \log(p_i)$$

subject to

$$y = Xp \tag{1.10}$$

$$1 = i_K' p \tag{1.11}$$

Typically, the constraints in Equation (1.10) are known as the *model* or consistency constraints, and Equation (1.11) is the *additivity* constraint required for all probability distributions. A formal solution to the problem is presented in the next chapter.

To demonstrate the ME approach, Jaynes devised a simple example known as the dice problem. Suppose you are given a six–sided die and are asked to estimate the probabilities for each possible outcome in the next roll of the die. The only information you are given is y, the average outcome from a large number of independent rolls of the die. Note that you are *not* given the observed frequency distribution of the sample, which is the MLE for a multinomial distribution. The problem is clearly ill–posed because there are six unknown probabilities, but only two pieces of information – the six probabilities must sum to one and the mean of the distribution is y. Although Kolmogorov's second SLLN (Spanos, 1986, p. 170) implies that the sample average will converge almost surely to the true mean of the distribution, there are an infinite number of distributions with a mean of y and supported on $\{1, \ldots, 6\}$.

As stated earlier, an ill–posed problem may become well–posed under additional regularity conditions. In the dice example, suppose we believe the die is 'fair' and has a discrete uniform distribution. If $y = 3.5$, the observed average matches the discrete uniform mean, and we would use the discrete uniform distribution. If $y \neq 3.5$, the underlying distribution is unlikely to be discrete uniform. However, it seems reasonable to select the most 'uniform' (uncertain) of the distributions with a mean of y. Jaynes uses Shannon's entropy to measure uniformity, and the maximum entropy distribution is the most uniform distribution with a mean of y. Trivially, the discrete uniform distribution maximizes entropy and has a mean of $y = 3.5$. Thus, Jaynes uses the uniformity (uncertainty) assumption to regularize the dice problem.

Formally, the ME solution to the dice problem selects p to maximize

$$H(p) = -\sum_{i=1}^{6} p_i \log(p_i) \tag{1.12}$$

subject to

$$\sum_{i=1}^{6} p_i X_i = y \tag{1.13}$$

$$\sum_{i=1}^{6} p_i = 1 \tag{1.14}$$

where $X_i = i$ for each $i = 1, \ldots, 6$. The constraint set is non-empty if $y \in (1, 6)$, and H is strictly concave in p. Thus, there is a unique, interior solution to the dice problem. Trivially, $p_1 = 1$ or $p_6 = 1$ if $y = 1$ or $y = 6$, respectively.

To calculate the interior ME solution, form the Lagrangian expression for the problem

$$\mathcal{L} = -\sum_{i=1}^{6} p_i \log(p_i) + \lambda \left(y - \sum_{i=1}^{6} p_i X_i \right) + \gamma \left(1 - \sum_{i=1}^{6} p_i \right) \tag{1.15}$$

The associated first-order conditions (FOC) are

$$\frac{\partial \mathcal{L}}{\partial p_i} = -1 - \log(\hat{p}_i) - X_i \hat{\lambda} - \hat{\gamma} = 0 \; \forall \; i \tag{1.16}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = y - \sum_{i=1}^{6} \hat{p}_i X_i = 0 \tag{1.17}$$

$$\frac{\partial \mathcal{L}}{\partial \gamma} = 1 - \sum_{i=1}^{6} \hat{p}_i = 0 \tag{1.18}$$

By solving the FOC, we find that the ME probability distribution places weight

$$\hat{p}_i = \frac{\exp(-X_i \hat{\lambda})}{\sum_{j=1}^{6} \exp(-X_j \hat{\lambda})} = \frac{\exp(-X_i \hat{\lambda})}{\Omega(\hat{\lambda})} \tag{1.19}$$

on the i^{th} outcome. Clearly, the ME probabilities are admissible because $\hat{p} \gg 0$ and Equation (1.11) is satisfied. However, $\hat{p}_i$ is a function of $\hat{\lambda}$, the Lagrange multiplier on the model constraint in Equation (1.10). Thus, the Maximum Entropy distribution does not have a closed-form solution, and the problem must be solved numerically. For various values of y, the ME solutions to the dice problem are presented in Table 1.1. The analytical and computational properties of the ME problem are discussed in greater detail in Chapter 2.

A dice–like problem has recently appeared statistics literature on bootstrapping, and an example is presented in Section 23.7 of Efron and Tibshirani (1993). For a

given sample, $\{X_1, \ldots, X_T\}$, the empirical mass function is constucted by placing a weight of T^{-1} on each observation. A variety of estimators and tests may be computed by resampling from the empirical distribution. If the bootstrap procedure is used to compute a test statistic based on the mean of the distribution, an empirical distribution that reflects the mean under the alternate hypothesis may be required. The alternate distribution must be constructed by shifting mass among the observations, and Efron and Tibshirani note that ME is one basis for recovering such a distribution. The ME solution is the distribution that is 'closest' to the empirical mass function, yet reflects the alternate hypothesis.

1.3.3 The Minimum Cross–Entropy Formalism

In the dice problem, suppose we believe that the die is not fair and has some non–uniform distribution, q. The problem may then be solved under the cross–entropy (CE) formalism. CE regularizes the problem by selecting the distribution, p, that has a mean of y and minimizes the Kullback–Liebler directed divergence, $I(p, q)$, between p and q. Intuitively, the CE distribution satisfies the observed information and is 'closest' to our prior beliefs. Alternately, the CE distribution minimizes the additional information reflected in p relative to q.

Formally, the CE method selects p to minimize Equation (1.8) subject to the previous constraint set, Equations (1.10) and (1.11). If the constraint set is non–empty, there is a unique interior solution to the problem because $I(p, q)$ is strictly convex in p. The CE probabilities for the dice problem are

$$\hat{p}_i = \frac{q_i \exp(X_i \hat{\lambda})}{\Omega(\hat{\lambda})} \tag{1.20}$$

where

$$\Omega(\hat{\lambda}) = \sum_n q_n \exp(X_n \hat{\lambda}) \tag{1.21}$$

As before, $\hat{\lambda}$ is the Lagrange multiplier on the model constraint, Equation(1.10). Recall that $I(p, q) = -H(p) + \log(K)$ if q is discrete uniform. In this case, the minimum

cross-entropy and the maximum entropy problems are equivalent, and ME may be viewed as a special case of CE.

The ME solution to the dice problem may be related to broader family of probability distributions with mass function

$$p_\lambda(x,q) = \frac{q(x)\exp(x\lambda)}{\Omega(\lambda)} \tag{1.22}$$

for $\lambda \in \Re$ and $x \in \{1,\ldots,6\}$. $p_\lambda(x,q)$ is a generalized version of the Maxwell-Boltzmann distribution (Rao, 1973, p. 173), which is a member of the univariate canonical exponential family

$$p(x,\theta) = h(x)\cdot\exp\left[\theta t(x) - c(\theta)\right] \tag{1.23}$$

where $h(x) = q(x)$, $\theta = \lambda$, $t(x) = x$, and $c(\theta) = \log\left[\Omega(\lambda)\right]$. The natural parameter space is $\Lambda = \Re$, and the canonical family is full rank (trivially).

The ME distribution, $p_{\hat{\lambda}}(x,q)$, is one member of the Maxwell-Boltzmann family, and the properties of the canonical exponential family may be used to evaluate the ME solution. In particular, the information matrix of $p_{\hat{\lambda}}(x,q)$ is

$$\begin{aligned} I(\hat{\lambda}) &= -E_\theta\left[p''_{\hat{\lambda}}(x,q)\right] \\ &= \sum_{i=1}^{6}\hat{p}_i X_i^2 - \left(\sum_{i=1}^{6}\hat{p}_i X_i\right)^2 = Var_{\hat{\lambda}}(x) \end{aligned} \tag{1.24}$$

which is strictly positive for an interior solution. Given that the moments of $p_{\hat{\lambda}}(x,q)$ match the observed moments, the moment generating function (Bickel and Doksum, 1977, Theorem 2.3.2)

$$\begin{aligned} K(s) &= \exp[c(\theta+s) - c(\theta)] \\ &= \frac{\Omega(\hat{\lambda}+s)}{\Omega(\hat{\lambda})} \end{aligned} \tag{1.25}$$

may be viewed as the *empirical* m.g.f. These properties may be extended to ME and CE solutions to other linear inverse problems.

1.3.4 Pros and Cons of Maximum Entropy

Other criteria may be used to regularize ill-posed problems like the dice example. For example, one could minimize the distance between p and q under Euclidean or other norms. Jaynes argued that the ME solution is analogous to the frequency distribution that could be generated in the largest number of ways and is consistent with the observed information. To see this, consider Boltzmann's derivation of the Maxwell-Boltzmann distribution. Let $p_i = n_i/N$ be the freqency distribution for N independent trials with K possible outcomes. The number of ways of observing a particular distribution is given by the multinomial coefficient

$$W = \frac{N!}{n_1! \ldots n_K!} \tag{1.26}$$

Given the logarithmic version of Stirling's approximation,

$$\begin{aligned} \log(n!) &\approx \frac{\log(2\pi)}{2} + [n + 0.5]\log(n) - n \\ &\propto n\log(n) - n \end{aligned} \tag{1.27}$$

the (monotonic) log-transform of W is

$$\begin{aligned} \log(W) &= \log(N!) - \sum_{i=1}^{K} n_i! \\ &\approx N\log(N) - N - \sum_{i=1}^{K} n_i \log(n_i) + \sum_{i=1}^{K} n_i \\ &= -\sum_{i=1}^{K} n_i \left[\log(n_i) - \log(N)\right] \end{aligned} \tag{1.28}$$

which implies that the average log-multiplicity is

$$\begin{aligned} N^{-1}\log(W) &= -\sum_{i=1}^{K} \left(\frac{n_i}{N}\right) \log\left(\frac{n_i}{N}\right) \\ &= -\sum_{i=1}^{K} p_i \log(p_i) \\ &= H(p) \end{aligned} \tag{1.29}$$

Thus, Shannon's entropy is asymptotically proportional to the multinomial coefficient, and the ME distribution may be associated with the 'most likely' (ML) frequency distribution.

Jaynes' further asserted that Shannon's entropy is the only objective that provides an image of the underlying system that is consistent with the observations. A portion of the assertion was proved by Khinchin (Theil, 1967). Later, Shore and Johnson (1980) and Tikochinsky, Tishby and Levine (1984) independently proved Jaynes' full proposition using axioms of information theory. A brief review of the axiomatic approach is provided by Skilling (1988), and a sketch of the proofs is presented by Theil (1967).

Critics readily point out the principle limitations of ME – it is only designed for pure inverse problems in which the unknowns are probabilities. By dropping the additivity constraint, Donoho, Johnstone, Hoch and Stern (1992) show that $H(p)$ may be used as an MOR penalty function to solve inverse problems with noise for $p \in \Re_+^K$. Although this approach has been used in many signal recovery problems throughout the physical and social sciences, it does not extend to problems with negative unknown parameters. Lacking a more general formulation, the ME and CE formalisms have not been widely accepted as means of information recovery.

1.4 Purpose and Objectives of the Dissertation

In summary, the traditional methods of information recovery provide a variety of bases for solving inverse problems. Although many well–posed experimental designs may fit at least one of the alternates, difficulties arise in ill–posed or ill–conditioned problems. If prior economic knowledge exists, it may be used to augment the limited indirect observations or to regularize the inverse problems. Unfortunately, the traditional methods generally require different types of information and may not accomodate the available information. Consequently, researchers often employ creative or simplifying assumptions to solve economic inverse problems within the traditional framework.

Despite the cited drawbacks, the entropy formalisms are an attractive means of information recovery. By using the available indirect observations and prior information, economists may use the entropy approach to solve inverse problems without additional or unnatural assumptions. These motives prompted Judge and Golan (1992)

to extend the original entropy framework to a generalized entropy formalism. The Generalized Maximum Entropy and Generalized Cross–Entropy problems account for the disturbances and real–valued parameters commonly found in economic models, and the solutions agree with the sample information and reflect the prior knowledge.

The purpose of the dissertation is to examine proposed extensions of the entropy formalism to inverse problems in economics. The objectives of the research are:

(i). to specify the generalized entropy formulations of the GLM,

(ii). to demonstrate the analytical and computational properties of the generalized entropy solutions, and

(iii). to examine the performance of the generalized entropy techniques in familiar cases of the GLM.

The generalized entropy methods for the GLM are specified and the mathematical and statistical properties of the inverse problems are presented in the next chapter. Using limited Monte Carlo evidence, the performance of the proposed methods are demonstrated in Chapter 3. Finally, a summary of the research, conclusions about the generalized entropy methods, and some additional extensions are discussed in Chapter 4.

y	p_1	p_2	p_3	p_4	p_5	p_6	$H(p)$
1.0	1.000	0.000	0.000	0.000	0.000	0.000	0.000
1.5	0.664	0.224	0.075	0.025	0.009	0.003	0.953
2.0	0.478	0.255	0.136	0.072	0.038	0.021	1.367
2.5	0.348	0.240	0.165	0.114	0.079	0.054	1.614
3.0	0.247	0.207	0.174	0.146	0.123	0.103	1.748
3.5	0.167	0.167	0.167	0.167	0.167	0.167	1.792
4.0	0.103	0.123	0.146	0.174	0.207	0.247	1.748
4.5	0.054	0.079	0.114	0.165	0.240	0.348	1.614
5.0	0.021	0.038	0.072	0.136	0.255	0.478	1.367
5.5	0.003	0.009	0.025	0.075	0.224	0.664	0.953
6.0	0.000	0.000	0.000	0.000	0.000	1.000	0.000

Table 1.1: ME Solution to the Dice Problem for Various y

Chapter 2

The Generalized Entropy Approach to Information Recovery

2.1 The GME–GCE Framework

As stated in Chapter 1, the information theoretic basis for solving economic inverse problems is attractive due to the limitations of the indirect observations and the presence of prior information. However, most economic inverse problems include a noise component, and many of the unknown parameters may take on real values. Consequently, the entropy formalisms are not directly applicable to cases such as the GLM. Judge and Golan (1992) extended the ME formalism to handle real-valued unknowns in linear inverse problems with noise by reformulating the GLM in terms of unknown probability distributions.

Assume $\mathcal{B}$ may be represented by a compact hyperrectangle, $\mathcal{Z} \subset \Re^K$. If Z_{k1} and Z_{k2} are the extreme possible values of β_k, there exists $p_k \in [0,1]$ such that

$$\beta_k = p_k Z_{k1} + (1 - p_k) Z_{k2} \tag{2.1}$$

In general, let Z_k be a set of $M \geq 2$ points that span the k^{th} dimension of $\mathcal{Z}$, and let p_k be the associated M-vector of weights on these points. Then, any $\beta \in \text{int}(\mathcal{Z})$ may be expressed as

$$\beta = Zp = \begin{bmatrix} Z_1 & 0 & \cdot & 0 \\ 0 & Z_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & Z_K \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \cdot \\ p_K \end{bmatrix} \tag{2.2}$$

where Z is a $(K \times KM)$ matrix and p is a KM-vector of weights such that $p_k \gg 0$ and $p_k' \iota_M = 1$ for each k.

Prior information about the unknowns may be expressed as a set of subjective probability distributions over Z. If the prior weights are q, the prior mean of the parameters is Zq. For example, let β_1 be an elasticity of supply. Suppose we believe $\beta_1 \in [0,4]$, and our prior expectation is $\beta_1 = 1$. Then, $Z_{11} = 0$ and $Z_{12} = 4$ may be used as supports on β_1, and the prior distribution may be $q = [0.75, 0.25]$. The number of support points for each parameter, M, may be increased to reflect the available prior information.

The disturbances may be treated in a similar fashion. Suppose there exist sets of error bounds, V_{t1} and V_{t2}, for each e_t so that $\Pr[V_{t1} < e_t < V_{t2}]$ may be made arbitrarily small. With positive probability, each disturbance may be written as

$$e_t = w_t V_{t1} + (1 - w_t) V_{t2} \tag{2.3}$$

for some $w_t \in (0, 1)$. Typically, V_{t1} and V_{t2} will be symmetric about zero. However, $J \geq 2$ points may be used to express or recover additional information about e_t (e.g. skewness). The T unknown disturbances may be written as

$$e = Vw = \begin{bmatrix} V_1 & 0 & \cdot & 0 \\ 0 & V_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & V_T \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ w_T \end{bmatrix} \tag{2.4}$$

where V is a $(T \times TJ)$ matrix and w is a TJ–vector of weights such that $w \gg 0$ and $w_t' \imath_J = 1$ for each t.

By defining $\beta = Zp$ and $e = Vw$, Judge and Golan rewrite the general linear model as

$$y = X\beta + e = XZp + Vw \tag{2.5}$$

Given Z and V, the Generalized Maximum Entropy (GME) solution to the linear inverse problem with noise selects $p, w \gg 0$ to maximize

$$H(p, w) = -p' \log(p) - w' \log(w) \tag{2.6}$$

subject to

$$y = XZp + Vw \tag{2.7}$$

$$\imath_K = (I_K \otimes \imath_M')p \tag{2.8}$$

$$\imath_T = (I_T \otimes \imath_J')w \tag{2.9}$$

where Equation (2.7) is the model constraint, and Equations (2.8) and (2.9) provide the additivity constraints. The optimal probability vector, $\hat{p}$, may be used to form a point estimate of the unknown parameter vector, $\hat{\beta} = Z\hat{p}$.

Analogous to Jaynes' explanation of Maximum Entropy, GME selects weights for the elements of Z and V that are most 'uncertain' *and* satisfy the observed information. Given informative prior distributions on Z and V, q and u, the cross-entropy criterion may be used to form the Generalized Cross Entropy (GCE) problem. The GCE solution is the set of weights, p and w, that are 'closest' to the prior weights and satisfy the observations. Conseqently, generalized entropy may be viewed as a form of minimum distance or GMM estimation. However, the observations are used as constraints, and GME–GCE can solve traditionally ill–posed inverse problems. Further, the method requires very little information about the noise process, and prior information about the unknown parameters may be included.

A large number of models may be written in the GLM form

$$\alpha = \Gamma\beta + \epsilon \tag{2.10}$$

where the associated generic GCE formulation is

$$\alpha = \Gamma Z p + V w \tag{2.11}$$

A list of the familiar special cases of the generic GLM is presented in Table 2.1. The suffices D, M, IV, and NM denote the data, moment, instrumental variable, and normed–moment formulations. The table also includes the associated GME problems, which are special cases of the GCE problems when q and u are discrete uniform distributions. For example, classical ME problems such as the dice example are a

Model	α	Γ	ϵ
CE	y	X	0
GCE–D	y	X	e
GCE–M	$X'y$	$X'X$	$X'e$
GCE–IV	$P'y$	$P'X$	$P'e$
GCE–NM	$\left(\frac{X'y}{T}\right)$	$\left(\frac{X'X}{T}\right)$	$\left(\frac{X'e}{T}\right)$

Table 2.1: Cases of the Generic GCE Problem

special cases of the CE model — y is the observed mean, X is the support of the die, and the noise term is excluded.

The formal solution to the generic GCE problem is presented in the next section. As in the ME–CE formulation, the GCE objective is strictly convex on the interior of the additivity constraint set, and a solution exists if the intersection of the consistency and additivity constraint sets is non-empty. As we shall see, the generic GCE probabilities are

$$\hat{p}_{km} = \frac{q_{km} \exp(Z_{km}\Gamma_k'\hat{\lambda})}{\Omega_k(\hat{\lambda})} \tag{2.12}$$

where

$$\Omega_k(\hat{\lambda}) = \sum_{n=1}^{M} q_{kn} \exp(Z_{kn}\Gamma_k'\hat{\lambda}) \tag{2.13}$$

The GCE problem does not have a closed–form solution, and the solution must be computed numerically.

Although $\beta \in \mathcal{Z}$ is the same for each version of the generic GCE problem, the choice of V clearly depend on the properties of ϵ. Chebychev's Inequality may be used as a conservative means of specifying sets of error bounds. For any random variable, ϵ, such that $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$, the inequality provides

$$\Pr[\ |\epsilon| < v\sigma] \geq v^{-2} \tag{2.14}$$

for arbitrary $v > 0$. Given some excluded tail probability, v^{-2}, the extreme error bounds will be $V_{t1} = -v\sigma$ and $V_{tJ} = v\sigma$. An example is the familiar 3-σ rule which excludes at most one–ninth of the mass for $v = 3$. If the ϵ has a unimodal Lebesgue density, the 3-σ rule excludes at most 5% of the tails. Pukelsheim (1994) provides a recent discussion of probability bounds and the 3-σ rule.

Suppose the elements of X are bounded and the GLM disturbances, $\{e_t\}$, are white noise disturbances with unit variance. In the GCE–M model, the variance of ϵ_i is $\sigma_i^2 = X_i'X_i$, and $\sigma_i = O(\sqrt{T})$. In the GCE–NM case,

$$\sigma_i = O\left(\frac{1}{\sqrt{T}}\right)$$

Consequently, the elements of V should reflect the variation (e.g. variance, support) of the underlying errors and may be a function of T.

An alternate version of the GCE formulation may be used if probability bounds are unattainable or inappropriate for a given inverse problem. Given support Z and prior q, choose p to minimize

$$\mathcal{L} = \|y - XZp\| + \eta I(p, q) \tag{2.15}$$

where $\|\cdot\|$ is some norm on $\mathcal{Y}$ and $I(p, q)$ is used as a penalty function. The objective function is clearly a member of the MOR class, and it may be viewed as a modified version of the entropy–penalized objective discussed by Donoho et al. (1992). Although this formulation avoids error bounds, the optimal probabilities do not take a closed–form (even as a function of the Lagrange muultipliers), and the researcher must choose the smoothing parameter, η, and the objective norm. The alternate formulation is not consistent with the generalized entropy framework and is only presented here for completeness.

2.2 Solving the Generic GCE Problem

The generic GCE problem selects $p, w \gg 0$ to minimize

$$I(p, w, q, u) = p' \log(p/q) + w' \log(w/u) \tag{2.16}$$

subject to

$$\alpha = \Gamma Z p + V w \tag{2.17}$$

$$\begin{aligned} \imath_K &= (I_K \otimes \imath_M')p \\ \imath_T &= (I_T \otimes \imath_J')w \end{aligned} \tag{2.18}$$

Note that the additivity constraint set (2.18) is composed of K unit simplices of dimension $M \geq 2$ and T unit simplices of dimension $J \geq 2$. Denote these simplices as S_M and S_J, respectively, so that the additivity constraint set (2.18) is $\mathcal{A} = S_M^K \times S_J^T$. Clearly, $\mathcal{A}$ is a non–empty and compact set. The model constraint set (2.17) further restricts $\mathcal{A}$ to those probabilities that are 'consistent' with the data. Let the fully restricted constraint set be

$$\mathcal{A}^* = \{(p, w) \in \text{int}(\mathcal{A}) : \alpha = \Gamma Z p + V w\} \tag{2.19}$$

To verify the uniqueness of the solution, note that the Hessian matrix of the objective function is

$$\nabla_{(p,w)(p',w')} I(p,w) = \begin{bmatrix} P^{-1} & 0 \\ 0 & W^{-1} \end{bmatrix} \tag{2.20}$$

where P^{-1} is a $(KM \times KM)$ diagonal matrix with elements p_{km}^{-1}, and W^{-1} is a $(TJ \times TJ)$ diagonal matrix with elements w_{tj}^{-1}. The matrix is positive definite for $p, w \gg 0$, which satisfies the sufficient condition for strict convexity. So, there is a unique global minimum (GM) for the problem if $\mathcal{A}^* \neq \emptyset$.

To find the interior solution, form the Lagrangean equation

$$\mathcal{L} = I(p,w) + \lambda'[\alpha - \Gamma Z p - V w] + \gamma'[\imath_K - (I_K \otimes \imath_M')p] + \tau'[\imath_T - (I_T \otimes \imath_J')w]$$

where $\lambda \in \Re^T$, $\gamma \in \Re^K$, and $\tau \in \Re^T$ are the associated Lagrange multipliers. Taking the gradient of $\mathcal{L}$ to derive the first-order conditions (FOC), we have

$$\nabla_p \mathcal{L} = \imath_{KM} + \log(\hat{p}/q) - Z'\Gamma'\hat{\lambda} - (I_K \otimes \imath_M)\hat{\gamma} = 0 \tag{2.21}$$

$$\nabla_w \mathcal{L} = \imath_{TJ} + \log(\hat{w}/u) - V'\hat{\lambda} - (I_T \otimes \imath_J)\hat{\tau} = 0 \tag{2.22}$$

$$\nabla_\lambda \mathcal{L} = \alpha - \Gamma Z \hat{p} - V\hat{w} = 0 \tag{2.23}$$

$$\nabla_\gamma \mathcal{L} = \imath_K - (I_K \otimes \imath_M')\hat{p} = 0 \tag{2.24}$$

$$\nabla_\tau \mathcal{L} = \imath_T - (I_T \otimes \imath_J')\hat{w} = 0 \tag{2.25}$$

Solving Equations (2.21) and (2.22) for $\hat{p}$ and $\hat{w}$, respectively,

$$\hat{p} = q \odot \exp(Z'\Gamma'\hat{\lambda}) \odot \exp[-\imath_{KM} + (I_K \otimes \imath_M)\hat{\gamma}] \tag{2.26}$$

$$\hat{w} = u \odot \exp(V'\hat{\lambda}) \odot \exp[-\imath_{TJ} + (I_T \otimes \imath_J)\hat{\tau}] \tag{2.27}$$

Considering just $\hat{p}_k$, the distribution for β_k, note that the term

$$\exp[-\imath_{kM} + (I_K \otimes \imath_M)\hat{\gamma}_k]$$

is the same for all m. So, $\hat{p}_{km} \propto q_{km} \odot \exp(Z_{km}\Gamma_k'\hat{\lambda})$ for all m of each k. Then, the additivity constraints can be satisfied by using the sum of the m kernels to normalize each $\hat{p}_{km}$. A similar normalization factor may be derived for $\hat{w}$.

More formally, substitute Equations (2.26) and (2.27) into Equations (2.24) and (2.25), respectively. Considering just $\hat{p}$, the normalization factor may be identified by

further premultiplication

$$
\begin{aligned}
(2.28) \qquad (I_K \otimes \imath_M)\imath_K &= (I_K \otimes \imath_M)(I_K \otimes \imath'_M)\hat{p} \\
\imath_{KM} &= (I_K \otimes J_M)\hat{p} \\
\imath_{KM} &= (I_K \otimes J_M)[q \odot \exp(Z'\Gamma'\hat{\lambda}) \\
&\qquad \odot \exp(-\imath_{KM} + (I_K \otimes \imath_M)\hat{\gamma})] \\
\imath_{KM} &= \{(I_K \otimes J_M)[q \odot \exp(Z'\Gamma'\hat{\lambda})]\} \\
&\qquad \odot \exp(-\imath_{KM} + (I_K \otimes \imath_M)\hat{\gamma})
\end{aligned}
$$

By inverting the bracketed term, the result may be rewritten as

$$
(2.29) \qquad \exp(-\imath_{KM} + (I_K \otimes \imath_M)\hat{\gamma}) = \{(I_K \times J_M)[q \odot \exp(Z'\Gamma'\hat{\lambda})]\}^{-1}
$$

and by substitution into Equation (2.26)

$$
(2.30) \qquad \hat{p} = q \odot \exp(Z'\Gamma'\hat{\lambda}) \odot \{(I_K \otimes J_M)[q \odot \exp(Z'\Gamma'\hat{\lambda})]\}^{-1}
$$

The individual probabilities take the form

$$
(2.31) \qquad \hat{p}_{km} = \frac{q_{km} \exp(Z_{km}\Gamma'_k\hat{\lambda})}{\Omega_k(\hat{\lambda})}
$$

where Γ_k is the k^{th} column of Γ, and

$$
(2.32) \qquad \Omega_k(\hat{\lambda}) = \sum_n q_{kn} \exp(Z_{kn}\Gamma'_k\hat{\lambda})
$$

is the *partition function* (normalization). In similar fashion, the vector of optimal noise probabilities is

$$
(2.33) \qquad \hat{w} = u \odot \exp(V'\hat{\lambda}) \odot \{(I_T \otimes J_J)[u \odot \exp(V'\hat{\lambda})]\}^{-1}
$$

with individual elements

$$
(2.34) \qquad \hat{w}_{tj} = \frac{u_{tj} \exp(V_{tj}\hat{\lambda})}{\Psi_t(\hat{\lambda}_t)}
$$

The partition function for $\hat{w}$ is

$$
(2.35) \qquad \Psi_t(\hat{\lambda}) = \sum_n u_{tn} \exp(V_{tn}\hat{\lambda}_t)
$$

Clearly, the GCE solutions, $\hat{p}$ and $\hat{w}$, satisfy the additivity constraints (2.18) and are strictly positive. However, the GCE solution depends on the Lagrange multipliers for the model constraints, $\hat{\lambda}$. The only remaining information in the FOC is the set of model constraints (2.17), which are not a function of λ. Hence, there is no known closed-form solution to the GCE or GME problems, as in Jaynes' dice problem. The GCE solution must be found numerically, and an efficient computing algorithm is presented in the next section.

2.3 Computing the Numerical Solution

Although computing power is no longer a serious limitation to empirical research, there are clear advantages to using efficient techniques that may be employed in a broad set of computing environments. The purpose of this section is to specify the 'dual' version of the generic GCE problem and solve it with simpler and more widely available unconstrained numerical techniques. As we shall see, the dual formulation is also a valuable tool for evaluating the properties of the GCE solution.

2.3.1 Minimal Value Function

For arbitrary $\lambda \in \Re^K$, let $p(\lambda)$ and $w(\lambda)$ represent the functional form of the optimal GCE probabilities, Equations (2.30) and (2.33). Then, substitute these into the original Lagrangean expression to form the *minimal value* function. Note that the optimal probabilities satisfy the additivity constraints $\forall\ \lambda \in \Re^T$, and the associated term may be dropped from $\mathcal{L}$ to yield

$$
\begin{aligned}
(2.36)\quad \mathcal{L}(\lambda) &= p(\lambda)'\log(p(\lambda)) + w(\lambda)'\log(w(\lambda)) + \lambda'[\alpha - \Gamma Z p(\lambda) - V w(\lambda)] \\
&= p(\lambda)'[Z'\Gamma'\lambda - \log(\Omega(\lambda))] + w(\lambda)'[V'\lambda - \log(\Psi(\lambda))] \\
&\qquad +[\alpha' - p(\lambda)'Z'\Gamma' - w(\lambda)'V']\lambda \\
&= \alpha'\lambda - p(\lambda)'\log(\Omega(\lambda)) - w(\lambda)'\log(\Psi(\lambda)) \\
&= \alpha'\lambda - \sum_k \log(\Omega_k(\lambda)) - \sum_t \log(\Psi_t(\lambda)) \equiv M(\lambda)
\end{aligned}
$$

In many optimization problems, the minimal value function is used to evaluate the optimal objective function $I(p, w)$ given alternate values of the constants (e.g Z or V). Without a closed–form solution for $\hat{p}$ and $\hat{w}$, the GCE minimal value function does not have a closed–form. However, the *saddle–point* properties of the GCE problem may be used to find $\hat{\lambda}$ by solving an *unconstrained* problem.

2.3.2 Saddle–point Properties and the Dual Problem

The ability to recover $\hat{\lambda}$ from an unconstrained optimization problem follows from the next result.

Proposition 2.1 *If $\mathcal{A}^* \neq \emptyset$, the optimal solution to the GCE problem, $(\hat{p}, \hat{w}, \hat{\lambda})$, satifies the saddle–point (SP) property:*

$$\mathcal{L}(p, w, \hat{\lambda}) \geq \mathcal{L}(\hat{p}, \hat{w}, \hat{\lambda}) \geq \mathcal{L}(\hat{p}, \hat{w}, \lambda)$$

Proof: If $\mathcal{A}^* \neq \emptyset$, the strict convexity of $I(p, w)$ ensures that the GCE problem has a unique GM, $(\hat{p}, \hat{w}, \hat{\lambda})$. Clearly, a GM is also a local minimum (LM). The linearity of the constraint set defined by Equations (2.17) and (2.18) satisfies the second condition of the Arrow–Hurwicz–Uzawa Constraint Qualification (Takayama, 1985, Theorem 1.D.4), and the LM is also a quasi–saddle–point (QSP) by the Kuhn–Tucker Theorem (Takayama, 1985, Theorem 1.D.3). Finally, the linearity (hence concavity) of the constraints satisfies result *(ii)* of Theorem 1.D.1 in Takayama (1985). Therefore, the unique GM for the GCE problem also satisfies SP. □

In terms of $M(\lambda)$, SP implies

$$M(\lambda) \leq M(\hat{\lambda}) \; \forall \; \lambda \in \Re^T \tag{2.37}$$

Further, the inequality is strict because M is a strictly concave function of λ. To see this, note that the gradient of the dual problem is

$$\nabla_\lambda M(\lambda) = \alpha - \Gamma Z p(\lambda) - V w(\lambda) \tag{2.38}$$

which is simply the model constraint from the generic GCE problem. The Hessian matrix of $M(\lambda)$ is

$$\begin{aligned} \nabla_{\lambda\lambda'} M(\lambda) &= -\Gamma Z \nabla_{\lambda'} p(\lambda) - V \nabla_{\lambda'} w(\lambda) \\ &= -\Gamma \Sigma_Z(\lambda) \Gamma' - \Sigma_V(\lambda) \end{aligned} \tag{2.39}$$

where $\Sigma_Z(\lambda)$ and $\Sigma_V(\lambda)$ are the variance–covariance matrices for distributions $p(\lambda)$ and $w(\lambda)$.

By the sufficient condition for strict concavity, it suffices to show that $\nabla_{\lambda\lambda'} M(\lambda)$ is a negative definite matrix. To evaluate the Hessian matrix, note that the t^{th} equation in $\nabla_\lambda M(\lambda)$ is

$$\alpha_t - \Gamma_t Z p(\lambda) - V_t w_t(\lambda) \tag{2.40}$$

The second–partial derivative of this equation with respect to λ_s is

$$\frac{\partial^2 M}{\partial \lambda_s \partial \lambda_t} = -\sum_k \Gamma_{tk} \sum_m Z_{km} \frac{\partial p_{km}(\lambda)}{\partial \lambda_s} - \sum_j V_{tj} \frac{\partial w_{tj}(\lambda_t)}{\partial \lambda_s} \tag{2.41}$$

where

$$\frac{\partial p_{km}(\lambda)}{\partial \lambda_s} = \Gamma_{sk} \left[Z_{km} p_{km} - p_{km} \sum_n Z_{kn} p_{kn} \right] \tag{2.42}$$

and

$$\frac{\partial w_{tj}(\lambda_t)}{\partial \lambda_s} = \begin{cases} V_{tj} w_{tj} - w_{tj} \sum_n V_{tn} w_{tn} & s = t \\ 0 & s \neq t \end{cases} \tag{2.43}$$

Finally, note that

$$\sigma^2_{Z_k} = \sum_m p_{km} Z^2_{km} - \left[\sum_m p_{km} Z_{km} \right]^2 \tag{2.44}$$

$$\sigma^2_{V_t} = \sum_j w_{tj} V^2_{tj} - \left[\sum_j w_{tj} V_{tj} \right]^2 \tag{2.45}$$

which implies

$$\frac{\partial^2 M}{\partial \lambda_s \partial \lambda_t} = -\sum_k \Gamma_{tk} \Gamma_{sk} \sigma^2_{Z_k} - \sigma^2_{V_t} \tag{2.46}$$

For any interior solution, $(\hat{p}, \hat{w})$, each of these variance terms is strictly positive, and Σ_Z and Σ_V are positive definite matrices.

Assembled in matrix form, these second-partials take the form of Equation (2.39). Although $\Gamma\Sigma_Z\Gamma'$ is positive semi-definite when $T > K$, Σ_V is a positive definite matrix. Hence, $\Gamma\Sigma_Z\Gamma' + \Sigma_V$ is positive definite, and (2.39) is a negative definite matrix. Therefore, $M(\lambda)$ is strictly concave in λ, and choosing λ to maximize $M(\lambda)$ will yield the unique solution, $\hat{\lambda}$. Under the dual formulation, the GCE solution may be computed with simpler unconstrained techniques.

2.3.3 A Simple Computer Algorithm

1. Specify the vector of starting values; $\lambda^0 = 0$ yields $\widehat{\beta}$ equal to its prior mean and is often a good choice.

2. Check for a feasible solution:

   (a) The eigenvalues of $\nabla_{\lambda\lambda'}M(\lambda^0)$ must be strictly positive for some λ^*, which should be near λ^0.

   (b) Use a search algorithm on the parameter space (primal or dual); an example is the linear programming subroutine suggested by Agmon, Alhassid and Levine (1979).

3. Form $M(\lambda)$ as in Equation (2.36):

   (a) Proceed with a derivative-free optimization method (e.g. downhill simplex of Nelder and Mead).

   (b) Form $\nabla_\lambda M(\lambda)$ as in Equation (2.38) and use a gradient-based method.

   (c) Also form $\nabla_{\lambda\lambda'}M(\lambda)$ as in Equation (2.39) to capture second-order improvements in the convergence rate.

4. Use $\hat{\lambda}$ to compute $p(\hat{\lambda})$ and $\widehat{\beta} = Zp(\hat{\lambda})$ as well as $w(\hat{\lambda})$, $\hat{e} = Vw(\hat{\lambda})$, or $I(\hat{p}, \hat{w}, q, u)$.

The Agmon et al. (1979) algorithm is a popular means for solving classical ME–CE problems, and it is a special case of the present algorithm. To see this, set $\epsilon \equiv 0$ and solve the pure linear inverse problem where $K = 1$, $XZ_{1m} = X_m$, and $\beta \equiv p$. Other approaches to computing the solutions to pure inverse problems are summarized by Shore and Johnson (1981). Based on a limited number of trials, the computing time for the dual formulation is roughly 35% less than for the constrained (primal) problem.

2.4 Sampling Properties of the GCE Solution

As explained in Chapter 1, the entropy–based methods of information recovery are not directly motivated by standard sampling theory. However, large– and small–sample properties have been used to compare competing estimators. For example, researchers may compute the bias of a Bayesian point estimator. Although the GCE solution does not have a closed–form, the dual formulation of the GCE problem may be used to evaluate the behavior of the solutions within the extremum or M–estimation framework developed by Huber (1981).

2.4.1 Asymptotic Behavior

The model constraints used in Jaynes' classical ME problems were implicitly assumed to be consistent sample moments with negligible noise components. As T increases, the model constraint converges almost surely, and the ME–CE problem is asymptotically non–stochastic. Although the GCE–D version of the GLM does not fit this pattern, large–sample properties for GCE–NM problem may be derived.

For the purpose of the following analysis, define the GCE-NM model as

$$\min_{p,w} I(p,q,w,u) = p'\log(p/q) + w'\log(w/u) \tag{2.47}$$

subject to

$$\frac{X'y}{T} = \left(\frac{X'X}{T}\right) Zp + Vw \tag{2.48}$$

$$\imath_K = (I_K \otimes \imath_M')p \tag{2.49}$$

$$\imath_K = (I_K \otimes \imath_J')w \tag{2.50}$$

where V is now a $(K \times KJ)$ matrix that specifies the support of the K-vector of residuals. Based on the results from the preceding section, the dual formulation of the problem is

$$\max_{\lambda} M_T(\lambda) = \left(\frac{y'X}{T}\right)\lambda + \sum_k \log\left[\Omega_k(\lambda)\right] + \sum_k \log\left[\Psi_k(\lambda)\right] \tag{2.51}$$

where $\lambda \in \Re^K$ and the partition functions are

$$\Omega_k = \sum_n q_{kn} \exp\left(Z_{kn}\left(\frac{X_k X'}{T}\right)\lambda\right) \tag{2.52}$$

$$\Psi_k = \sum_n u_{kn} \exp\left(V_{kn}\lambda_k\right) \tag{2.53}$$

As we shall see, the GCE–NM solution is consistent under the following assumptions:

(A1) $\beta \in \text{int}(\mathcal{Z})$

(A2) There exists a finite, positive definite matrix Q such that

$$\lim_T \left(\frac{X'X}{T}\right) = Q$$

(A3) $E(e) = 0$, $\text{Var}(e) = \Sigma_e$, and $F(e)$ satisfies the Lindeberg condition (Billingsley, 1986, Eq. 27.8)

$$T^{-1}\sum_{t=1}^{T}\int_{\mathcal{E}} \|e\|^2 dF(e) \to 0$$

where $\mathcal{E} = \{e : \|e\| > \varepsilon\sqrt{T}\}$ for $\varepsilon > 0$.

(A4) The variance–covariance matrix of $\epsilon = X'e/\sqrt{T}$ converges to a finite, positive definite matrix

$$\lim_T \left(\frac{X'\Sigma_e X}{T}\right) = \Sigma^*$$

The Chebychev and 3-σ rules provide error bounds that are proportional to the underlying standard error of the disturbances. Under the preceding assumptions, $\epsilon/\sqrt{T} \xrightarrow{p} 0$. Accordingly, the error bounds used in the GCE–NM problem should collapse on 0 as T increases. In the present discussion, we will consider $V = O(\sqrt[-1]{T})$, but the rate of convergence will be altered later.

Before proceeding, define the following items:

(D1) $\lambda_0 \in \Re^K$ is uniquely and implicitly defined as $\beta = Zp(\lambda_0)$ for some $\beta \in \text{int}(\mathcal{Z})$.

(D2) $\Lambda \subset \Re^K$ is an open neighborhood of λ_0.

(D3) $\bar{\Lambda}$ is the topological closure of Λ.

(D4) $M_T(\lambda)$ is the GCE–NM dual objective function (refer to Equation (2.36)) for sample size T.

(D5) $\hat{\lambda}_T = \max_{\lambda \in \bar{\Lambda}} M_T(\lambda)$, which exists for all T by the Weierstrass Theorem.

(D6) $\widehat{\beta}_T = Zp(\hat{\lambda}_T)$

(D7) $\mathcal{Z}_\Lambda$ is the range of $Zp(\lambda)$ for all $\lambda \in \Lambda$.

(D8) $\mathcal{Y}_\Lambda$ is the range of $QZp(\lambda)$ for all $\lambda \in \Lambda$.

Finally, consider three preliminary results that will be useful in demonstrating the large–sample properties of the GCE–NM solution. The first result relates the original parameter space, $\mathcal{B}$, to the solution space for the dual formulation, Λ. The conceptual basis for these operations is also used to derive large sample results for ML solutions in exponential families (Brown, 1986; Johansen, 1979).

Lemma 2.1 *$Zp(\lambda)$ is a diffeomorphism from Λ to $\mathcal{Z}_\Lambda$ for all sufficiently large T.*

Proof: A *diffeomorphism* is a mapping from one set to another which is one-to-one, differentiable, and invertible in each direction. Let $\omega(\lambda) = Zp(\lambda)$, where $\omega : \Lambda \to \mathcal{Z}_\Lambda$. To establish the local properties, note that the function is continuously differentiable with Jacobian matrix

$$\omega' = \nabla_{\lambda'} Zp(\lambda) = \left(\frac{X'X}{T}\right) \Sigma_Z(\lambda) \tag{2.54}$$

as a special case of Equation (2.39). Note that for all sufficiently large T, ω' is positive definite for all $\lambda \in \Lambda$ by Assumption A2. By the Inverse Function Theorem, ω has a unique continuously differentiable inverse, $\rho : \mathcal{Z}_\Lambda \to \Lambda$, with Jacobian matrix

$$\left[\left(\frac{X'X}{T}\right) \Sigma_Z(\lambda)\right]^{-1} \tag{2.55}$$

which is also positive definite in large samples. Thus, ρ also satisfies the Inverse Function Theorem, and the local mapping is a diffeomorphism in large samples.

The global relation may be demonstrated by contradiction. Suppose there exists two vectors, $\lambda_1 \neq \lambda_2$, such that

$$\beta = Zp(\lambda_1) = Zp(\lambda_2) \tag{2.56}$$

The equality holds for both distributions on Z, so the analysis may be restricted to the probabilities. WLOG, let $\lambda_1 = 0$, and note that Z_{km} in each term in the denominator does not cancel. Consequently, just consider the numerator of the individual probabilities

$$q_{km} \exp\left(Z_{km}\left(\frac{X_k X'}{T}\right)\lambda_2\right) = \exp(0) \tag{2.57}$$

This only holds for $\lambda_2 \neq 0$ if $X'X$ is rank deficient, which contradicts A2 in large samples. □

Lemma 2.2

$$\lim_T \Pr\left(\frac{X'e}{T}\right) = 0$$

Proof: By Assumption A4, $E[X'e] = 0 \; \forall \; T$, and

$$\lim_T \mathrm{Var}\left(\frac{X'e}{T}\right) = \lim_T T^{-1}\Sigma^* = 0 \tag{2.58}$$

So, $T^{-1}X'e$ converges (in quadratic mean) to a null vector, which implies $T^{-1}X'e \xrightarrow{p} 0$ by Chebychev's Inequality. □

Lemma 2.3

$$\left(\frac{X'e}{\sqrt{T}}\right) \Rightarrow N\left[0, \Sigma^*\right]$$

Proof: The result follows from Assumptions A3 and A4, and implicitly the Lindeberg–Feller CLT (Spanos, 1986, p. 177). □

Existence

The first step is to show that an interior solution to the GCE–NM problem exists for all sufficiently large sample sizes.

Proposition 2.2 *Under Assumptions A1–A3,*

$$\lim_T \Pr\left[\hat{\lambda}_T \in \Lambda\right] = 1$$

for all sufficiently large T.

Proof: By Lemma 3.1, the event

$$\left(\frac{X'y}{T}\right) \in \mathcal{Y}_\Lambda$$

is equivalent to the event of interest, $\hat{\lambda}_T \in \Lambda$, in sufficiently large samples. Assumption A1 and Lemma 3.2 imply

$$\begin{aligned} \frac{X'y}{T} &= \left(\frac{X'X}{T}\right)\beta + \left(\frac{X'e}{T}\right) \\ &\xrightarrow{p} Q\beta \end{aligned} \tag{2.59}$$

by Slutsky's Theorem. Definitions D1 and D8 further provide

$$Q\beta \equiv QZp(\lambda_0) \in \mathcal{Y}_\Lambda \tag{2.60}$$

Clearly,

$$\lim_T \Pr\left[\frac{X'y}{T} \in \mathcal{Y}_\Lambda\right] = 1 \tag{2.61}$$

and the equivalence of the events proves the proposition. □

Consistency

Proposition 2.3 *Under Assumptions A1–A3,*

$$plim(\hat{\beta}_T) = \beta$$

Proof: First, show that $\hat{\lambda}_T \xrightarrow{p} \lambda_0$. To do this, evaluate the three sets of terms in the GCE–NM dual objective function, Equation (2.51). By Lemma 3.2,

$$\left(\frac{y'X}{T}\right)\lambda \xrightarrow{p} \beta' Q'\lambda \tag{2.62}$$

By Assumption A2 and Slutsky's Theorem,

$$\begin{aligned} \Omega_k(\lambda) &= \sum_n q_{kn} \exp\left[Z_{kn}\left(\frac{X_k'X}{T}\right)\lambda\right] \\ &\to \sum_n q_{kn} \exp\left[Z_{kn} Q_k'\lambda\right] \end{aligned} \tag{2.63}$$

Given $V = O(\sqrt[-1]{T})$, Slutsky's Theorem implies

$$\begin{aligned} \Psi_k(\lambda) &= \sum_j u_{kj} \exp\left[V_k'\lambda\right] \\ &\to \sum_j u_{kj} \exp\left[0\right] = 1 \end{aligned} \tag{2.64}$$

By additional application of Slutsky's Theorem, the three results may be combined to yield

$$\begin{aligned} M_T(\lambda) &\xrightarrow{p} \beta' Q\lambda - \sum_k \log\left[\sum_m q_{km} \exp\left(Z_{km} Q_k'\lambda\right)\right] - \sum_k \log[1] \\ &= \beta' Q\lambda - \sum_k \log\left[\sum_m q_{km} \exp\left(Z_{km} Q_k'\lambda\right)\right] \\ &\equiv M_\infty(\lambda) \end{aligned} \tag{2.65}$$

The limiting objective function, $M_\infty(\lambda)$, is non–stochastic, strictly concave in λ, and has gradient

$$\nabla_\lambda M(\lambda) = Q\beta - QZp(\lambda) \tag{2.66}$$

By D1, $\lambda_0 \in \Lambda$ is the unique solution to the FOC for the limiting objective function. Further,

$$\lim_T \Pr\left[\lambda_0 \in \partial(\bar{\Lambda})\right] = 0 \tag{2.67}$$

by the existence result.

Given that $M_T(\lambda)$ is uniformly continuous in λ for all T, $M_T(\lambda) \xrightarrow{p} M_\infty(\lambda)$ necessarily implies $\hat{\lambda}_T \xrightarrow{p} \lambda_0$. Further, $Zp(\lambda)$ is a continuous function of λ, and Slutsky's Theorem implies that $Zp(\hat{\lambda}_T) \xrightarrow{p} Zp(\lambda_0) \equiv \beta$. Thus, $\widehat{\beta}_T \xrightarrow{p} \beta$. □

Asymptotic Normality

The asymptotic distribution of the GCE–NM solution can be derived by finding the distribution of $\hat{\lambda}_T$. Given that $\widehat{\beta}_T = Zp(\hat{\lambda}_T)$ is a continuous function of $\hat{\lambda}_T$, the δ–method (Spanos, 1986, p. 201) may be used to approximate the distribution of $\widehat{\beta}_T$. For example, suppose x is a random vector such that $x \sim N[0, \Sigma]$. For any continuous function, $h(\cdot)$, the distribution of $h(x)$ is $N[0, \bigtriangledown h(x)\Sigma \bigtriangledown h'(x)]$ where $\bigtriangledown h(x)$ is the Jacobian matrix of $h(\cdot)$ with respect to x.

Theorem 4.1.3 in Amemiya (1985) was modified to state and prove the following proposition.

Proposition 2.4 *Under Assumptions A1–A4,*

$$\sqrt{T}(\widehat{\beta}_T - \beta) \Rightarrow N\left[0, Q^{-1}\Sigma^* Q^{-1}\right]$$

if the GCE–NM error bounds are $V = O(T^{-1})$.

Proof: If $\hat{\lambda}_T \in \Lambda$, the first-order Taylor expansion of the FOC is

$$\bigtriangledown_\lambda M_T(\hat{\lambda}_T) = \bigtriangledown_\lambda M_T(\lambda_0) + \bigtriangledown_{\lambda\lambda'} M_T(\lambda^*)(\hat{\lambda}_T - \lambda_0) \tag{2.68}$$

for some λ^* between $\hat{\lambda}_T$ and λ_0 by the Mean Value Theorem. The lefthand side (LHS) is 0 by D5, and the approximation may be rewritten as

$$\sqrt{T}(\hat{\lambda}_T - \lambda_0) = -\left[\bigtriangledown_{\lambda\lambda'} M(\lambda^*)\right]^{-1} \cdot \left[\sqrt{T}\bigtriangledown_\lambda M(\lambda_0)\right] \tag{2.69}$$

Note that the inverse of the Hessian matrix exists because it is a positive definite matrix for all T and λ^* by the previous discussion.

To evaluate this expression, note that

$$\begin{aligned} \sqrt{T}\bigtriangledown_\lambda M_T(\lambda_0) &= \sqrt{T}\left[\frac{X'y}{T} - \left(\frac{X'X}{T}\right) Zp(\lambda_0) - Vw(\lambda_0)\right] \\ &= \sqrt{T}\left[\frac{X'(y - X\beta)}{T} + O(T^{-1})\right] \\ &= \left(\frac{X'e}{\sqrt{T}}\right) + O(\sqrt[-1]{T}) \end{aligned} \tag{2.70}$$

which converges in law to $N[0, \Sigma^*]$ by Assumption A4, Lemma 3.3, and Slutsky's Theorem. Further, $\text{plim}(\hat{\lambda}) = \lambda_0$ implies $\text{plim}(\lambda^*) = \lambda_0$, which provides

$$\begin{aligned} \text{plim} \bigtriangledown_{\lambda\lambda'} M_T(\lambda^*) &= \lim_T \bigtriangledown_{\lambda\lambda'} M(\lambda_0) \\ &= \lim_T \left[\left(\frac{X'X}{T} \right) \Sigma_Z(\lambda_0) \left(\frac{X'X}{T} \right) + \Sigma_V \right] \\ &= Q \Sigma_Z Q' \end{aligned} \tag{2.71}$$

because $V = O(T^{-1})$. The limiting distribution for $\hat{\lambda}_T$ is

$$\begin{aligned} \sqrt{T}(\hat{\lambda}_T - \lambda_0) &\Rightarrow N\left[0, (Q\Sigma_Z Q')^{-1} \Sigma^* (Q\Sigma_Z Q')^{-1}\right] \\ &\equiv N[0, \Sigma_{\lambda_0}] \end{aligned} \tag{2.72}$$

Now, the continuity of $Zp(\lambda)$ and the δ-method may be used. For $\beta \equiv Zp(\lambda_0)$,

$$\sqrt{T}(Zp(\hat{\lambda}_T) - \beta) \Rightarrow N[0, \bigtriangledown_{\lambda'} Zp(\lambda) \Sigma_{\lambda_0} \bigtriangledown_{\lambda} Zp(\lambda)] \tag{2.73}$$

By the GCE–NM version of Equation (2.39),

$$\bigtriangledown_{\lambda'} Zp(\lambda) = Q\Sigma_Z \tag{2.74}$$

which yields the desired result by substitution. □

Specifying $V = O(T^{-1})$ only affects the existence of an interior solution for a finite sample, and does not change the existence or consistency results as $T \to \infty$. If the rate of convergence were not changed, the remainder in Equation (2.70) would be $O(\sqrt[-1]{T})$, and multiplying by $\sqrt{T}$ yields a remainder that is $O(1)$. Thus, the asymptotic bias in $\hat{\lambda}_T$ would be

$$-(Q\Sigma_Z Q)^{-1} Vw(\lambda_0) \tag{2.75}$$

where $Vw(\lambda_0)$ is a vector of constants.

A Note on the Pure GCE–NM Problem

The GCE–NM solution is asymptotically equivalent to the LS or normal ML estimators because the FOC are equivalent in the limit of T — the limiting model

constraints are identical to the limiting normal equations. In the pure formulation of the GCE–NM problem, $V = 0$ for all T and the model constraint is

$$\frac{X'y}{T} = \left(\frac{X'X}{T}\right) Zp(\lambda) \tag{2.76}$$

If $\widehat{\beta}_{LS} \in \mathcal{Z}$ for a particular T, $\hat{\lambda}_T \in \Lambda$ and $\widehat{\beta}_T = \widehat{\beta}_{LS}$. Otherwise, there is no interior solution, and $\hat{\lambda}_T \in \partial(\bar{\Lambda})$. The existence result implies that the pure solution, $\hat{\lambda}_T$, will exist for sufficiently large samples, which are at least as large as those required for the noise formulation. Consequently, the preceding results also hold for the pure GCE–NM problem, but the probability of having an interior solution to the noise problem is larger for a given T.

2.4.2 Small–sample Behavior

Although the large–sample properties of the GCE solution are helpful, the entropy formalisms were were developed to solve ill–posed problems. As well, the generalized entropy approach is motivated by inverse problems with limited data and prior information. Consequently, the finite–sample properties of the GME–GCE solutions are of greater concern and are discussed in the present section.

Impact of Error Bounds, V

The noise terms, Vw, effectively 'loosen' the model constraints for a given set of observations, and an interior solution is more likely. If we view the GCE objective as a directed divergence function, wider error bounds provide a posterior that is 'closer' to the prior distribution. The width of the error bounds affects the amount of shrinkage toward the prior, and the degree of shrinkage may be measured in the Lagrange multipliers.

The shrinkage property may be demonstrated for a special subset of the generic GCE problems. For each disturbance, let the support be symmetric about 0 and limit the number of support points to $J = 2$. In this case, e_t may be written as

$$\begin{aligned} e_t &= \frac{v_t\left[\exp(-v_t\lambda_t) - \exp(v_t\lambda_t)\right]}{\exp(v_t\lambda_t) + \exp(-v_t\lambda_t)} \\ &= v_t \tanh(-v_t\lambda_t) \end{aligned} \tag{2.77}$$

where $v_t > 0$ is the scalar bound on e_t.

The following proposition describes the general impact of changes in the vector of error bounds, $v \in \Re^T$, on the optimal Lagrange multipliers, $\hat{\lambda}$.

Proposition 2.5 *For the noise specification in Equation (2.77),*

$$\nabla_{v'}\hat{\lambda} = D \cdot [\Gamma\Sigma_Z\Gamma + \Sigma_V]^{-1}$$

where D is a diagonal matrix with elements

$$D_{tt} = \tanh(-v_t\hat{\lambda}) - v_t\hat{\lambda}\left[1 - \tanh^2(-v_t\hat{\lambda})\right]$$

Proof: The proposition is a comparative statics result obtained by taking the total differential of the FOC for the dual version of the generic GCE problem,

$$\alpha - \Gamma Z p(\hat{\lambda}) - V w(\hat{\lambda}) = 0 \tag{2.78}$$

which has total differential

$$[\text{SOC}]\, d\hat{\lambda} + \left[\nabla_{v'} V w(\hat{\lambda})\right] dv = 0 \tag{2.79}$$

The first bracketed term is simply the Hessian matrix for the SOC, Equation (2.39). The second term will be a diagonal matrix with elements

$$\frac{de_t}{dv_t} = \tanh(-v_t\hat{\lambda}) - v_t\hat{\lambda}\left[1 - \tanh^2(-v_t\hat{\lambda})\right] \tag{2.80}$$

by the properties of the $\tanh(\cdot)$ operator. Rearranging the arguments yields the proposed result. □

A special case arises when Γ is an orthogonal matrix. The Hessian matrix from the SOC is also diagonal, and the impact of v_t on $\hat{\lambda}_t$ is

$$\frac{d\hat{\lambda}_t}{dv_t} = \frac{\tanh(-v_t\hat{\lambda}_t) - v_t\hat{\lambda}_t\left[1 - \tanh^2(-v_t\hat{\lambda}_t)\right]}{\sum_k \sigma_{Z_k}^2 + \sigma_{V_t}^2} \tag{2.81}$$

which can be easily computed for a particular problem.

To illustrate the result, consider a simple noise formulation of Jaynes' dice problem from Chapter 1. Let the bounds on the noise term be $[-v, v]$ for some $v > 0$. Then, the GME probability of observing i on the next roll of the die is

$$\hat{p}_i = \frac{\exp(-X_i\hat{\lambda})}{\Omega(\hat{\lambda})} \tag{2.82}$$

with the associated error probability

$$\hat{w} = \frac{\exp(-v\hat{\lambda})}{\Psi(\hat{\lambda})} \tag{2.83}$$

where $\hat{\lambda}$ is the optimal Lagrange multiplier.

To compute the impact of a change in v on $\hat{\lambda}$, write the dual objective function as

$$\begin{aligned} M(\lambda) &= y\lambda + \log[\Omega(\lambda)] + \log[\Psi(\lambda)] \\ &= y\lambda + \log[\Omega(\lambda)] + \log[2 \cdot \cosh(-v\lambda)] \end{aligned} \tag{2.84}$$

by employing the definition of the hyperbolic cosine. The FOC for the unconstrained problem is

$$\nabla_\lambda M(\hat{\lambda}) = y - \sum_i X_i p_i(\hat{\lambda}) - v \cdot \tanh(-v\hat{\lambda}) = 0 \tag{2.85}$$

Now, take the total differential of the FOC

$$d\hat{\lambda}\{\nabla^2_\lambda M(\hat{\lambda})\} + dv\{\tanh(-v\hat{\lambda}) - v\lambda[1 - \tanh^2(-v\hat{\lambda})]\} = 0 \tag{2.86}$$

where $\nabla^2_\lambda M(\hat{\lambda}) > 0$ because $M(\lambda)$ is strictly convex. Finally, solve for the desired ratio

$$\frac{d\hat{\lambda}}{dv} = \frac{\tanh(-v\hat{\lambda}) - v\hat{\lambda}[1 - \tanh^2(-v\hat{\lambda})]}{\nabla^2_\lambda M(\hat{\lambda})} \tag{2.87}$$

To evaluate this term, note that $\tanh(\cdot)$ is an odd function and that $\tanh(x) \in [-1, 1]\ \forall x$. We find that

$$\frac{d\hat{\lambda}}{dv} \begin{cases} > 0 & \forall\ \hat{\lambda} < 0 \\ = 0 & \hat{\lambda} = 0 \\ < 0 & \forall\ \hat{\lambda} > 0 \end{cases} \tag{2.88}$$

As expected, 'widening' the error bound by increasing v reduces the absolute value of $\hat{\lambda}$. This action corresponds with a solution that is 'more uniform', or closer to the prior distribution. By considering the potential noise in the observed average, the posterior distribution is 'shrunk' toward the prior.

Finally, consider the impact of using infinitely wide error bounds. Intuitively, the noise term now 'swamps' the signal, and the GME solution is simply the prior distribution. To confirm the intution, recall the dual objective function

$$M(\lambda) = \alpha'\lambda + \sum_k \log\left[\Omega_k(\lambda)\right] + \sum_t \log\left[\Psi_t(\lambda)\right] \tag{2.89}$$

Note that for finite λ_t, the error partition functions are

$$\begin{aligned} \Psi_t(\lambda) &= 2 \cdot \cosh(v_t \lambda_t) \\ &\rightarrow \begin{cases} 0 \text{ if } \lambda_t = 0 \\ \infty \text{ o.w.} \end{cases} \end{aligned} \tag{2.90}$$

Clearly, the disturbance terms dominate the remainder of the objective, and $M(\lambda)$ takes on a minimum at $\lambda = 0$ as $v_t \rightarrow \infty$.

Approximate Distributions

The asymptotic normality property may be used to approximate the distribution of the GCE point estimate for finite samples. The basic idea is to use the δ–method to transform the asymptotic distribution of $\hat{\lambda}_T$ and form the asymptotic approximation.

Recall the limiting distribution of $\hat{\lambda}_T$

$$\sqrt{T}(\hat{\lambda}_T - \lambda_0) \Rightarrow N\left[0, \Sigma_{\lambda_0}\right] \tag{2.91}$$

and the asymptotic approximation for small samples is

$$\hat{\lambda}_T \sim N\left[\lambda_0, T^{-1}\Sigma_{\hat{\lambda}}\right] \tag{2.92}$$

where

$$\Sigma_{\hat{\lambda}} = A^{-1}B(A')^{-1} \tag{2.93}$$

$$A = \left(\frac{X'X}{T}\right)\Sigma_Z(\hat{\lambda}_T)\left(\frac{X'X}{T}\right) + \Sigma_V(\hat{\lambda}_T) \tag{2.94}$$

$$B = \left(\frac{X'\Sigma_e X}{T}\right) \tag{2.95}$$

Consequently,

$$\hat{\lambda}_T \sim N\left[\lambda_0, T^2 C^{-1} D (C')^{-1}\right] \tag{2.96}$$

$$C = X'X\Sigma_Z(\hat{\lambda}_T)X'X + \Sigma_V(\hat{\lambda}) \tag{2.97}$$

$$D = X'\Sigma_e X \tag{2.98}$$

Then, the distribution of $\widehat{\beta}_T$ may be approximated by the δ-method. The required Jacobian matrix is

$$\nabla_{\hat{\lambda}_T} Zp(\hat{\lambda}_T) = \Sigma_Z(\hat{\lambda}_T)\left(\frac{X'X}{T}\right) \tag{2.99}$$

which is pre- and post-multiplied about $\Sigma_{\hat{\lambda}_T}$ to yield

$$\widehat{\beta}_T \sim N\left[\beta, \Sigma_Z(\hat{\lambda}_T)(X'X)C^{-1}D(C')^{-1}(X'X)\Sigma_Z(\hat{\lambda}_T)\right] \tag{2.100}$$

If the GCE-NM problem is specified as a pure inverse problem, the Σ_V terms disappear. For an interior solution (i.e. Σ_Z is full-rank), the approximate variance-covariance matrix is

$$(X'X)^{-1}X'\Sigma_e X(X'X)^{-1} \tag{2.101}$$

which is identical to the variance of the LS estimator. If the noise specification is used, the presence of Σ_V in the inverted terms reduces the variance of the estimator. Following the discussion in the preceding section, variance reduction is a property of shrinkage rules.

A Note on Efficiency

The concept of efficient estimation is usually tied to a particular model, $F(e)$, or relative performance in a family of estimators (e.g. the Gauss-Markov result for linear unbiased rules). Consequently, efficiency considerations may be of little concern in inverse problem that are ill-posed or have otherwise limited information.

If efficiency is of concern, note that the limiting model constraints for the GCE-NM problem are equivalent to the limiting normal equations for the LS or normal ML problems, which are efficient if $\Sigma_e = \sigma^2 I_T$. Otherwise, the estimators may be

improved by transforming the data to a scalar–identity error distribution and using the GLS rule. The same concept also applies to the GCE–GME problems. Suppose there exists some matrix, P, such that $\text{Var}(Pe) = \sigma^2 I_T$. Then, the transformed GLM

$$Py = PXZp + PVw \tag{2.102}$$

may be used to form the GCE model constraints. The transformation matrix, P, may be recovered by inverting the Cholesky decomposition of Σ_e. If Σ_e is unknown, it may be estimated in a consistent fashion (as in the feasible GLS case) to retain the large–sample properties.

Chapter 3

Applications and Performance of Generalized Entropy in Cases of the General Linear Model

3.1 Introduction

The analytical results for the GCE solutions presented in Chapter 2 are useful, but limited to fairly restrictive model assumptions. Given that the GCE solution does not take a closed form, the performance of the entropy methods for particular problems is difficult to assess. The finite–sample distributions presented at the end of Chapter 2 are only approximate, and many of the traditional methods do not have well–developed small sample properties. For these reasons, Monte Carlo sampling experiments provide a useful basis for examining the properties of GCE relative to other methods of information recovery.

The purpose of the present chapter is to demonstrate the behavior of the GCE solutions using three sampling exercises. The first problem attempts to recover a bounded mean from a single observation. Two traditional estimators, restricted ML and Bayes under normality, are compared to GME, and the robustness of estimators is examined under alternate error distributions. Second, a vector of unknown, real–valued parameters must be recovered from an ill–conditioned inverse problem. The problem is regularized by imposing bounds on the parameters, and GME is compared to LS, RLS, and the ridge estimators. Finally, alternate GME specifications for non–i.i.d. errors are demonstrated under an AR(1) error process. Unless otherwise stated, each experiment is based on 5000 Monte Carlo trials.

3.2 Recovering a Bounded Mean

The basic properties of the generalized entropy solutions may be demonstrated using a simple problem. Consider a single observation, $x = \beta + e$, where e and $\beta \in [-m, m]$ are unknown. The problem of recovering an image of β from x is a familiar linear inverse problem in statistics.

3.2.1 Normal Errors

A common assumption is $e \sim N[0,1]$, and the sampling theory estimator is the restricted ML rule

$$\widehat{\beta}_{ML} = \begin{cases} -m & \text{if } x < -m \\ x & \text{if } x \in [-m, m] \\ m & \text{if } x > m \end{cases} \tag{3.1}$$

The ML estimator may also be derived as the restricted LS estimator, which does not require the normality assumption.

Alternately, consider a prior distribution with equal mass on points $-m$ and m. Casella and Strawdermann (1981) show that the Bayesian posterior mean under squared-error loss (SEL) is

$$\widehat{\beta}_B = m \tanh(mx) \tag{3.2}$$

Motivated by an earlier result from Ghosh (1964), they also show that $\widehat{\beta}_B$ is minimax if $m < 1.06$. For larger values of m, the minimax property may be extended by increasing the number of elements in the support of β. Bickel (1981) examines the minimax character of the problem when $\beta \in \Re^K$.

As in the LS case, the generalized entropy approach does not require the normality assumption. Using the Bayes support, $\beta \in \{-m, m\}$, the GME model constraint may be written as

$$\begin{aligned} x = &= Zp + Vw \\ &= -mp + m(1-p) - vw + v(1-w) \end{aligned} \tag{3.3}$$

where $v \geq 0$ is the error bound. Taken as a special case of the generic GCE solution from Chapter 2, the GME probability is

$$\hat{p} = \frac{\exp(m\hat{\lambda})}{\exp(m\hat{\lambda}) + \exp(-m\hat{\lambda})} \tag{3.4}$$

where $\hat{\lambda}$ is the optimal Lagrange multiplier on the model constraint. The GME

posterior mean is

$$\widehat{\beta}_{\text{GME}} = \frac{-m\exp(m\hat{\lambda}) + m\exp(-m\hat{\lambda})}{\exp(m\hat{\lambda}) + \exp(-m\hat{\lambda})} = m\tanh(-m\hat{\lambda}) \tag{3.5}$$

Clearly, the GME and Bayes solutions are related through their common functional form. The estimates are equal if $\hat{\lambda} = -x$, which only occurs when $x = 0$. If the GME problem is treated as a pure inverse problem (i.e. $v = 0$), the GME solution is the posterior distribution on $-m$ and m with a mean of x. For $x \in (-m, m)$, the pure GME solution is simply x. If the violated boundary is used in case of an infeasible solution, the pure GME solution is identical to $\widehat{\beta}_{ML}$. Thus, the ML approach is a special case of generalized entropy. For $v > 0$, the results from the small-sample section in Chapter 2 imply that the optimal Lagrange multiplier is smaller in absolute value than $\hat{\lambda}$ for the pure inverse problem. The posterior distribution will be closer to the prior (i.e. more uniform), and the solution to the noise formulation is a form of shrinkage estimator.

To compare the competing rules, let $m = 1$ so that $\widehat{\beta}_B$ is minimax. For the GME with noise formulation, specify $v = 3$ according to the 3-σ rule. Also, infeasible solutions will be evaluated at the violated bound. The risk functions of the three estimators were recovered as the mean SEL (MSEL), $\|\widehat{\beta} - \beta\|^2$, and the results for $\beta \in [0, 1]$ are plotted in Figure 3.1. The risk functions for the ML and Bayes estimators are nearly identical to those presented in Figure 2 of the Casella and Strawdermann article, and the Bayes estimator risk-dominates the ML solution. The GME solution risk-dominates the Bayes solution for most of the parameter space, and only surrenders its advantage for very large values of β.

Again, the analysis of $d\hat{\lambda}/dv$ in Chapter 2 implies that increasing v shrinks the pure GME solution away from the sample and toward the prior mean. In this example, the prior mean is 0, and the pure or sample solution is $\widehat{\beta}_{ML}$. Conceptually, each $v > 0$ corresponds to some $\phi(x) \in [0, 1]$ such that $\widehat{\beta}_{GME} = \phi(x)\widehat{\beta}_{ML}$. In other inferential settings, several methods have been devised for choosing the generic shrinkage factor, $\phi(x)$, and most methods depend on the underlying signal-noise ratio for the model.

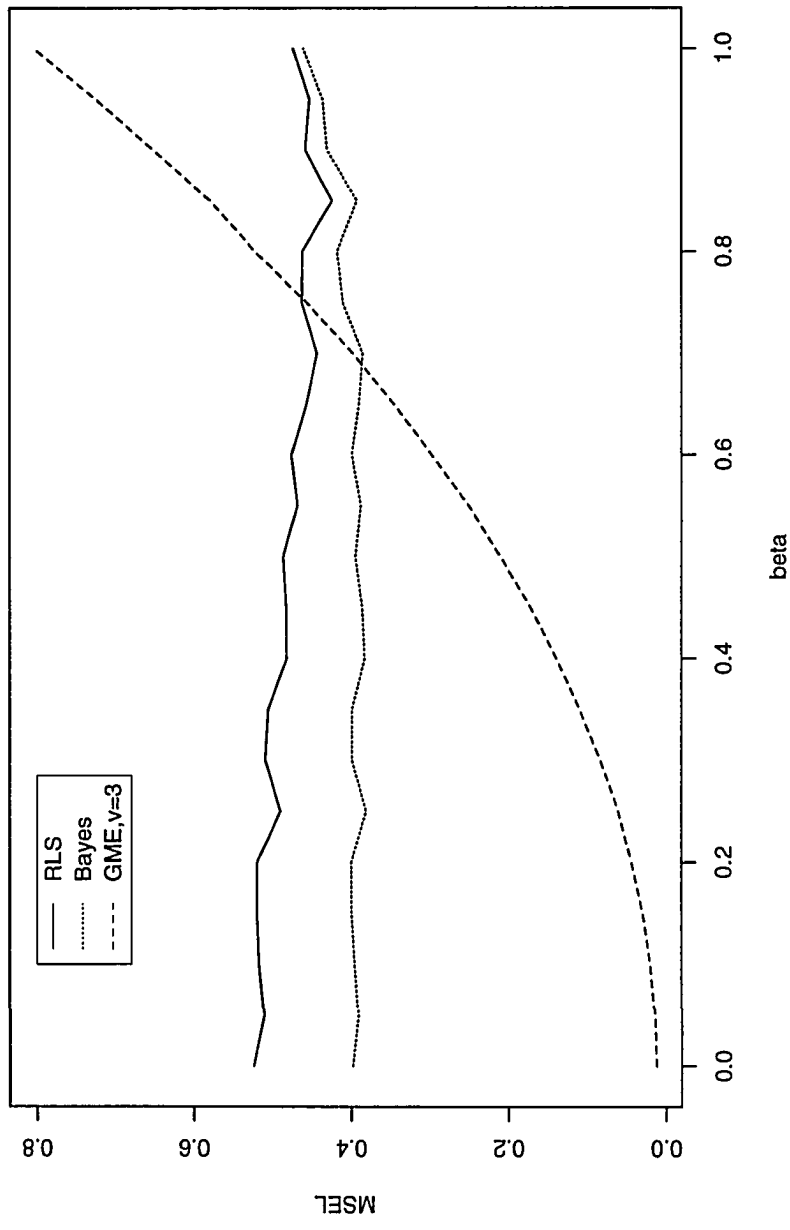

Figure 3.1: Empirical Risk of Bounded Normal Mean Estimators

However, the GME approach is relatively easy to implement because v is simply viewed as an error bound, and information about the signal–noise ratio is directly employed.

In the present example, an error bound of $v = 3$ allows for a feasible solution if $x \in (-4, 4)$. The probability of a boundary solution is nearly zero, even if $|\beta| = 1$, and the probability that $\phi(x) = 1 \; \forall \; x$ is nearly 0. As v decreases, $\Pr[\phi = 1 \; \forall \; x] \rightarrow 1$, and the GME solution will behave more like the restricted sampling estimator, $\widehat{\beta}_{ML}$. For $v \in \{0.5, 1, 3\}$, the empirical risk functions of the GME and Bayes solutions are presented in Figure 3.2. As expected, the wider error bounds allow for a GME solution that is 'closer' to the prior distribution. As v decreases, the GME risk becomes 'flatter' as it uses less of the prior information. Note that for $v = 0.5$, the GME solution behaves very much like the Bayes estimator.

Alternately, the performance of the GME solution may be explained by considering the minimum distance interpretation of the GME problem. The Bayesian posterior distribution for β is derived by combining the normal likelihood with the discrete, uniform prior under Bayes Rule. Without a likelihood function, GME solves for the posterior that is 'closest' to the prior and satisfies the model constraint, Equation (3.3). By using $v > 0$, the constraint is effectively loosened, and the posterior may be 'shrunk' closer to the prior. Consequently, GME will do very well when the prior information is correct, but the other estimators dominate GME when β is large. In this example, an error bound of $v = 0.5$ provides a GME posterior distribution that is nearly equivalent to using Bayes Rule and a normal likelihood function to evaluate the sample information. Although it is not a true Bayesian method, GME may be informally viewed as a nonparametric (i.e. *sans* likelihood) Bayesian technique for recovering information about β.

3.2.2 Alternate Error Distributions

If the restricted ML approach is viewed as a restricted LS estimator, the Bayes estimator is the only rule based on a distributional assumption. To examine the robustness of the competing methods, the sampling experiments were repeated using

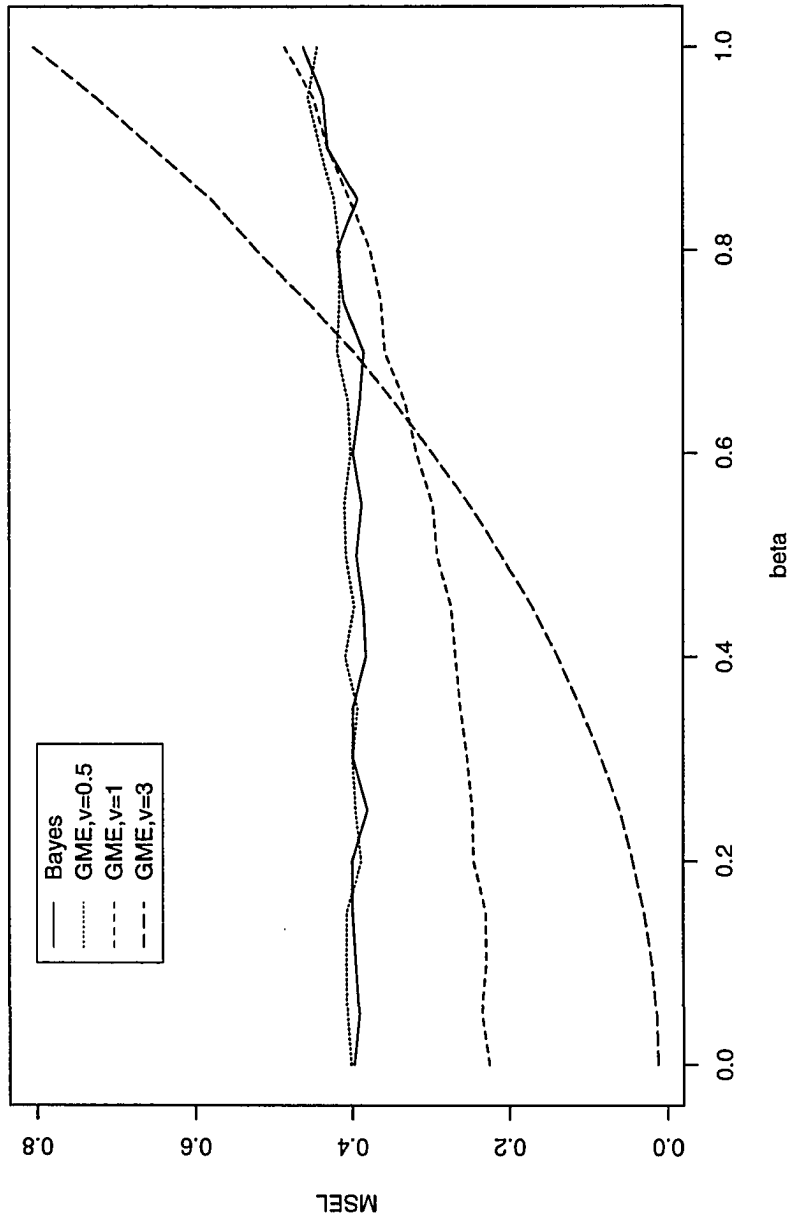

Figure 3.2: GME Risk under Various Error Bounds

two non–normal alternatives. First, the Student–t distribution with 3 degrees of freedom was used to evaluate performance under a heavy–tailed distribution. To maintain a unit variance, all of the drawings from the t(3) pseudo–random number generator were scaled. The risk functions of the restricted ML, Bayes, and GME estimators appear in Figure 3.3.

For $v = 3$, the risk functions for the ML and Bayes solutions maintain the same relationship, but the GME solution does not shrink the sample as strongly as before. Using $v = 6$ and $v = 9$, the risk functions for the GME solution were computed, and these appear in Figure 3.4. As before, decreasing v provides less shrinkage toward the prior, and the risk function is 'flatter' as it uses more of the sample information.

Another variation is to consider a skewed error distribution. The following results are based on errors drawn from a $\chi^2(4) - 4$ distribution, which has a mean of 0 due to mean–centering. As before, the errors were also scaled to have unit variance. By incorrectly assuming that the errors are standard normal, the GME error support is $v = 3$. Given that the error distribution is no longer symmetric, Figure 3.5 presents the risk functions for $\beta \in [-1, 1]$.

Again, the restricted ML and Bayes estimators maintain the same relative performance, although the risk functions take a different general shape. Further, the GME solution continues to dominate both ML and Bayes over much of the parameter space. Intuitively, the favorable performance of the GME solution follows from its reliance on sample and prior information. Although the ML result is also the restricted LS rule, which does not require normality, the shrinkage behavior of the GME rule improves the risk of the estimator.

Finally, suppose we know the disurbances are skewed, and use this information to shift the entropy error supports. Using an informative prior distribution of $u = [0.667, 0.333]$ on $V = [-\sqrt{2}, 2\sqrt{2}]$, the inverse problem may be solved under the GCE framework. The resulting risk function is also presented in Figure 3.5. In this case, the additional information used in the GCE problem uniformly improves the entropy risk.

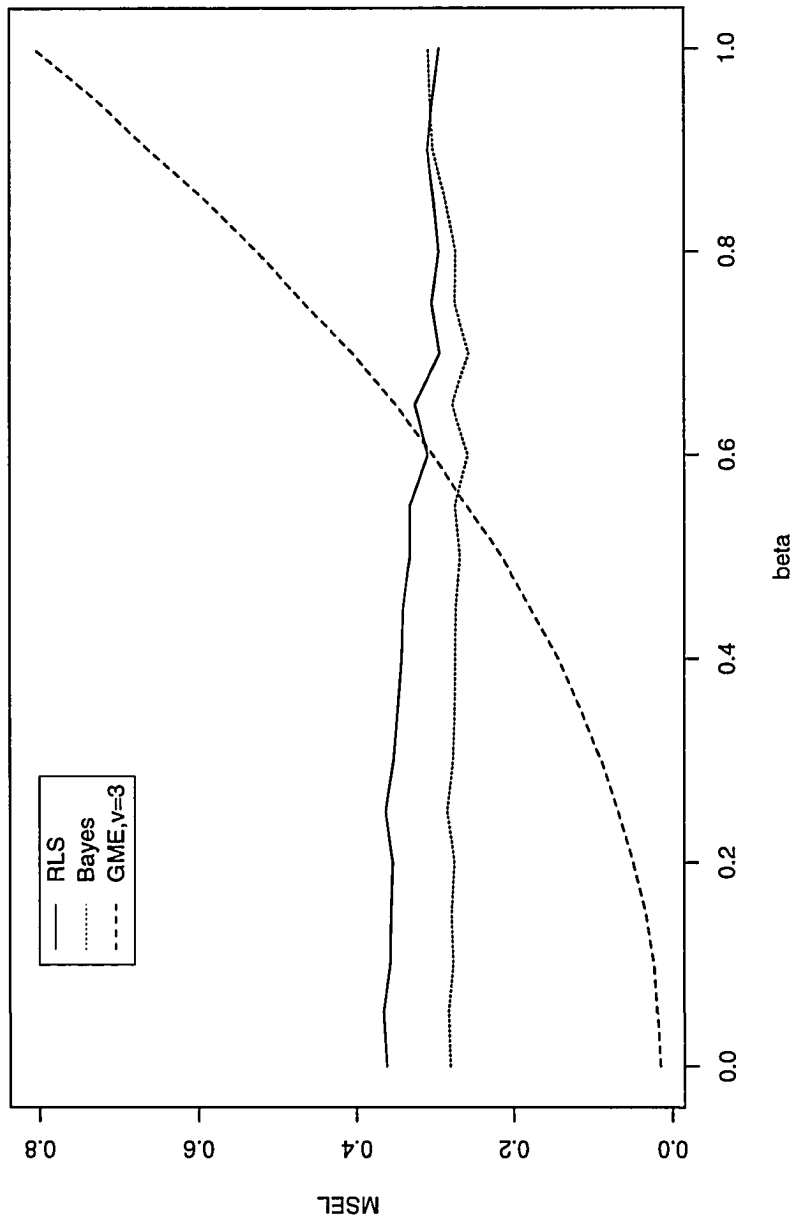

Figure 3.3: Empirical Risk of Bounded t(3) Mean Estimators

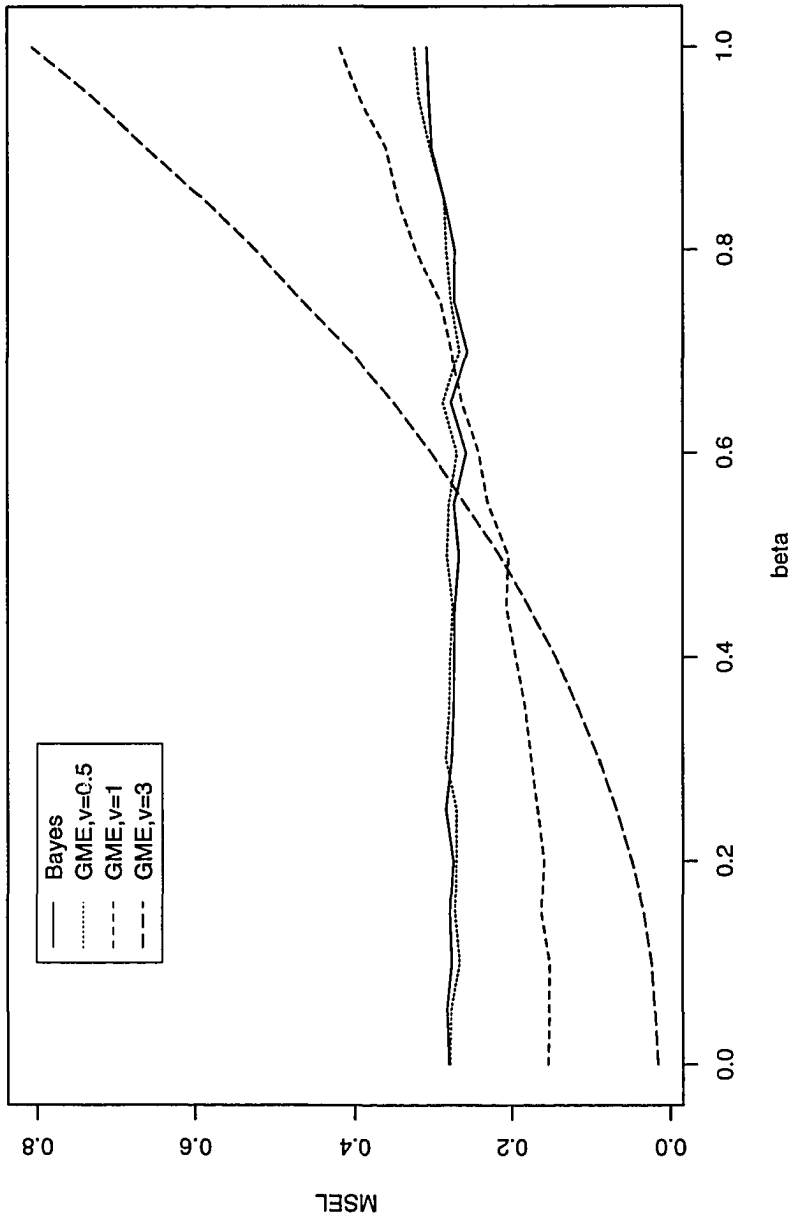

Figure 3.4: GME Risk under Various Error Bounds

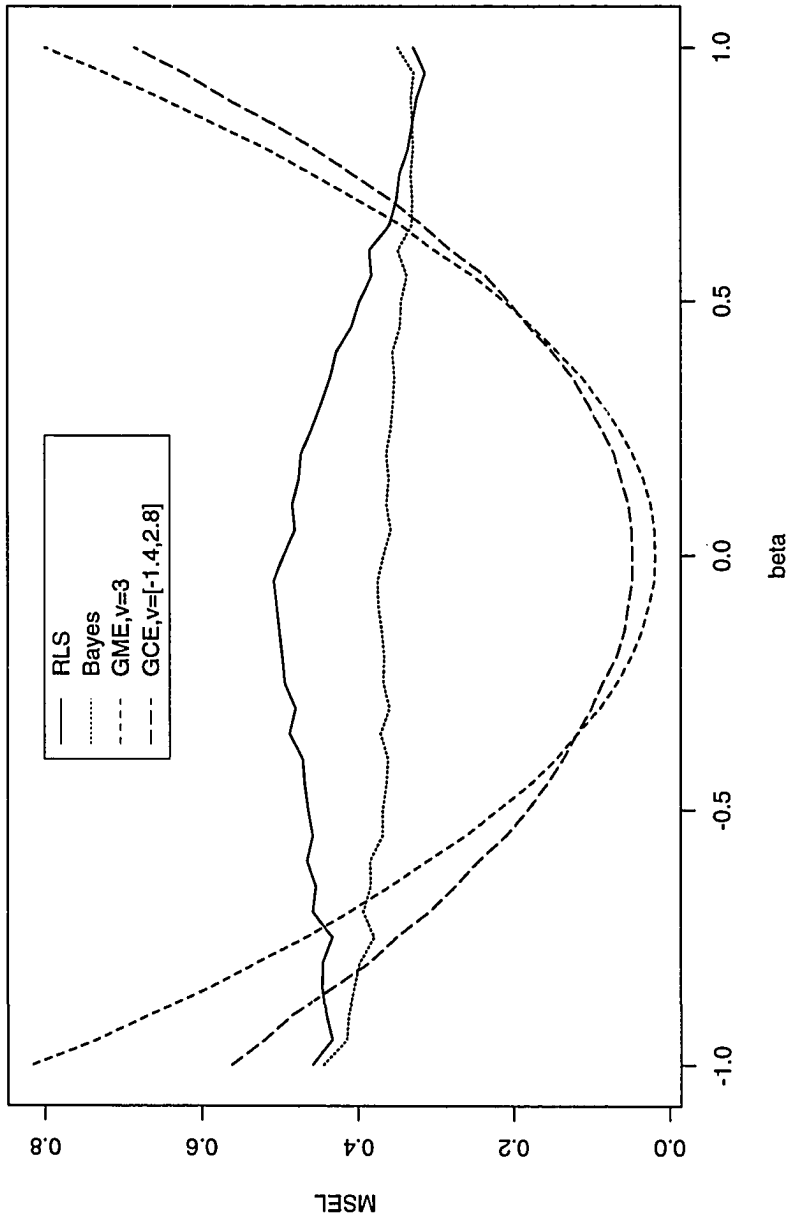

Figure 3.5: Empirical Risk of Bounded $\chi^2(4)$ Mean Estimators

3.3 An Ill–Conditioned Problem

As stated in Chapter 1, the data available for solving economic inverse problems are often limited to non–experimental observations. In the absence of an orthogonal experimental design, there may exist one or more exact or near–exact linear dependencies among the explanatory variables. Such dependencies are fairly common in problems like the demand example where the set of explanatory variables may include real prices of related goods, which tend to move together over time. Also, linear dependencies may be induced in finite samples by certain data transformations (e.g. linear spline models). In either case, X does not have full column rank or a numerically stable inverse matrix, and the problem is ill–conditioned or collinear.

3.3.1 Symptoms and Treatment

A common symptom of ill–conditioning in the GLM is unstable estimates of β, and small changes in the indirect observations result in large changes in the recovered image. The problem is easily demonstrated for the LS estimator, $\widehat{\beta}_{LS} = (X'X)^{-1}X'y$, which is an unbiased estimator of β. The LS estimator is also the ML estimator if $e \sim N[0, \Sigma_e]$ with known variance–covariance structure. If the problem is severely ill–conditioned, X does not have full column rank and the LS estimator is not uniquely defined.

In moderately ill–conditioned problems, $(X'X)^{-1}$ may exist, but $\widehat{\beta}_{LS}$ may have a very large elements in its variance–covariance matrix. Consider the singular value decomposition (SVD) of the design matrix, $X = QLR$, where Q is a $(T \times K)$ orthogonal matrix, L is a $(K \times K)$ diagonal matrix, and R is a $(K \times K)$ orthogonal matrix. The diagonal elements of L, $\pi_{(i)}$, are the singular values of X, and the columns of R are the associated eigenvectors of X. Using the spectral decomposition of $(X'X)^{-1}$, the variance matrix is

$$Var(\widehat{\beta}_{LS}) = \sigma^2(X'X)^{-1} = \sum_k \frac{\sigma^2}{\pi_k} R_k R_k' \tag{3.6}$$

if $\Sigma_e \equiv \sigma^2 I_T$. As the problem becomes more ill–conditioned, one or more of the $\{\pi_k\}$ approaches 0, and the variance of $\widehat{\beta}_{LS}$ increases.

Although the degree of collinearity present in a given design matrix may be measured in a variety of ways, the singular values are especially useful. Belsley (1991) recommends the condition number

$$\kappa(X'X) = \frac{\pi_{(1)}}{\pi_{(K)}} \tag{3.7}$$

which is the ratio of the largest and smallest singular values of X (with columns scaled to unit length). If the design matrix is orthogonal and the columns of X are linearly independent, $\pi_{(i)} = 1\ \forall i$ and $\kappa(X'X) = 1$. As the degree of collinearity increases, $\pi_{(K)} \to 0$ and $\kappa(X'X) \to \infty$. where Q is a $(T \times K)$ orthogonal matrix, L is a $(K \times K)$ diagonal matrix in which the i^{th} element is $\pi_{(i)}$, and R is a $(K \times K)$ orthogonal matrix.

Belsley notes that potentially harmful collinearity may arise if $\kappa(X'X)$ is as small as 25, but such cases are rarely encountered in practice. Although $\kappa(X'X)$ is only an ordinal measure of collinearity, Belsley recommends using $\kappa(X'X) > 900$ as a sign for the presence of potentially harmful collinearity. Given that $\kappa(X'X)$ only measures the most severe linear relationship in X, he further recommends examining the full set of K condition indices

$$\left\{\frac{\pi_{(i)}}{\pi_{(1)}}\right\}$$

for the presence of two or more potential harmful linear dependencies.

Of course, an erratic estimate of β is only recognized as such if it does not conform to our prior beliefs. In the demand example, a positive own–price elasticity is not very plausible, and a positive estimate may indicate the presence of significant collinearity (or a number of other problems with the data or the model). As in the ill–posed case, prior information may be used to regularize the ill–conditioned problem and reduce the variation in the estimates.

Linear inverse problems associated with the GLM may be regularized or augmented in variety of way. The parameter space may be restricted to subsets of $\Re^K$ under the restricted GMM (e.g. LS) or ML approaches. Using subjective probabilities, an informative prior distribution for β may be used in a Bayesian analysis. To reflect bounds on the unknown parameters, the Bayesian prior distributions may be

discrete (as in the bounded mean example). Alternately, the support of a Lebesgue prior density may be truncated, and the posterior distribution may be evaluated as shown by Geweke (1986). MOR techniques are another means of penalizing solutions that are inconsistent with the prior information.

A special case of MOR is quadratic regularization (QR) which chooses β to minimize

$$\mathcal{L} = \|y - X\beta\|^2 + \eta\beta' C\beta \tag{3.8}$$

Here, C is a positive definite matrix, $\beta' C\beta$ is the square of the weighted Euclidean norm of β, and η is a tuning or smoothing parameter used to establish the trade-off between the two criteria. In effect, we are penalizing solutions whose 'norm' exceeds some prior bound, and the problem is now well-posed. The penalized solution,

$$\beta_{QR} = (X'X + \eta C)^{-1} X'y \tag{3.9}$$

is commonly known as the *ridge regressor*. From the Bayesian perspective, β_{QR} is the posterior mean associated with a $N[X\beta, I_T]$ likelihood function and prior distribution $g(\beta) = N[0, (\eta C)^{-1}]$. Given Lebesgue prior and likelihood distributions, the variance–covariance structure is analogous to the MOR 'norm' restictions. In either the frequentist or Bayesian world, the researcher must still choose the smoothing or prior parameters to reflect their information for a particular problem.

Unfortunately, economists rarely have information about the 'norm' of their parameter vectors, and the choice of a likelihood function is more a matter of convenience than of using actual prior knowledge. The set of available information for economic inverse problems is often limited to signs or magnitudes of individual parameters (e.g. elasticities). The GME–GCE approach is a feasible alternative as it only requires modest assumptions about the error structure, does not require a likelihood function, and employs the available prior information to specify $\mathcal{Z}$.

3.3.2 A Sampling Experiment

To examine the relative performance of the GME–GCE techniques, the results from a series of Monte Carlo sampling experiments are presented. The experiments

involve the recovery of real-valued parameters from a linear model under an ill-conditioned design matrix, and the performance of the entropy-based methods is compared with the traditional methods for handling collinear problems: the least squares (LS), restricted least squares (RLS), and ridge estimators.

To form the signal for the MC experiment, a (10×4) design matrix was drawn from an i.i.d. $N(0,1)$ pseudo-random number generator. Then, the SVD of X was recovered, and the eigenvalues in L were replaced with the K-vector,

$$a = \left[\sqrt{\frac{2}{1+\mu}}, 1, 1, \sqrt{\frac{2\mu}{1+\mu}}\right] \tag{3.10}$$

which has length $K = 4$. The new design matrix, $X_a = QL_aR$, is characterized by $\kappa(X_a'X_a) = \mu$ so that the degree of collinearity may be specified *a priori*. For convenience, the columns of X_a are not scaled to unit length. For $\beta = [2, 1, -3, 2]'$, the mean vector of the dependent variables, $X_a\beta$, was formed, and T i.i.d. $N(0,1)$ pseudo-random errors, e, were added to form vector of noisy observations, y.

For each of 5000 MC trials, the LS, RLS, and ridge methods were used to estimate β from the relevant X_a and y. The RLS estimates were restricted to $\beta_k \in [-10, 10]$ for each k, and the Hoerl, Kennard and Baldwin (1975) iterative ridge estimator was used with $C \equiv I_K$ and

$$\hat{\eta} = \frac{\hat{\sigma}^2(K-2)}{\hat{\beta}'\hat{\beta}} \tag{3.11}$$

In addition, estimates of β were recovered by the GME-D method with parameter supports $Z_k = [-10, 10]$ for each k. For the error parameter space, the 3-σ rule was used, and $v_t = [-3, 3]$ was specified for each t.

To gauge the performance of the competing methods, the precision of the estimators was evaluated under squared error loss, SEL $= \|\beta - \tilde{\beta}\|^2$. The average SEL (MSEL) for the competing methods of information recovery are prsented in Figure 3.6. The horizontal axis of each plot is expressed in units of $\kappa(X'X)$ ranging from 1 to 100. Recall that $\kappa(X'X) > 900$ indicates a potentially harmful degree collinearity.

The empirical risk of LS is 4.02 when $\kappa(X'X) = 1$, which is very close to its theoretical value. The empirical risk of RLS is close to the LS risk in the nearly

orthogonal case, but it declines relative to LS as $\kappa(X'X)$ increases due to the variance inflation effect of collinearity. That is, the LS estimates are more likely to fall outside the restricted parameter space. Further, the iterative ridge estimator is biased, and its empirical risk is 4.4 when $\kappa(X'X) = 1$. As the degree of collinearity increases, the empirical risk functions cross and the ridge estimator risk–dominates LS. Again, the regularity conditions reduce the risk of the ridge estimator relative to LS.

In contrast, the empirical risk of GME is nearly invariant to the degree of ill-conditioning, and it dominates the traditional estimators over the range of $\kappa(X'X)$. The superior performance of the entropy-based method in the region of very slight collinearity was unexpected, although its superiority in the ill-conditioned area was anticipated.

One possible explanation of the entropy performance follows from the intuition behind the ridge estimator. The ridge estimator is often viewed as a variance-reduction technique that restricts the least squares solution to an ellipsoid in the parameter space,

$$\left\{x \in \Re^K : \|x\|^2 \leq \beta' C \beta\right\}$$

As the degree of collinearity increases, the restricted ellipsoid expands with the ridge parameter, η, and the average SEL will rise. In the GME case, the parameter vector is restricted to a fixed hypercube about the origin. Thus, the relevant information about β is introduced through the support space, Z, and the performance is not affected by the degree of collinearity. Although the same information is used to restrict the RLS solution, the risk of the GME–D rule is lower due to the additional shrinkage provided by the noise terms.

It is important to note that risk–consequences of the GME estimator are offset in other areas. For example, the average sum of squared errors (MSSE) may be used to measure the empirical prediction loss, and Figure 3.7 presents the MSSE for each of the estimators. As expected, the least squares rules dominate GME under this measure because the estimates are selected on this basis, and the restricted LS rules (RLS and ridge) have larger MSSE than LS.

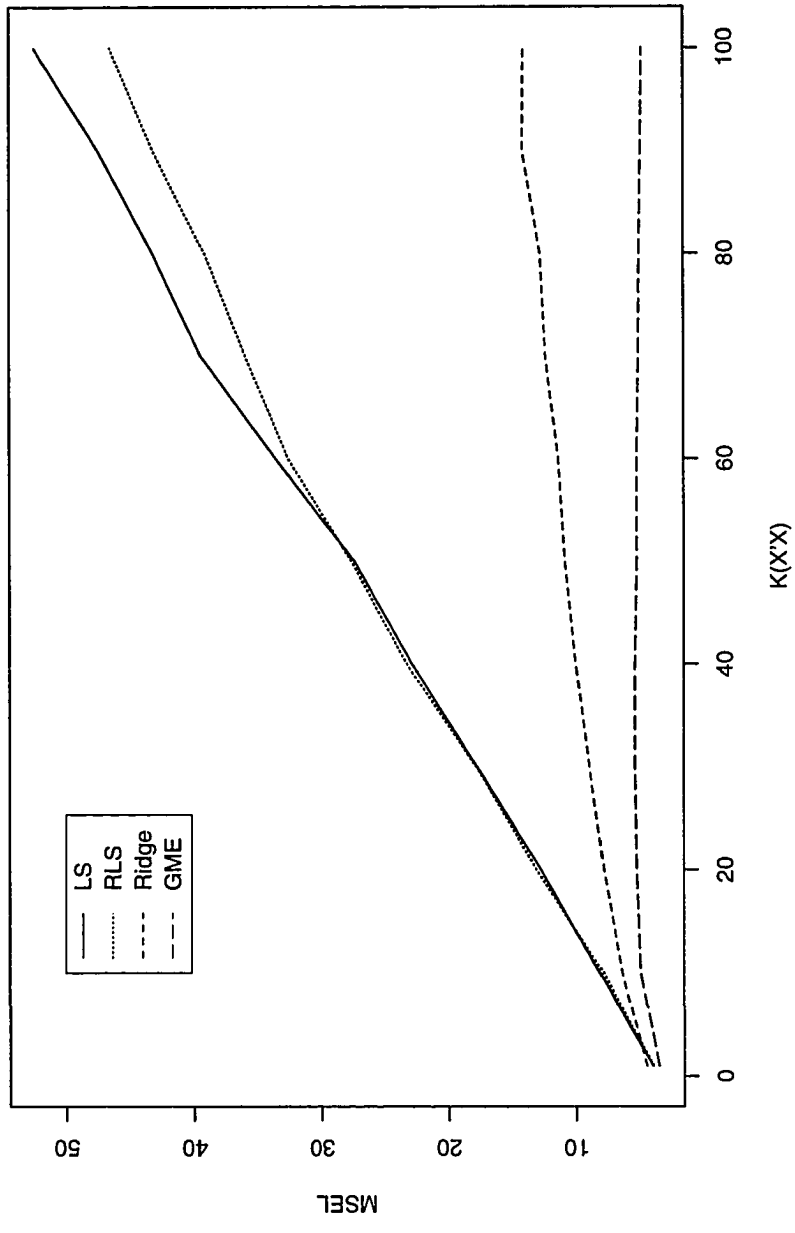

Figure 3.6: MSEL in Ill-Conditioned Problems

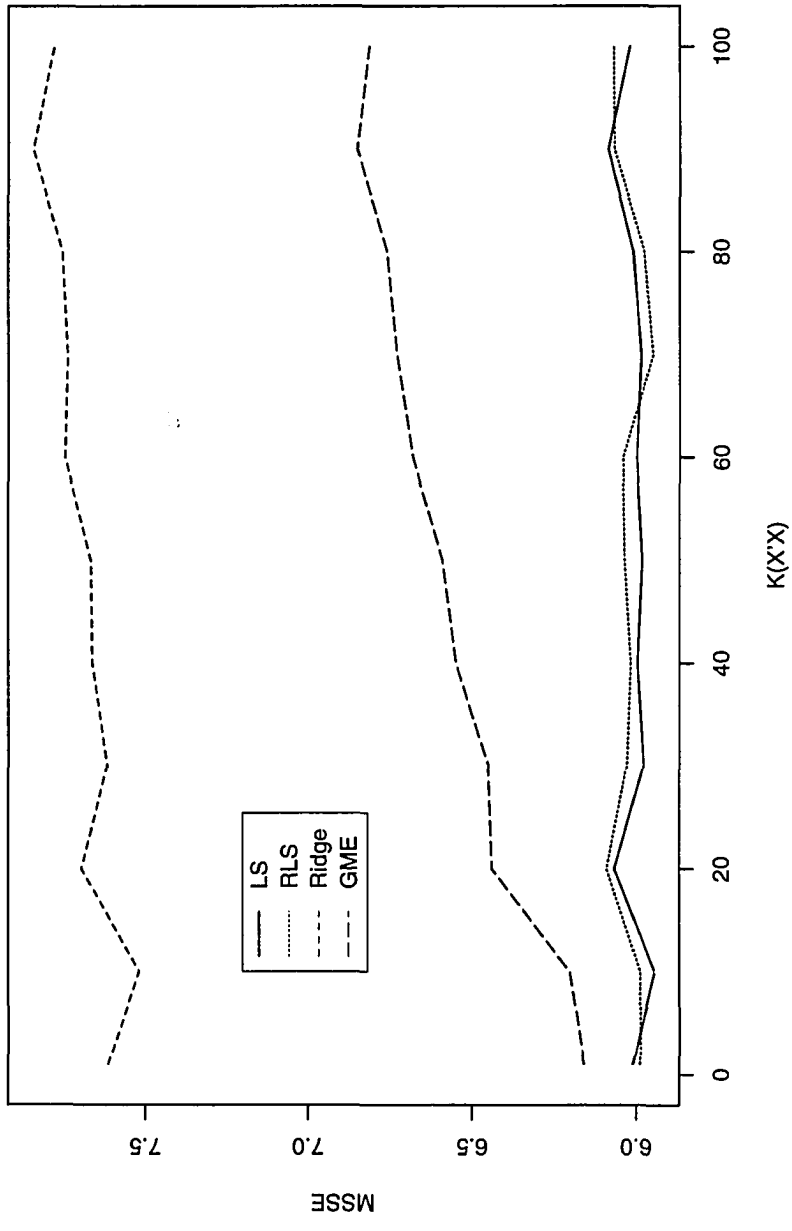

Figure 3.7: MSSE in Ill-Conditioned Problems

Empirical Distribution of β_3

Although the finite sample properties of the ridge and GME methods are unknown, especially in the collinear cases, the empirical distribution of the recovered parameters may be used to gather some information. Figures 3.8 and 3.9 present smoothed densities of the LS, ridge, and GME point estimates for β_3. The RLS estimates are excluded given their relation to the LS rule. The impact of ill–conditioning is represented by an orthogonal design matrix, $\kappa(X'X) = 1$, and a moderately collinear design, $\kappa(X'X) = 90$.

In the orthogonal case, the mean (-3.02) and variance (0.99) of the LS estimator are very close to their theoretical values. As expected, the data–based, iterative method of selecting the ridge parameter provides for some bias and variance–reduction relative to LS (mean = -2.54, variance = 0.96). However, the GME distribution is centered between the true value ($\beta_3 = -3$) and the center of its parameter support, 0. Consequently, the shrinkage property of the GME solution provides less bias (mean = -2.78) *and* less variance (0.83).

As the degree of collinearity increases, the LS distribution has greater variance (as expected) but remains centered over the true parameter value, $\beta_3 = -3$ (mean = -3.05, variance = 21.18). The ridge and GME distributions shift right as the degree of shrinkage (toward zero) increases, and their sample means are -1.82 and -2.17, respectively. Further, the ridge technique has a greater sample variance (4.77) than GME (1.33). Consequently, the entropy technique continues to reflect less bias than the ridge estimator, and it has smaller variance than either of the traditional methods in this sampling study.

3.3.3 Alternate Entropy Formulations

The restrictions imposed on the parameter space through Z reflect prior knowledge about the unknown parameters. However, such knowledge is not always certain, and researchers may want to entertain a variety of plausible bounds on β. The preceding sampling experiments were repeated using two additional Z matrices, $Z_k = [-5, 5]$ and $Z_k = [-15, 15]$ for each k. The empirical risk functions for these alternatives are

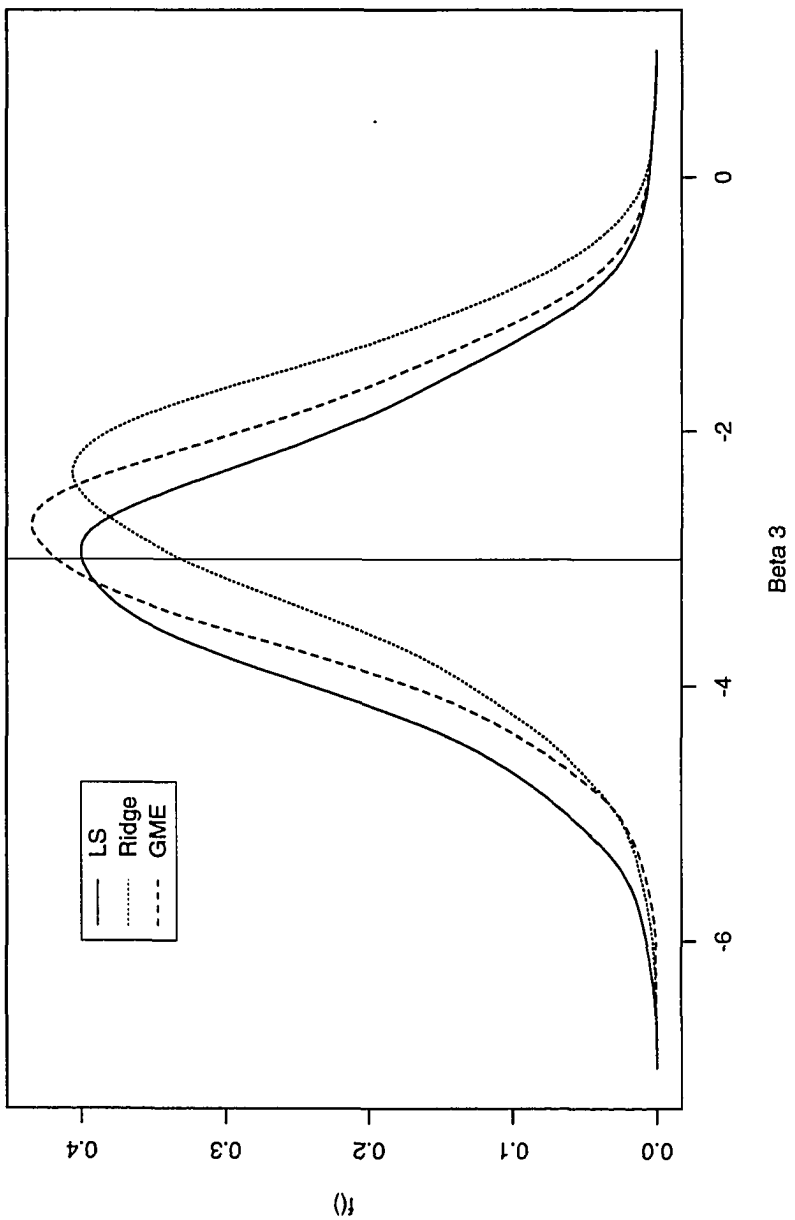

Figure 3.8: Empirical Distribution of β_3, $\kappa(X'X) = 1$

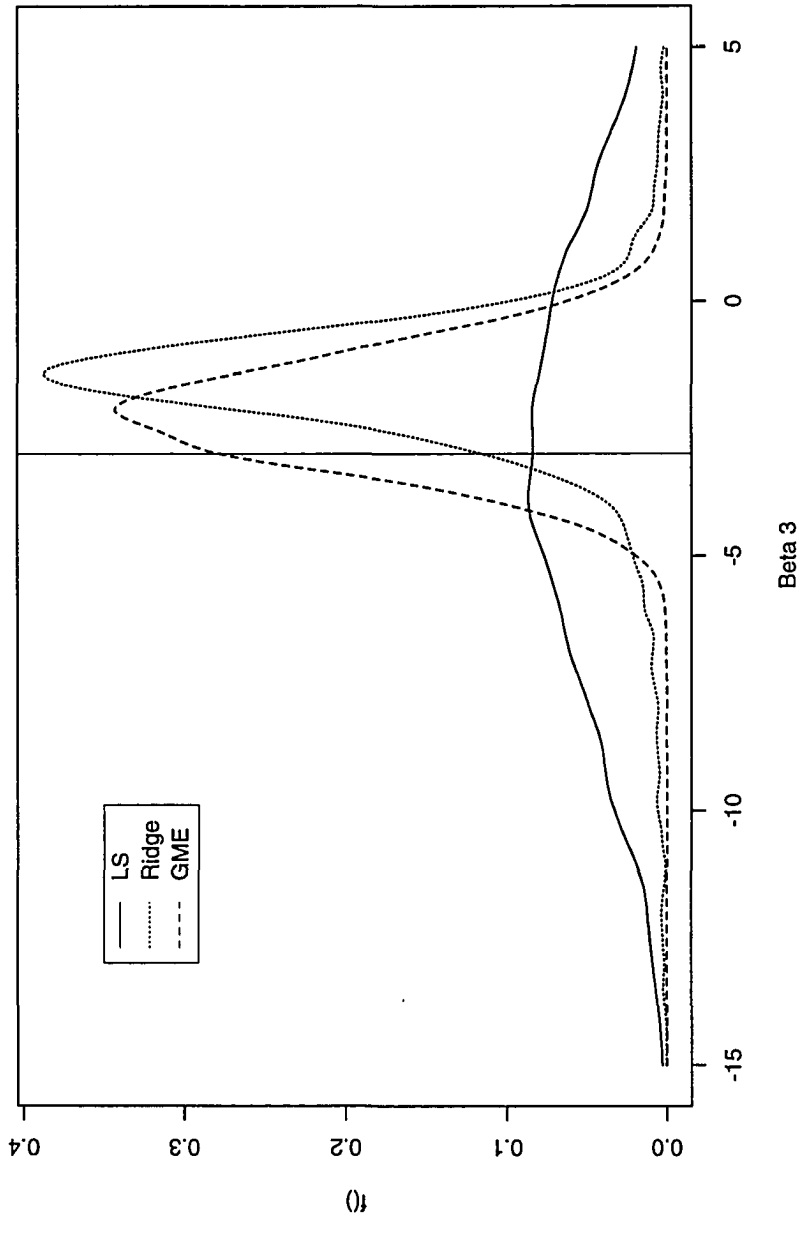

Figure 3.9: Empirical Distribution of β_3, $\kappa(X'X) = 90$

presented in Figure 3.10.

As the parameter supports are widened, the GME risk functions modestly shift upward reflecting the reduced constraints on the parameters space. Hence, wide bounds may be used without extreme risk consequences if our knowledge is minimal and we want to ensure that Z contains β. Intuitively, increasing the bounds increases the impact of the data and decreases the impact of the support. Of course, narrowing the parameter supports only improves risk as long as the true parameter vector is well in the interior of $\mathcal{Z}$. Although the results are not included here, the corresponding MSSE shifts in the opposite direction, highlighting the trade–off between the precision and prediction losses.

Finally, the cross–entropy criterion may be used to recover information about β given non–uniform prior distributions on Z. Using the GCE–D formulation, the Monte Carlo trials were repeated under two conflicting priors. First, the correct sign of the true parameters was included by specifying prior weights of $q_k = [0.375, 0.625]$ for $\beta_k > 0$, and the $q_3 = [0.625, 0.375]$ for β_3. Thus, the prior mean of each parameter is 2.5 in absolute value. Next, the prior weights were reversed so that the prior means are also 2.5 in absolute value, but with the wrong sign. The empirical risk functions appear in Figure 3.11.

As expected, including the correct prior signs improves risk for all values of $\kappa(X'X)$. However, the penalty for using the wrong prior information is not very large relative to the risk of the alternate estimators. The reason undoubtedly lies in the model constraint, which must be satisfied for any interior solution to the GCE–D problem. Although the incorrect prior weights affect the results, the entropy method cannot stray too far from the true parameters because it must also satisfy the sample information. This property is another benefit of the generalized entropy approach – incorrect prior information is effectively ignored if it does not agree with the sample.

3.3.4 Summary

Based on the limited evidence presented here, the GME–GCE framework is a feasible alternative to the standard methods of information recovery in ill–conditioned

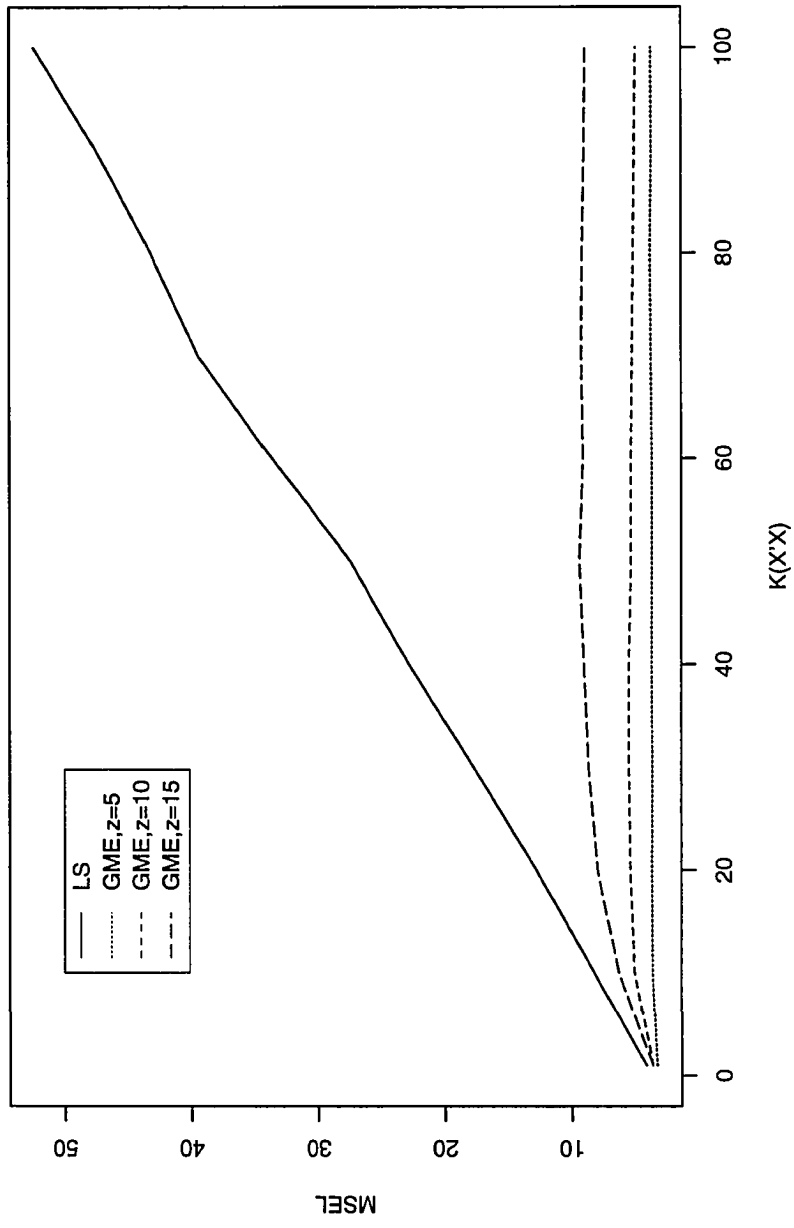

Figure 3.10: GME Risk under Alternate Z

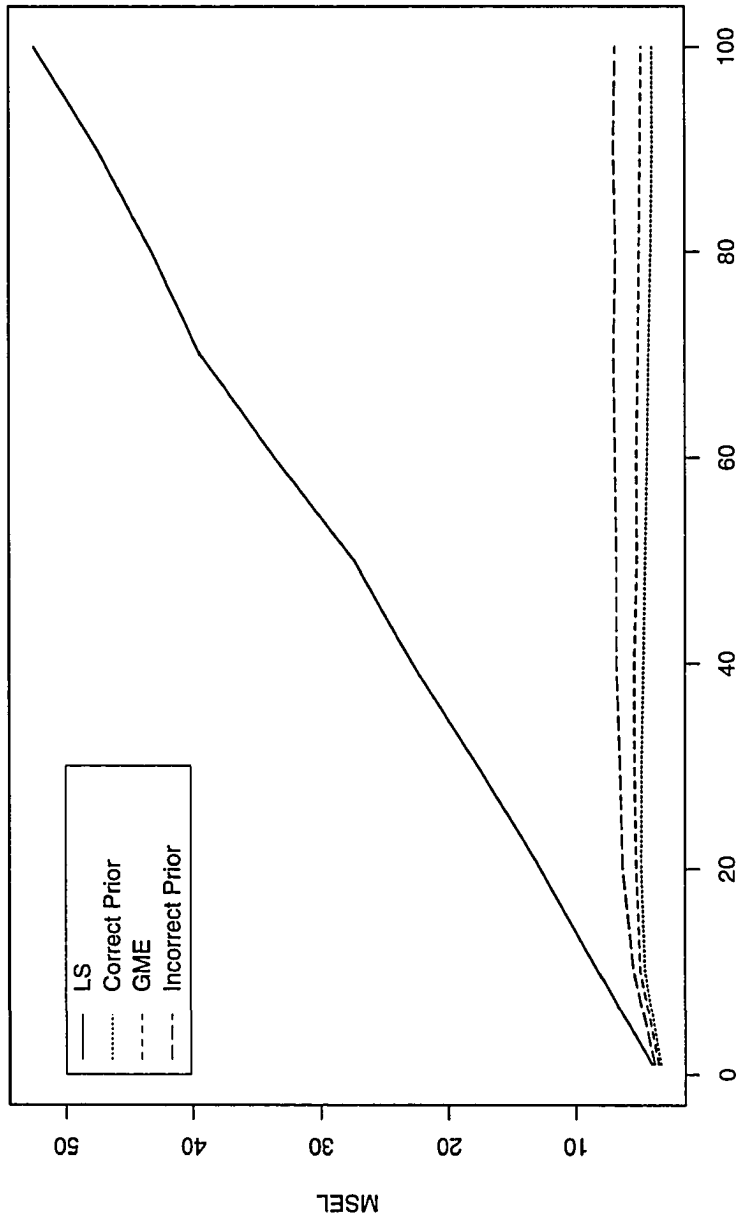

Figure 3.11: GCE Risk under Alternate Priors

linear inverse problems. Naturally, the traditional least squares techniques provide smaller average prediction losses, MSSE. However, the entropy–based methods are nearly invariant to collinearity, and they provide smaller average risk (MSEL) in moderately ill–conditioned problems. The trade–off between these two losses is clearly a subjective issue.

3.4 Dependent Error Structure

To this point, the GLM disturbances have been treated as i.i.d. random variables (i.e. $\Sigma_e = \sigma^2 I_T$). Consequently, the generalized entropy error bounds for each disturbance may be specified without regard for underlying dependencies or changes in magnitude or variation. In the more general case, Σ_e may be a diagonal matrix with non–stationary variances, σ_t^2, that reflect heteroskedastic behavior. Alternately, the errors may be correlated, and the off–diagonal elements of Σ_e will be non–zero. For example, heteroskedastic errors may appear in economic models due to changes in economic policies or varying amounts of market information. Correlated or autocorrelated errors may result from host of sources, including dependent economic actions, habitual behavior, lead–lag relationships, or biological cycles.

In general, the traditional methods of information recovery are still feasible estimators of β. However, the properties of the estimators may be improved by accounting for the error structure. For example, the LS estimator is no longer best unbiased if Σ_e is not a scalar–identity matrix, but the data may be transformed to derive a best unbiased estimator under the GLS framework. Hansen (1982) shows that GMM estimators may be asymptotically efficient if the objective norm is appropriately weighted (as in weighted LS).

The concepts of 'best' and 'efficient' are of secondary importance in the generalized entropy framework. In ill–posed problems, most linear estimators are infeasible, and the problem of finding the best member of the class is also ill–posed. Further, efficiency requires knowledge of the underlying distribution, $F(e)$, and this information is rarely available in practice. However, the GME–GCE problems may be formulated to include information about the variance–covariance structure of e.

To demonstrate the various GCE formulations for dependent error structures, consider a simple AR(1) version of the GLM

$$y = X\beta + e \tag{3.12}$$

$$e_t = \rho e_{t-1} + v_t \tag{3.13}$$

where $\rho \in (-1, 1)$ and v_t is a white noise disturbance with variance σ_v. If the error process is assumed to extend into the infinite past, it is well-known that Σ_e has elements

$$\sigma_{st} = \frac{\sigma^2 \rho^j}{1 - \rho^2} \tag{3.14}$$

where $j = |s - t|$.

As previously stated, the GME solution may be viewed as a shrinkage version of the LS estimator. By ignoring the correlated errors, GME will be a feasible, yet inferior rule. In fact, it is somewhat likely that the GME problem does not have an interior solution, especially if ρ is large or the error bounds are narrow relative to σ^2.

One approach to solving the AR(1) model is to use the GLS transformation to derive an i.i.d. error distribution. As discussed at the end of Chapter 2, there exists some matrix P such that $P'\Sigma_e P = I_T$. Then, $\text{Var}(P'e) = \sigma^2 I_T$, and we may proceed as we did in the i.i.d. case. If ρ and σ are known, the transformation matrix, P, may be recovered by inverting the Cholesky decomposition of Σ_e.

To evaluate this approach, consider the orthogonal design case of the model used in the preceding section, $\kappa(X'X) = 1$. For $\rho \in [0, 1]$, a set of standard normal errors were drawn and correlated according to Equation (3.13).[1] Given the resulting noisy signal, the Cholesky result was inverted and used to transform the data for each Monte Carlo trial and each ρ. Then, the GME–D solution was computed using $Z_k = [-10, 10]$ for each k and $V_t = [-3, 3]$ for each t. The risk functions for the LS, GLS, and GME estimators are presented in Figure 3.12.

As expected, GLS provides a vast improvement on LS as ρ increases. Also, the GME risk is lower than the GLS risk due to the shrinkage property of the entropy

[1] To represent the infinite history of the series, the first error was drawn from a normal distribution with variance $(1 - \rho^2)^{-1}$.

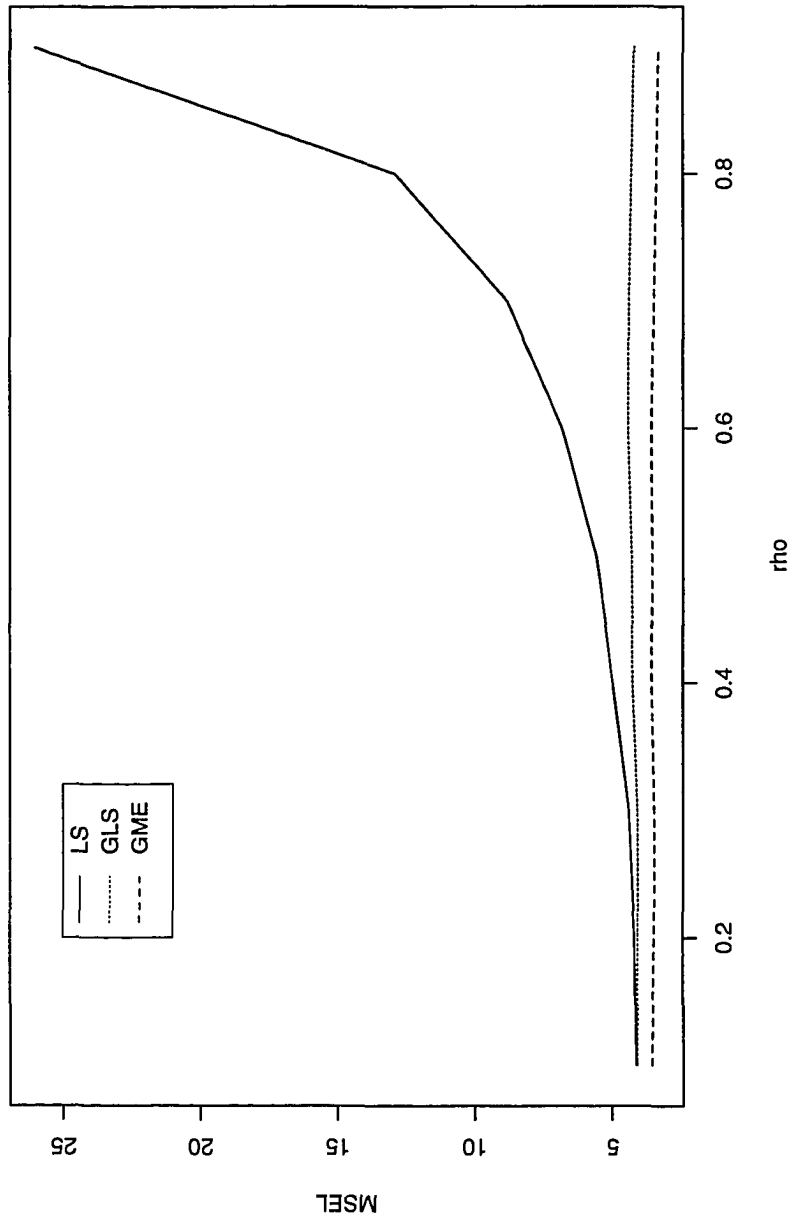

Figure 3.12: GME Risk under GLS Transformation

method. If the disturbances are ignored and a pure GME formulation is used, the GLS and GME solutions would be identical for interior solutions, $\beta \in \mathcal{Z}$.

An alternate entropy formulation is to specify the AR(1) relation in the model constraint and recover ρ as we do for the other unknowns. Here, a support of $[-1, 1]$ was specified for ρ and embedded in V, and the entropy of the resulting distribution on the correlation coefficient was added to the objective function. The prior information about ρ was first specified under a uniform distribution to give a prior mean of 0 (GME). Then, cross–entropy was used to solve the problem when the true value of ρ is used as the prior mean (GCE). Note that the model constraints are now nonlinear functions of the unknown probabilities. Although the computational properties of the problem are more complex, the analytical properties are largely unchanged (Shore and Johnson, 1980).

Due to the computational burden of the nonlinear problem, the number of Monte Carlo trials was reduced to 500. The resulting risk function of the GME–D and GCE–D solutions are presented in Figure 3.13, and the average value of the recovered ρ is plotted in Figure 3.14. As expected, the informative priors improve the risk behavior of the GME solution, and the average estimates of ρ are nearly identical to the true values.

One final alternative is to directly recover ρ rather than probabilities on the support of ρ. Here, ρ appears in the model constraints, but not the objective function. The optimal ρ is simply selected as a system parameter, and this approach follows the entropy formulation of dynamic optimization problems used by Golan, Judge and Karp (1993).

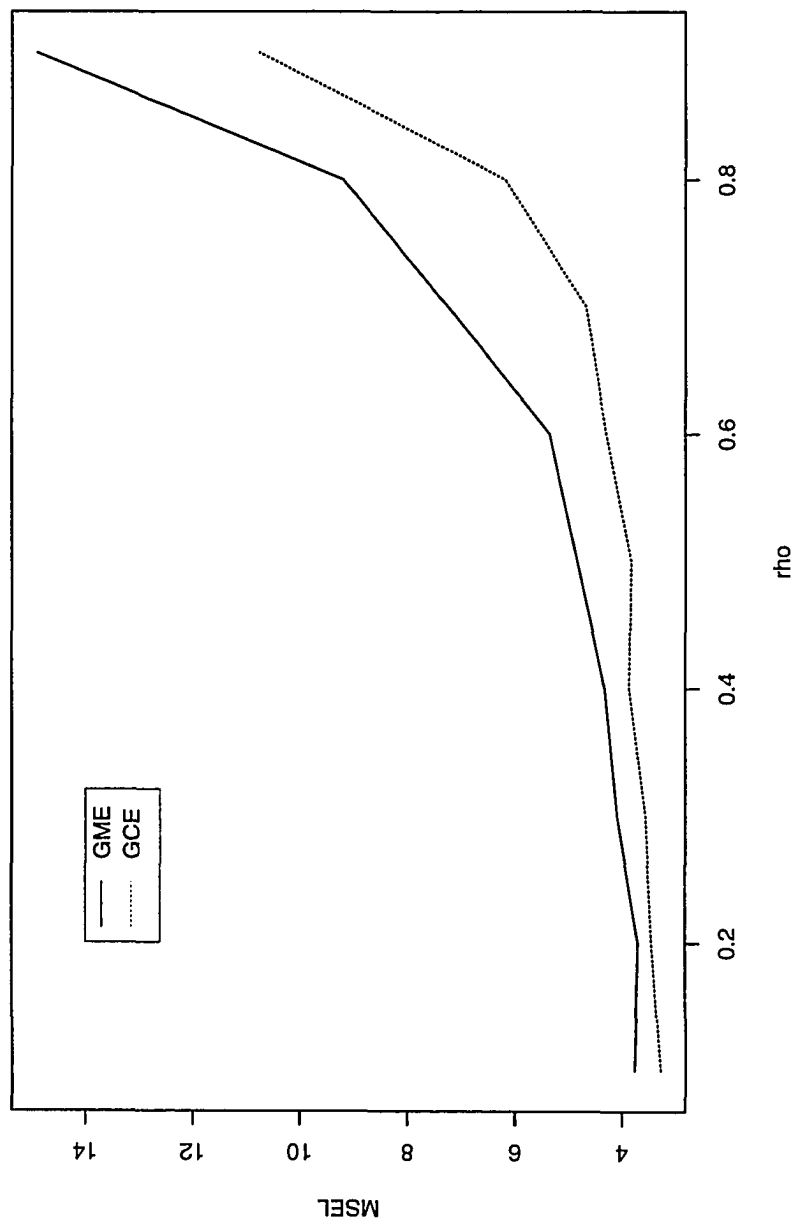

Figure 3.13: Nonlinear GCE Risk under AR(1) Errors

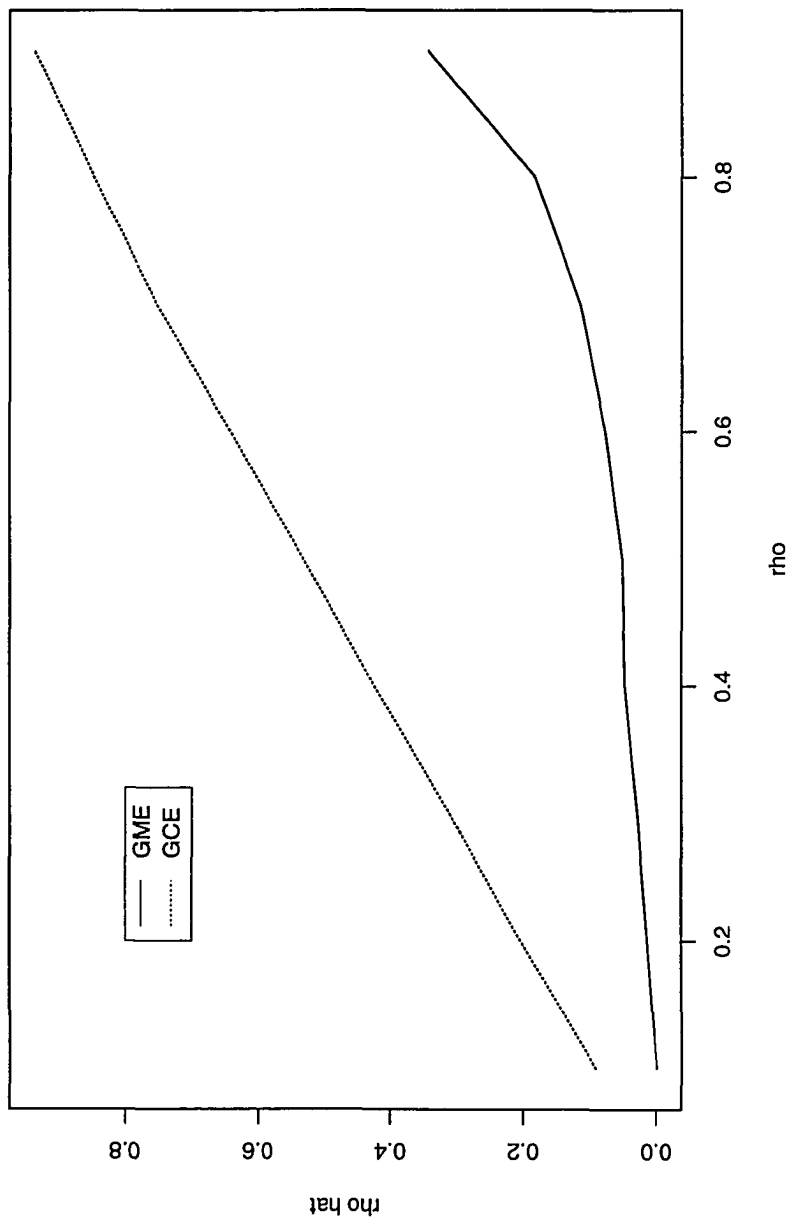

Figure 3.14: Average ρ from Nonlinear GCE

Chapter 4

Summary and Conclusions of the Research, with Extensions to Other Models

4.1 Summary of the Research Results

In the preceding chapters, the generalized entropy methods of information recovery were developed as extensions of the ME–CE formalisms. Unlike many of the traditional methods of inference, the entropy techniques require very little information and may be adapted to include the available sample or prior information. Given that economic data are characteristically noisy, limited, partial, or incomplete, the entropy approach is an attractive means for solving economic inverse problems.

The generic GCE problem was specified in Chapter 2, and it includes many of the members of the GLM family as special cases. The problem was solved analytically, and the solution was shown to be unique and admissible if the constraint set is non–empty. Although the solution does not take a closed form, the primal (constrained) problem satisfies the saddle–point property, and the dual formulation may be used to solve the problem in an unconstrained fashion. A brief computer algorithm was outlined, and it takes previously published computing rules as special cases.

The dual formulation may also be used to derive basic properties of the generalized entropy solutions. For the GCE–NM problem, tools developed to study M–estimators were used to show that the optimal Lagrange multipliers, $\hat{\lambda}_T$, and hence the point estimator, $\widehat{\beta}_T$, are consistent and asymptotically normal. In small samples, the approximate distribution of $\widehat{\beta}_T$ was derived from the large–sample results. Also, the FOC of the dual problem were used to show that the error bounds serve as shrinkage factors

Three major sampling exercises were presented in Chapter 3 to demonstrate the performance of the GCE–GME solutions in simple cases of the GLM. First, a bounded mean was recovered under different error distributions. The GME performance under SEL is relatively robust, and the shrinkage property provides for risk improvements in the presence of limited prior information. Next, the parameters of a linear model subject to collinearity were recovered using LS, RLS, ridge and GME estimators. The GME risk is nearly invariant to the degree of ill–conditioning, although the alternate estimators provide better prediction loss performance. The GME results may be improved by using more informative parameter supports or prior weights, but

incorrect information does little damage to the GME results. This is undoubtedly due to the use of the model constraints, which ensure that the GME solution must satisfy the observed information. Finally, a model with AR(1) errors was used to demonstrate various methods of handling dependent error structures in the generalized entropy framework. For example, the data may be transformed as in the GLS format, or the correlation coefficient may be recovered with the other unknowns.

In summary, the GCE–GME framework is a feasible approach for recovering information, especially if the inverse problem of interest is ill–posed, ill–conditioned, or reflects prior knowledge about the underlying economic system. Even if the inverse problem is well–posed and amenable to estimation under traditional methods, the GME–GCE solutions may provide better performance based on the limited evidence included in the Chapter 3.

4.2 Interpreting Generalized Entropy

Although the generalized entropy approach is not strictly Bayesian or frequentist, it may be related to methods employed by either group. In the sampling theory world, GME–GCE is a form of minimum distance estimation. However, the distances are not measured in the sample space, $\mathcal{Y}$. Rather, the sample information is used to constrain the solutions, which are selected to minimize Kullback–Liebler directed divergence in the parameter (probability) space.

As noted throughout the dissertation, the pure GCE solutions may be identical to restricted sample–based estimators. For example, the pure GME and the restricted ML–LS rules are the same for the bounded mean problem in Chapter 3. By including the disturbances, the GCE solutions to inverse problems with noise may be viewed as shrinkage rules that reduce the influence of the sample information based on the underlying signal–noise ratio. Unlike other shrinkage estimators, the GCE shrinkage factor is controlled by v, which is directly interpretted as an error bound.

From the Bayesian perspective, generalized entropy may be informally viewed as a non–parametric technique. True Bayes point estimators employ a likelihood function, which is incorporated with the prior information through Bayes rule. The

GCE approach does not use a likelihood, and recovers the posterior distribution that is closest to the prior, yet satifies the sample information. The width of the error support essentially serves as the variance of the likelihood function – wider bounds imply greater variation in the data, and the posterior more closely reflects the prior information.

Finally, it is important to reiterate that the generalized entropy methods should be viewed as an alternate method of information recovery. Although GCE shares properties with many of the traditional methods, it is not strictly frequentist or Bayesian.

4.3 Extensions of the Dissertation Research

The dissertation research may be extended in two principle ways. First, some of the analytical results used to demonstrate the properties of the generalized entropy estimators are new to the entropy literature. Examples include the computing algorithm, which extends previously published efforts, and the large–sample properties of the GCE–NM model. However, these results are relatively primative when compared to the current state of estimation theory. The behavior of more complex entropy formulations is of interest, and a great deal of work lies ahead. Nonetheless, the concept of generalized entropy is itself an innovation, and it will take some time before these methods are understood as deeply and as fully as the traditional methods, which have been with us for decades.

Second, the simplicity and convenience of the GLM are largely responsible for its popularity in applied research. A large number of models may be expressed in the GLM form, and only a few were examined in this research. Another linear model that has been studied under the generalized entropy framework is the first–order, finite and discrete Markov chain (stationary and otherwise) (Lee and Judge, 1992). The basic formulation of the Markov problem is to choose the set of transition probabilities, $\{p_{ij}\}$, that satisfy the observed transition relations *and* are closest to the prior

transition probabilities. Formally, the GME–GCE solution minimizes

$$I(p,q,w,u) = p'\log(p/q) + w'\log(w/u) \tag{4.1}$$

subject to

$$X = X_{-1}P + Vw \tag{4.2}$$

$$\imath_K = (I_K \otimes \imath_K)p \tag{4.3}$$

$$\imath_T = (I_T \otimes \imath_J)w \tag{4.4}$$

where X is the matrix of observed frequency distributions and P is the unknown transition matrix. Although a variety of estimation techniques have been devised for this problem, most require enough transitions to form a well–posed, stationary problem. Further, the flexible nature of the GME–GCE framework allows researchers to account for nonstationary chains by specifying a dynamic set of model constraints, Equation (4.2).

Other linear models that take this form include input–output relations, market share and size–distribution of firms, and qualitative choice models. In the latter case, the set of explanatory variables are used as weights to form first moments from the observations of the multinomial random variables. Specifically, the model equation is formed as

$$(I_J \otimes X')\,y = (I_J \otimes X')\,p \tag{4.5}$$

where y is the TJ-vector of observed multinomial choices, and p is the associated vector of probability distributions over choices for each observation (individual). Note that this is the FOC for the multinomial logit ML problem, which may be viewed as a special case of GME–GCE (as with LS–ML). By including a noise term, Vw, the GME–GCE solution is a shrinkage version of the standard sampling estimator. As in the preceding models, the GME–GCE solution to the qualitative choice problem may be computed with limited data or informative prior distributions — refer to Golan, Judge and Perloff (1994) for additional details.

A large number of linear models have yet to be considered under the GME–GCE framework, and these include simultaneous systems of equations, model selection

problems, and time series models in the time domain (e.g. ARMAX models). Although some features of these problems distinguish them from those cases already examined, there is no reason to believe that the properties and performance of the entropy methods will be vastly different. Finally, as demonstrated in Chapter 3, the GME–GCE specifications may be extended to handle nonlinear constraints as long as the constraint qualifications of nonlinear programming (Kuhn–Tucker or Arrow–Hurwicz–Uzawa) are satisfied. Thus, non–linear, non–stationary, or dynamic specifications may be included as additional model constraints.

Bibliography

Agmon, N., Alhassid, Y. and Levine, R. D. (1979). An algorithm for finding the distribution of maximal entropy, *Journal of Computational Physics* **30**: 250–8.

Amemiya, T. (1985). *Advanced Econometrics*, Harvard, Cambridge, MA.

Arrow, K. J. (1984a). Information and economic behavior, *The Economics of Information*, Vol. 4 of *Collected Papers of Kenneth J. Arrow*, Belknap Press, Cambridge, MA.

Arrow, K. J. (1984b). The value of and demand for information, *The Economics of Information*, Vol. 4 of *Collected Papers of Kenneth J. Arrow*, Belknap Press, Cambridge, MA.

Belsley, D. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regressions*, Wiley, New York.

Berger, J. and Wolpert, R. (1988). *The Likelihood Principle*, Vol. 6, IMS, Hayward, CA. Institute of Mathematical Statistics Lecture Notes—Monograph Series.

Bickel, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted, *Annals of Statistics* **9**: 1301–9.

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics*, Holden–Day, Oakland, CA.

Billingsley, P. (1986). *Probability and Measure*, Wiley, New York.

Brooke, A., Kendrick, D. and Meeraus, A. (1992). *GAMS: A User's Guide, Release 2.25*, Scientific Press, South San Francisco, CA.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families*, Vol. 9, IMS, Hayward, CA. Institute of Mathematical Statistics Lecture Notes–Monograph Series.

Casella, G. and Berger, R. (1990). *Statistical Inference*, Wadsworth and Brooks/Cole, Pacific Grove, CA.

Casella, G. and Strawdermann, W. E. (1981). Estimating a normal bounded mean, *Annals of Statistics* **9**: 870–8.

Csiszár, I. (1991). Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems, *Annals of Statistics* **19**: 2032–66.

Davidson, R. and Solomon, D. (1974). Moment–type estimation in the exponential family, *Communications in Statistics* **3**: 1101–8.

Denbigh, K. G. and Denbigh, J. S. (1985). *Entropy in Relation to Incomplete Knowledge*, Wiley, New York.

Donoho, D. (1994). Statistical estimation and optimal recovery, *Annals of Statistics* **22**: 238–70.

Donoho, D., Johnstone, I., Hoch, J. and Stern, A. (1992). Maximum entropy and the near black object, *Journal of the Royal Statistical Society, Series B* **54**: 41–81.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Fisher, R. A. (1950). On the mathematical foundations of theoretical statistics, *Contributions to Mathematical Statistics*, Wiley, New York.

Georgescu-Roegen, N. (1971). *The Entropy Law and the Economic Process*, Harvard, Cambridge, MA.

Geweke, J. (1986). Exact inference in the inequality constrained normal linear regression model, *Journal of Applied Econometrics* **1**: 127–142.

Ghosh, M. N. (1964). Uniform approximation of minimax point estimates, *Annals of Mathematical Statistics* **35**: 1031–47.

Golan, A. and Judge, G. G. (1992). Recovering and processing information in the case of underdetermined economic models, University of California at Berkeley.

Golan, A., Judge, G. G. and Karp, L. (1993). Recovering information in the case of ill–posed dynamic inverse problems, University of California at Berkeley.

Golan, A., Judge, G. G. and Perloff, J. (1994). A generalized maximum entropy approach to recovering information from multinomial response data, University of California at Berkeley.

Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables, *Annals of Mathematical Statistics* **34**: 911–34.

Gull, S. F. and Daniell, G. J. (1978). Image reconstruction from incomplete and noisy data, *Nature* **272**: 686–90.

Hansen, L. P. (1982). Large sample properties of method of moments estimators, *Econometrica* **50**: 1029–54.

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge, New York.

Hoerl, A., Kennard, R. and Baldwin, K. (1975). Ridge regression: Some simulations, *Communications in Statistics* **5**: 105–23.

Huber, P. (1981). *Robust Statistics*, Wiley, New York.

Jaynes, E. T. (1957a). Information theory and statistical mechanics, I, *Physics Review* **106**: 620–30.

Jaynes, E. T. (1957b). Information theory and statistical mechanics, II, *Physics Review* **108**: 171–90.

Jaynes, E. T. (1968). Prior probabilities, *IEEE Transactions on Systems Science and Cybernetics* **SSC–4**: 227–41.

Jaynes, E. T. (1985). Where do we go from here?, *Maximum Entropy and Bayesian Methods in Inverse Problems*, D. Reidel Publishing Co., Boston. Edited by C. R. Smith and W. T. Grandy.

Johansen, S. (1979). *Introduction to the Theory of Regular Exponential Families*, Vol. 3, Institute of Mathematical Statistics, University of Copenhagen, Copenhagen, Denmark. Lecture Notes.

Judge, G. G. and Golan, A. (1992). Recovering information in the case of ill–posed inverse problems with noise, University of California at Berkeley.

Judge, G. G., Golan, A. and Miller, D. (1994). Recovering information in the case of inverse problems with discrete noisy data, University of California at Berkeley.

Judge, G. G., Griffiths, W., Hill, R., Lütkepohl, H. and Lee, T.-C. (1985). *The Theory and Practice of Econometrics*, Wiley, New York.

Judge, G. G., Hill, R., Griffiths, W., Lütkepohl, H. and Lee, T.-C. (1988). *Introduction to the Theory and Practice of Econometrics*, Wiley, New York.

Kapur, J. N. (1989). *Maximum Entropy Models in Science and Engineering*, Wiley, New York.

Kapur, J. N. and Kesavan, H. K. (1992). *Entropy Optimization Principles with Applications*, Academic Press, Boston.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*, Academic Press, San Diego.

Kullback, J. (1959). *Information Theory and Statistics*, Wiley, New York.

Laffont, J.-J. (1989). *The Economics of Uncertainty and Information*, MIT Press, Cambridge, MA.

Lee, T.-C. and Judge, G. G. (1992). Entropy and cross-entropy estimation of non-stationary transition probabilities from aggregate data, University of California at Berkeley.

Lehmann, E. (1983). *Theory of Point Estimation*, Wiley, New York.

Levine, R. D. (1980). An information theoretical approach to inversion problems, *Journal of Physics, A* **13**: 91–108.

Lindley, D. (1972). *Bayesian Statistics, A Review*, Vol. 2, SIAM, Philadelphia.

Luce, R. D. (1960). The theory of selective information and some of its behavioral applications, *Developments in Mathematical Psychology*, Free Press, Glencoe, IL. Edited by R. D. Luce.

Maasoumi, E. (1993). A compendium to information theory in economic and econometrics, *Econometric Reviews* **12**: 137–81.

Manski, C. F. (1988). *Analog Estimation Methods in Econometrics*, Chapman and Hall, New York.

Miller, D. (1994). Solving generalized maximum entropy problems with unconstrained numerical techniques, University of California at Berkeley.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems, *Statistical Science* **1**: 502–27.

Press, W., Teukolsky, S., Vetterling, W. and Flannery, B. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge, New York.

Pukelsheim, F. (1994). The three sigma rule, *The American Statistician* **48**(4): 88–91.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Wiley, New York.

Rosenkrantz, R. D. (ed.) (1983). *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, D. Reidel Publishing Co., Boston.

Ryu, H. K. (1993). ME estimation of density and regression functions, *Journal of Econometrics* **56**: 397–440.

Sabatier, P. C. (1987). A few geometrical features of inverse and ill–posed problems, *Inverse and Ill–Posed Problems*, Academic Press, Boston. Edited by H. W. Engl and C. W. Groetsch.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*, Wiley, New York.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technical Journal* **27**: 379–423.

Shore, J. E. and Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross–entropy, *IEEE Transactions on Information Theory* **IT-26**: 26–37.

Shore, J. E. and Johnson, R. W. (1981). Properties of cross-entropy minimization, *IEEE Transactions on Information Theory* **IT-27**: 472–82.

Shore, J. E. and Johnson, R. W. (1983). Comments on and corrections to 'Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross–entropy', *IEEE Transactions on Information Theory* **IT-29**: 942–3.

Skilling, J. (1988). The axioms of maximum entropy, *Maximum Entropy and Bayesian Methods in Science and Engineering*, Vol. 1, Kluwer, Amsterdam, pp. 173–87.

Spanos, A. (1986). *Statistical Foundations of Econometric Modelling*, Cambridge, New York.

Takayama, A. (1985). *Mathematical Economics*, Cambridge, New York.

Theil, H. (1967). *Economics and Information Theory*, North–Holland, New York.

Theil, H. (1971). *Principles of Econometrics*, Wiley, New York.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*, Winston Publishing, Washington, D. C.

Tikochinsky, Y., Tishby, N. and Levine, R. D. (1984). Consistent inference of probabilities for reproducible experiments, *Physics Review Letters* **52**: 1357–60.

Titterington, D. M. (1984). The maximum entropy method for data analysis, *Nature* **312**: 381–2.

van Campenhout, J. M. and Cover, T. M. (1981). Maximum entropy and conditional probability, *IEEE Transactions on Information Theory* **IT-27**: 483–9.

Zellner, A. (1991). Bayesian and non–bayesian estimation using balanced loss functions, Graduate School of Business, University of Chicago.

Zellner, A. (1994). Bayesian method of moments / instrumental variables (BMOM/IV) analysis of mean and regression models, Graduate School of Business, University of Chicago.

Zellner, A. and Highfield, R. (1988). Calculation of ME distributions and approximation of marginal posterior distributions, *Journal of Econometrics* **37**: 195–209.