Extended Empirical Likelihood Estimation and Inference

Marco van Akkeren and George Judge[1]

University of California, Berkeley

Abstract

We extend the empirical likelihood method of estimation and inference proposed by Owen and others and demonstrate how it may be used in a general linear model context and to mitigate the impact of an ill-conditioned design matrix. A dual loss information theoretic estimating function is used along with extended moment conditions to yield a data based estimator that has the usual consistency and asymptotic normality results. Limiting chi-square distributions may be used to obtain hypothesis test or confidence intervals. The estimator appears to have excellent finite sample properties under a squared error loss measure.

Keywords: Empirical likelihood, semiparametric models, extended estimating equations, Kullback-Leibler information criterion, asymptotic estimator properties, Lagrange multiplier and pseudo likelihood ratio tests.

Field of Designation: Econometric Theory; Semiparametrics

AMS 1991 Classifications: Primary 62E20

JEL Classifications: C10, C24

[1] 207 Giannini Hall, The University of California, Berkeley, CA 94720; Phone: (510)642-0791; Fax: (510)643-8911; E-mail: akkeren@are.berkeley.edu, judge@are.berkeley.edu
It is a pleasure to acknowledge valuable discussions with Rudy Beran, Douglas Miller, Ron Mittelhammer, Art Owen, and Arnold Zellner during the course of this research.

1 Introduction

Empirical likelihood (EL) is a semiparametric method of estimation and inference that uses a multinomial likelihood supported on the dependent sample observations to yield limiting sampling properties similar to its parametric counterpart. In *iid* settings its properties have been investigated by Owen (1988, 1990), Hall (1990), DiCiccio, Hall and Romano (1991), Qin and Lawless (1994) and others. Owen (1991) and Kolaczyk (1994) have extended its applicability to the larger class of generalized linear models for univariate data that follow exponential family distributions (McCullagh and Nelder, 1989).

In the context of using EL to reason with a sample of data, consider the traditional noisy inverse problem where we are unable to measure the unknown k-dimensional vector $\boldsymbol{\beta} \in \mathcal{B}$ directly and instead observe a n-dimensional vector of noisy sample observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ that are consistent with the underlying data sampling process in the linear statistical model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{1.1}$$

where $\mathbf{X}$ is a $(n \times k)$ design matrix known to the experimenter. The unobservable components are the $(k \times 1)$ vector of unknown parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$ and the n-dimensional noise vector $\mathbf{e}$, which reflects a drawing from a multivariate (possibly normal) distribution with mean zero and precision matrix $\boldsymbol{\Sigma}_{\mathbf{e}} = \sigma^2 \mathbf{I}_n$. The objective given the linear statistical model (1.1) is to find estimator $\boldsymbol{\delta}(\mathbf{y}) \in \mathcal{D}$ of the unknown parameter vector $\boldsymbol{\beta} \in \mathcal{B}$ that yields small expected squared error loss (SEL)

$$\rho(\boldsymbol{\delta}(\mathbf{y})|\boldsymbol{\beta}) = \mathbb{E}_y \|\boldsymbol{\delta}(\mathbf{y}) - \boldsymbol{\beta}\|^2 \tag{1.2}$$

relative to conventional estimators. Under model (1.1), when $\mathbf{e}$ is a normal random vector, the maximum likelihood (ML) estimator, $\boldsymbol{\delta}^0(\mathbf{y})$ of $\boldsymbol{\beta}$, where

$$\boldsymbol{\delta}^o(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \sim N_k(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \tag{1.3}$$

under measure (1.2) is a best-unbiased and minimax estimator with constant risk $\rho(\boldsymbol{\delta}(\mathbf{y})|\boldsymbol{\beta}) = \sigma^2 tr(\mathbf{X}'\mathbf{X})^{-1} = tr(\boldsymbol{\Sigma}_\delta)$.

For parametric models such as this, the likelihood concept is central to estimation and inference. The great appeal of the likelihood-based methods in parametric estimation and hypothesis testing is due to their wide range of applications and well developed asymptotic theory that provides a foundation for their use when certain regularity conditions are fulfilled. If not enough information about the underlying data sampling process is available to specify the form of the likelihood function, traditionally non-likelihood data based methods such as least squares, method of moments, and empirical likelihood have been used to cope with a range of semiparametric inference problems. Other alternatives for avoiding the likelihood are the quasi maximum likelihood (White, 1982; 1993) and estimating equations (EE) approach (Godambe, 1960; Heyde and Morton, 1998). Qin and Lawless (1994) link EE and EL and demonstrate how to make use of information in the form of unbiased estimating equations. From a Bayesian estimation perspective, Zellner's (1994) Bayesian method of moments (BMOM) permits post data densities for parameters while avoiding the likelihood.

In practice much of the sample data may come from a badly designed experiment or one that is non-experimentally generated. Consequently, the design matrix $\mathbf{X}$ used in the context of sampling model (1.1) may be ill-conditioned. Therefore, traditional estimating rules may result in a situation where, (i) there may be arbitrary parameters, (ii) the solution may be undefined and/or, (iii) the estimates can be highly unstable giving rise to high variance or low precision for the recovered parameters. Given this result, one alternative is to rely on the shrinkage properties of the method of regularization (MOR) and penalized likelihood (PL) estimators (Hoerl and Kennard, 1970a; 1970b; O'Sullivan, 1986; Titterington, 1985). In general, MOR estimators fit the data subject to a penalty function and can be formulated as the solution of an optimization problem involving a measure of lack of fit (prediction) of the data, a convex measure of roughness or plausibility relating to the precision of the estimator and an unknown regularization or tuning parameter. Because regularization procedures involve the use of both data and prior notions about the unknown parameters, if a likelihood formulation is relevant, a Bayesian interpretation of the MOR-ridge estimator is possible.

Given the range of traditional and non-traditional estimators for trying to avoid the likelihood and coping with ill posed inverse problems with noise, in this paper we extend the empirical likelihood concept and demonstrate an estimator that exhibits the following characteristics: (i) belongs to the family of extremum, M-type, estimators, (ii) has asymptotic sample properties similar to those for parametric likelihoods, (iii) maintains superior risk behavior relative to conventional estimators in finite samples when the design matrix is ill-conditioned, (iv) is robust to variations in the sampling process, and (v) is computationally tractable.

This paper is organized as follows: to establish a notational base, we review in Section 2 the empirical likelihood (EL) and general estimating equations (EE) concepts. In Section 3, a data based information theoretic (DBIT) estimator is demonstrated. Sections 4 and 5 provide the corresponding asymptotic results and an illustration of the finite sample properties of the estimator. Concluding remarks are given in Section 6 and mathematical derivations for the theorems and lemmas are given in an appendix.

2 EL and EE Conceptual Base

Our objective in this section is to link the estimating equations and empirical likelihood concepts and to recognize the possibility of alternative criterion functions. If we let $\mathbf{y}$ denote the $(n \times 1)$ vector of sample observations from a distribution F that depends on the k-dimensional vector $\boldsymbol{\beta} \in \mathcal{B}$, then the empirical likelihood of this parameter is defined by considering the distributions supported on the sample where each y_i is assigned probability π_i. We assume information exists in the form of $m \geq k$ unbiased estimating functions, $\mathbf{g}(\mathbf{y}, \boldsymbol{\beta}) = \mathbf{0}$. The profile empirical likelihood is consequently defined as

$$l_E(\boldsymbol{\beta}) = \sup_{\beta \in \mathcal{B}} \left\{ \prod_{i=1}^{n} \pi_i \;\; | \;\; \sum_{i=1}^{n} \pi_i g_m(y_i, \boldsymbol{\beta}) = 0, \sum_{i=1}^{n} \pi_i = 1, \pi_i \geq 0 \qquad \forall \; i \right\}. \tag{2.1}$$

The solution to this constrained optimization problem yields an optimal estimate $\hat{\pi}_i = n^{-1}[\sum_{j=1}^{m} \hat{\lambda}_j g_j(y_i, \boldsymbol{\beta}) + 1]^{-1}$ which is generally not in closed form. When $m = k$, the resulting estimator of $\boldsymbol{\beta}$ and F are those provided by Owen (1988; 1990). For the overidentified case when $m > k$, the formulation and solution of Qin and Lawless (1994) results. Corresponding likelihood ratio statistics have limiting Chi-square distributions that results in an inference base analogous to parametric likelihoods.

DiCiccio and Romano (1990) and Jing and Wood (1996) have noted that an alternative EL formulation is to choose an estimating criterion that minimizes the Kullback-Leibler (KL) distance between the post data probabilities and uniform prior weights $q_i = 1/n$. This criterion results in the following maximum entropy empirical likelihood (MEEL) problem, (Shannon, 1948; Jaynes, 1957a; 1957b; Shore and Johnson, 1980; Skilling, 1989; and Csiszar, 1991): $\sup_{\boldsymbol{\beta} \in \mathcal{B}} S(\pi) = -\sum_{i=1}^{n} \pi_i \log(\pi_i)$ subject to $\sum_{i=1}^{n} \pi_i g_m(y_i, \boldsymbol{\beta}) = 0$ and $\sum_{i=1}^{n} \pi_i = 1$. The solution for the ith optimal weight takes the form

$$\tilde{\pi}_i = \frac{\exp[-\sum_{j=1}^{m} \tilde{\lambda}_j g_j(y_i, \boldsymbol{\beta})]}{\sum_{i=1}^{n} \exp[-\sum_{j=1}^{m} \tilde{\lambda}_j g_j(y_i, \boldsymbol{\beta})]} \equiv \frac{\exp[-\sum_{j=1}^{m} \tilde{\lambda}_j g_j(y_i, \boldsymbol{\beta})]}{\Omega(\tilde{\boldsymbol{\lambda}})} \tag{2.2}$$

where $\hat{\lambda}_j$ is the jth optimal Lagrange multiplier associated with the m moment constraints. Again, for the MEEL formulation, the usual asymptotic inference properties for the likelihood ratios follow. If $m = k$ both the traditional MEL and MEEL formulations yield the same solution. This, of course, is not surprising since the estimates $\boldsymbol{\delta}(\mathbf{y})$ can be obtained as roots of the corresponding estimating equations $\mathbf{g}(\mathbf{y}, \boldsymbol{\beta}) = 0$.

If prior information $\mathbf{q}$ exists for the unknown $\boldsymbol{\pi}$, one alternative is to minimize the Kullback-Leibler (K-L) distance between post data weights and the priors. This criterion is known as cross entropy (CE), (pp.29-31 Golan, Judge, and Miller, 1996; Gokhale and Kullback, 1978). Under the CE metric, the CEEL estimating criterion of the ME criterion (2.1) becomes

$$I(\boldsymbol{\pi}, \mathbf{q}) = \sum_{i=1}^{n} \pi_i \log(\pi_i/q_i) = \sum_{i=1}^{n} \pi_i \log(\pi_i) - \sum_{i=1}^{n} \pi_i \log(q_i) \tag{2.3}$$

The resulting optimal solution is derived from minimizing $I(\boldsymbol{\pi}, \mathbf{q})$ subject to the unbiased estimating equations and adding up conditions,

$$\tilde{\pi}_i = \frac{q_i \exp[-\sum_{j=1}^{m} \tilde{\lambda}_j g_j(y_i, \boldsymbol{\beta})]}{\sum_{i=1}^{n} q_i \exp[-\sum_{j=1}^{m} \tilde{\lambda}_j g_j(y_i, \boldsymbol{\beta})]} \equiv \frac{q_i \exp[-\sum_{j=1}^{m} \tilde{\lambda}_j g_j(y_i, \boldsymbol{\beta})]}{\Omega(\tilde{\boldsymbol{\lambda}})} \tag{2.4}$$

When the prior information is uniform and $q_i = 1/n \; \forall \; i$, the optimal CE solution (2.3) is equivalent to the optimal ME solution (2.1). If one uses the traditional Owen MEL criterion, one chooses the feasible weights $\tilde{\pi}_i$ that maximizes the probabilities assigned to the observed set of observations. Alternatively, under the MEEL criterion the feasible weight π_i involves the maximum of the expected value of all possible log-likelihoods, consistent with the structural constraints.

It is important to note that the MEL, MEEL and CEEL estimating criterions are embedded in the general measure

$$I(\mathbf{p}, \mathbf{q}, \alpha) = \frac{1}{\alpha(\alpha+1)} \sum_{i=1}^{n} p_i \Big[\Big(\frac{p_i}{q_i} \Big)^{\alpha} - 1 \Big] \tag{2.5}$$

As α goes from 0 to -1, the MEL and MEEL problems are nested in the maximum α-entropy framework and thus determine the estimation and inference principles. A third distance measure, the log-Euclidean likelihood, is obtained for $\alpha = -2$ and may be useful when the unknown parameters lie outside the convex hull spanned by the sample points. For a more complete discussion of the α-entropy functional, see Renyi (1961) and Cressie and Read (1984).

The EL and EE concepts offer a viable basis for semiparametric estimation and inference when the design matrix is well conditioned. In the case of ill-conditioning, estimates from EE, EL and LS can be highly unstable in small samples, resulting in high variance or low precision for the recovered parameters. Typically, the method of regularization (MOR) has been used (O'Sullivan, 1986) that yields an estimator

$$\boldsymbol{\delta}_{mor}(\phi) = arg \min_{\beta \in \mathcal{B}} [\psi(\mathbf{y}, \boldsymbol{\beta}) + k\Phi(\boldsymbol{\beta})] \tag{2.6}$$

where $\psi(\mathbf{y}, \boldsymbol{\beta})$ is some measure of lack of fit (prediction) relative to the data and $\Phi(\boldsymbol{\beta})$ is a convex measure of roughness or plausibility of the estimates. In general, an optimal value of k must be chosen to form a stable estimate and hence the MOR optimization problem involves a dual loss function of prediction and estimation precision (Zellner, 1994; Dey, Ghosh, and Strawderman, 1999). Given the uncertain sampling properties of MOR-like estimators, in the next section we propose an extension and interpretation of the EL and EE concepts that leads to a data based estimator that has the usual asymptotic behavior and performs well in finite samples under the SEL measure.

3 Data Based Information Theoretic Estimator

In developing an information theoretic alternative to MOR, we generalize the EL and EE concepts and focus on ill-conditioned inverse problems. In particular, we replace the traditional EL estimating function assumption of $\mathbf{g}(\mathbf{y}, \boldsymbol{\beta}) = \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$ and make use of information about the unknown distribution and parameters in the form of the following condition-structural data constraint:

$$\mathbf{h}(\mathbf{y}, \boldsymbol{\beta}, \mathbf{e}) = \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{e}) = \mathbf{0} \tag{3.1}$$

or equivalently,

$$n^{-1}[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{e}] = \mathbf{0}. \tag{3.2}$$

The k-dimensional structural data constraint contains $(k + n)$ unobservables, $(\boldsymbol{\beta}, \mathbf{e})$, and thus is ill-posed. Consequently, traditional matrix procedures do not yield a unique solution which satisfies the equation.

Traditional estimation methods may be described to belong to a general class of estimators $\Psi = \left\{\boldsymbol{\psi} \ : \ \boldsymbol{\delta}(\mathbf{y}) = \sum_{i=1}^{n}(\boldsymbol{\psi}_i \odot y_j)\right\}$ that are some weighted sum of the y_i values, given appropriate choices of $\boldsymbol{\psi}_i$. For example, in case of least squares estimation, we have $\boldsymbol{\psi}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$ which reduces to $\boldsymbol{\psi}_i = \mathbf{x}_i$ for the orthonormal linear model. In addition, the instrumental variables estimator, IV $(h = k)$, may be represented as $\boldsymbol{\psi}_i = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{z}_i$ or $\boldsymbol{\psi}_i = [(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_i$ for IV$(h > k)$. To solve the problem, we seek estimates $\boldsymbol{\delta}(\mathbf{y})$ and $\boldsymbol{\gamma}(\mathbf{y})$ that satisfies (3.1) or (3.2). In particular, centering the observables $\mathbf{y}$ and $\mathbf{X}$ we let $\boldsymbol{\delta}(\mathbf{y}) = (\mathbf{X}' \odot \mathbf{P})\mathbf{y}$ with $\mathbf{P} = [\mathbf{p}_1', \mathbf{p}_2', \ldots, \mathbf{p}_k']'$ and $\boldsymbol{\gamma}(\mathbf{y}) = (\mathbf{I}_n \otimes \mathbf{y}')\mathbf{w}$ and seek $\mathbf{p}$ and $\mathbf{w}$ defined by

$$n^{-1}[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}' \odot \mathbf{P})\mathbf{y} - \mathbf{X}'(\mathbf{I}_n \otimes \mathbf{y}')\mathbf{w}] = \mathbf{0} \tag{3.3}$$

In estimating (3.2) with (3.3), we make use of the Hadamard product $\odot$ and let the weights $\mathbf{p}_k$ and $\mathbf{w}_i$ represent $(k + n)$ univariate probability distributions, one for each element of $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$. Therefore, in the extended empirical likelihood formulation, we apply the traditional

EL weights $\boldsymbol{\pi}$ marginally and specify a univariate probability distribution $\mathbf{p}_k$ for each of the data based information theoretic (DBIT) estimates $\delta_k(\mathbf{y})$. Correspondingly, for $\boldsymbol{\gamma}(\mathbf{y})$, a univariate probability distribution $\mathbf{w}_i$ is specified for each $\gamma_i(\mathbf{y})$.

Traditional inversion procedures cannot be used to obtain a unique solution that satisfies (3.1). In seeking a solution, we make use of the information theory K-L distance measure between the estimated and reference distributions along with information in (3.3), we have the dual criterion extremum problem

$$\min_{\mathbf{p},\mathbf{w}} I(\mathbf{p},\mathbf{q},\mathbf{w},\mathbf{u}) = \mathbf{p}' \log(\mathbf{p}/\mathbf{q}) + \mathbf{w}' \log(\mathbf{w}/\mathbf{u}) \tag{3.4}$$

subject to

$$\frac{\mathbf{X}'\mathbf{y}}{n} = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)(\mathbf{X}' \odot \mathbf{P})\mathbf{y} + \left(\frac{1}{n}\right)\mathbf{X}'(\mathbf{I}_n \otimes \mathbf{y}')\mathbf{w} \tag{3.5}$$

$$\mathbf{i}_k = (\mathbf{I}_k \otimes \mathbf{i}_n')\mathbf{p} \tag{3.6}$$

$$\mathbf{i}_n = (\mathbf{I}_n \otimes \mathbf{i}_n')\mathbf{w} \tag{3.7}$$

where $\mathbf{q}$ and $\mathbf{u}$ are the reference distributions composed of uniform probabilities in the case of noninformative priors. In this extremum problem, the objective is to recover $\mathbf{p}$ and $\mathbf{w}$ and hereby derive an estimate of $\boldsymbol{\beta}$ consistent with (3.3).

Under the K-L distance, minimization of the objective function subject to the model's constraints under noninformative priors then draws each DBIT estimate of the error terms, $\gamma_i(\mathbf{y})$, towards the central tendency of zero which is simply the expected value of each noise component. Also, it should be observed that the normalization of the weights to have unit sum is analogous to the least squares counterpart where weights are defined by the positive definite nature of the idempotent matrix, $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$, that premultiplies the vector of dependent observations in defining the LS estimate of each error term.

In terms of the parameter space $\mathcal{B}$, the feasible values are defined by $(\mathbf{X}' \otimes \mathbf{P})\mathbf{y}$. In the same manner as before, the objective has a tendency to draw all the convexity weights to n^{-1}. Thus, there is a tendency to draw the estimate of $\boldsymbol{\beta}$ towards $n^{-1}\mathbf{X}'\mathbf{y}$ which are the sample covariances between $\mathbf{X}$ and $\mathbf{y}$. Also, it is observed that the row sums of the matrix weighting the $\mathbf{y}$ elements each equal zero for the DBIT method as with least squares.

3.1 Solution to the DBIT Problem

To solve this extremum problem, we form the Lagrangian function,

$$\begin{aligned}
\mathcal{L} = & \sum_{k=1}^{\kappa}\sum_{m=1}^{n} p_{km}\log(p_{km}/q_{km}) + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(w_{ij}/u_{ij}) + \\
& \sum_{l=1}^{\kappa}\lambda_l\left\{\frac{1}{n}\sum_{i=1}^{n} x_{il}y_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{\kappa}\sum_{m=1}^{n} x_{il}x_{ik}(x_{km}y_m p_{km}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{il}y_j w_{ij}\right\} + \\
& \sum_{k=1}^{\kappa}\mu_k(1 - \sum_{m=1}^{n} p_{km}) + \sum_{i=1}^{n}\gamma_i(1 - \sum_{j=1}^{n} w_{ij}) \qquad (3.8)
\end{aligned}$$

with the following FOCs,

$$\frac{\partial\mathcal{L}}{\partial p_{km}} = 1 + \log(\hat{p}_{km}/q_{km}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{\kappa}\hat{\lambda}_l x_{il}x_{ik}(x_{km}y_m) - \hat{\mu}_k = 0 \qquad \forall\, k, m \qquad (3.9)$$

$$\frac{\partial\mathcal{L}}{\partial w_{ij}} = 1 + \log(\hat{w}_{ij}/u_{ij}) - \frac{1}{n}\sum_{l=1}^{\kappa}\hat{\lambda}_l x_{il}y_j - \hat{\gamma}_i = 0 \qquad \forall\, i, j \qquad (3.10)$$

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial \lambda_l} = & \frac{1}{n}\sum_{i=1}^{n} x_{il}y_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{\kappa}\sum_{m=1}^{n} x_{il}x_{ik}(x_{km}y_m\hat{p}_{km}) - \\
& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{il}y_j w_{ij} = 0 \qquad \forall\, l \qquad (3.11)
\end{aligned}$$

$$\frac{\partial\mathcal{L}}{\partial \mu_k} = 1 - \sum_{m=1}^{n}\hat{p}_{km} = 0 \qquad \forall\, k \qquad (3.12)$$

$$\frac{\partial\mathcal{L}}{\partial \gamma_i} = 1 - \sum_{j=1}^{n}\hat{w}_{ij} = 0 \qquad \forall\, i \qquad (3.13)$$

Solving the above allows us to obtain the solutions,

$$\hat{p}_{km} = \frac{q_{km}\exp\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{\kappa}\hat{\lambda}_l x_{il} x_{ik}(x_{km}y_m)\right\}}{\sum_{m=1}^{n} q_{km}\exp\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{\kappa}\hat{\lambda}_l x_{il} x_{ik}(x_{km}y_m)\right\}} \equiv \frac{q_{km}\exp(.)}{\Omega_k(\hat{\boldsymbol{\lambda}})} \qquad \forall\, k,m \tag{3.14}$$

$$\hat{w}_{ij} = \frac{u_{ij}\exp\left\{\frac{1}{n}\sum_{l=1}^{\kappa}\hat{\lambda}_l x_{il} y_j\right\}}{\sum_{j=1}^{n} u_{ij}\exp\left\{\frac{1}{n}\sum_{l=1}^{\kappa}\hat{\lambda}_l x_{il} y_j\right\}} \equiv \frac{u_{ij}\exp(.)}{\Psi_i(\hat{\boldsymbol{\lambda}})} \qquad \forall\, i,j \tag{3.15}$$

The solution to this problem provides optimal weights $\hat{\mathbf{p}}$ and $\hat{\mathbf{w}}$ that are used to recover the estimates $\boldsymbol{\delta}(\mathbf{y})$ and $\boldsymbol{\gamma}(\mathbf{y})$.

3.2 Characteristics of the Solution

The optimization problem formulated in equations (3.4) through (3.7) has a solution for $\mathbf{p}$ and $\mathbf{w}$ if the intersection of the constraint sets is non-empty. Formally, we can define the sets $\mathcal{A} = \{\mathbf{w} > 0 : \mathbf{i}_n = (\mathbf{I}_n \otimes \mathbf{i}'_n)\mathbf{w}\}$, $\mathcal{B} = \{\mathbf{p} > 0 : \mathbf{i}_k = (\mathbf{I}_n \otimes \mathbf{i}'_n)\mathbf{p}\}$ and $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$ where $\mathcal{C}$ must consequently be non-empty and compact. In addition, we can restrict the elements of $\mathcal{C}$ to be strictly positive as to assure the optimal solution is global and unique. Considering the moment constraint (3.5), the fully restricted constraint set is defined to be,

$$\mathcal{C}^* = \{(\mathbf{p},\mathbf{w}) \in int(\mathcal{C}) : \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}(\mathbf{X}' \odot \mathbf{P})\mathbf{y} + \mathbf{X}'(\mathbf{I}_n \otimes \mathbf{y}')\mathbf{w}\} \tag{3.16}$$

where the notation $int(\mathcal{C})$ denotes the interior of set $\mathcal{C}$ such that all elements contained in the interior are strictly positive.

As noted before, if $\mathbf{q}$ and $\mathbf{u}$ are non-informative and therefore composed of discrete uniform distributions, then the objective of the Kullback-Leibler distance reduces to Shannon's entropy metric. Uniqueness of the optimal solution can be ensured from the positive definite property of the Hessian matrix of the objective function, (p.92, Golan, Judge, and Miller, 1996)

$$\nabla_{(p,w)(p',w')} I(\mathbf{p}, \mathbf{w}) = \begin{bmatrix} \mathbf{P}_*^{-1} & 0 \\ 0 & \mathbf{W}_*^{-1} \end{bmatrix} \tag{3.17}$$

where $\mathbf{P}_*^{-1}$ is a $(kn \times kn)$ diagonal matrix with elements p_{km}^{-1} and $\mathbf{W}_*^{-1}$ is a $(nn \times nn)$ diagonal matrix with elements equal to w_{ij}^{-1}. If the restriction is imposed that $(\mathbf{p}, \mathbf{w}) \in \mathcal{C}^*$ where $\mathcal{C}^* \neq \emptyset$ then the Hessian matrix is insured to be positive definite, $\mathbf{aHa}' > 0$ for arbitrary $\mathbf{a}$, implying a strictly convex objective function which is a sufficient condition for a global optimum. Finally we note that any constrained optimization problem may be formulated in its unconstrained dual form. Reasons for specifying the problem as such often point to efficiency gains of computing, however as will become evident from later sections, this approach also represents a considerable advantage when investigating the DBIT estimator limiting properties as an extremum estimator. In terms of the maximal value function for the extremum problem, we have the following lemma.

Lemma 3.1. *The maximal value function of the normed moment constrained DBIT optimization problem is,*

$$M_n(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \sum_{l=1}^{\kappa} \left(\frac{x_{il} y_i}{n} \right) \lambda_l - \sum_{k=1}^{\kappa} \log(\Omega_k(\boldsymbol{\lambda})) - \sum_{i=1}^{n} \log(\Psi_i(\boldsymbol{\lambda}))$$

where,

$$\begin{aligned} \Omega_k(\boldsymbol{\lambda}) &= \sum_{m=1}^{n} q_{km} \exp\left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{\kappa} \lambda_l x_{il} x_{ik} (x_{km} y_m) \right\} \\ \Psi_i(\boldsymbol{\lambda}) &= \sum_{j=1}^{n} u_{ij} \exp\left\{ \frac{1}{n} \sum_{l=1}^{\kappa} \lambda_l x_{il} y_j \right\}. \end{aligned}$$

Proof. The maximal value function $M_n(\boldsymbol{\lambda})$ is obtained by substitution into the Lagrangian as follows:

$$\begin{aligned}
\mathcal{L}^c(\boldsymbol{\lambda}) = & \sum_{k=1}^{\kappa}\sum_{m=1}^{n} p_{km}\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{\kappa}\lambda_l x_{il}x_{ik}(x_{km}y_m) - \log(\Omega_k(\hat{\boldsymbol{\lambda}}))\right\} + \\
& \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left\{\frac{1}{n}\sum_{l=1}^{\kappa}\lambda_l x_{il}y_j - \log(\Psi_i(\boldsymbol{\lambda}))\right\} + \\
& \sum_{l=1}^{\kappa}\lambda_l\left\{\frac{1}{n}\sum_{i=1}^{n} x_{il}y_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{\kappa}\sum_{m=1}^{n} x_{il}x_{ik}(x_{km}y_m p_{km}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{il}y_j w_{ij}\right\} \\
= & \sum_{i=1}^{n}\sum_{l=1}^{\kappa}\left(\frac{x_{il}y_i}{n}\right)\lambda_l - \sum_{k=1}^{\kappa}\log(\Omega_k(\boldsymbol{\lambda})) - \sum_{i=1}^{\kappa}\log(\Psi_i(\boldsymbol{\lambda})) \\
= & \left(\frac{\mathbf{y}'\mathbf{X}}{n}\right)\boldsymbol{\lambda} - \sum_{k=1}^{\kappa}\log(\Omega_k(\boldsymbol{\lambda})) - \sum_{i=1}^{n}\log(\Psi_i(\boldsymbol{\lambda})) \equiv M_n(\lambda) \qquad (3.18)
\end{aligned}$$

□

The maximal value function $M_n(\boldsymbol{\lambda})$ may be interpreted as a constrained expected log-likelihood function. Having defined the maximum value function for the reformulated problem, we turn to the limiting properties of the resulting extremum estimator.

4 Large Sample Properties and Statistical Tests

In the next two sections we demonstrate the asymptotic and finite sample properties as a basis for evaluating the effectiveness of our estimation rule. The fact that the estimator cannot be expressed in closed form and the highly nonlinear functions of the data that characterize the optimal solution introduce interesting complications. In this section we develop the limiting properties of consistency, asymptotic normality and the formulations required to derive asymptotic tests such as the Wald, likelihood ratio, and the Lagrange multiplier test. Finite sample results are developed in Section 5. In order to derive the above mentioned results we use the framework of extremum estimation (Huber, 1981; Newey and McFadden 1994). Specifically, for the consistency property we demonstrate that the maximal value function taking $\boldsymbol{\lambda} \in \Lambda$ as its argument converges to a non-stochastic limit in probability uniformly which is maximized at the true parameter $\boldsymbol{\lambda}_0$. The result that the DBIT estimator $\boldsymbol{\delta} \in \mathcal{D}$ of the unknown parameter vector $\boldsymbol{\beta}_0 \in \mathcal{B}$ is consistent follows.

4.1 Consistency

To demonstrate consistency property of our estimator we utilize the following regularity conditions:

Assumption 1.

1. *The disturbance terms* $e_1, \ldots, e_n$ *are iid with zero expectation,* $\mathbb{E}(e_i) = 0 \ \forall\, i$*, and covariances* $\boldsymbol{\Sigma}_{\mathbf{e}} = \sigma^2 \mathbf{I}_n$.

2. *The parameter space* $\mathcal{B}$ *is a compact subset of the Euclidean k-space,* $\mathbb{R}^k$*, and the true parameter* $\boldsymbol{\beta}_0$ *is in the interior of* $\mathcal{B}$.

3. *The non-stochastic* $(n \times k)$ *design matrix* $\mathbf{X}$ *has the property*

$$\lim_{n \to \infty} \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right) = \mathbf{Q}$$

   where $\mathbf{Q}$ *is a finite symmetric positive definite matrix.*

4. *There exists a parameter* $\boldsymbol{\lambda}_0 \in \Lambda$ *such that the following condition holds; plim* $[\mathbf{X}' \odot \mathbf{P}(\boldsymbol{\lambda}_0)]\mathbf{y} = \boldsymbol{\beta}_0$.

5. *The parameter space* Λ *is also a compact subset of* $\mathbb{R}^k$*, and the true parameter* $\boldsymbol{\lambda}_0$ *is in the interior of* Λ.

Keeping in mind all observables are centered, the first three assumptions are familiar basic regularity conditions that justify some of the fundamental asymptotic results on moment estimators and likelihood functions. The fourth regularity condition states that in the limit there exist specific weights $\mathbf{p}$ which when evaluated at the correct parameter $\boldsymbol{\lambda}_0$, the probability that the estimate differs from the true vector of unknowns, $\boldsymbol{\beta}_0$, becomes arbitrarily small. This set of assumptions is not the most general one that leads to the desired consistency results, however they are useful for our purpose. We may for example, relax the condition that $\mathcal{B}$ is a compact set, the *iid* property of the disturbance terms, or the finite value assumption on $\mathbf{Q}$, but with corresponding consequences. (Amemiya, 1993; Mittelhammer, Judge, and Miller, 1999).

Lemma 4.1. *Given the conditions of Assumption 1 are satisfied, $(\frac{\mathbf{y}'\mathbf{X}}{n})\boldsymbol{\lambda}$ converges to $\boldsymbol{\beta}_0'\mathbf{Q}\boldsymbol{\lambda}$ in probability uniformly.*

Proof. See the Appendix. □

Lemma 4.2. *As n approaches ∞, the maximal value function $M_n(\boldsymbol{\lambda})$ converges in probability uniformly in $\boldsymbol{\lambda} \in \Lambda$ to the nonstochastic function $M(\boldsymbol{\lambda})$ which attains a unique global maximum at $\boldsymbol{\lambda}_0$.*

Proof. See the Appendix. □

Theorem 1. *The estimator $\hat{\boldsymbol{\lambda}}_n$ defined by $M_n(\hat{\boldsymbol{\lambda}}_n) = \max_{\lambda\in\Lambda} M_n(\boldsymbol{\lambda})$ converges to the true parameter $\boldsymbol{\lambda}_0$ in probability.*

Proof. Let $\mathcal{G} \subset \mathbb{R}^k$ be an open neighborhood which contains $\boldsymbol{\lambda}_0$ and $\bar{\mathcal{G}}$ be its complement. By assumption 1.5, the parameter space Λ is closed and bounded and it follows that the intersection $\bar{\mathcal{G}} \cap \Lambda$ is compact as well, implying that $\max_{\lambda\in\bar{\mathcal{G}}\cap\Lambda} M(\boldsymbol{\lambda})$ exists. Let's define

$$\delta = M(\boldsymbol{\lambda}_0) - \max_{\lambda\in\bar{\mathcal{G}}\cap\Lambda} M(\boldsymbol{\lambda}) \qquad \forall\, \boldsymbol{\lambda} \tag{4.1}$$

and consider E_n to be the event such that $|M_n(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})| < \delta/2 \;\forall\, \boldsymbol{\lambda}$. This means that E_n

implies,

$$M(\hat{\boldsymbol{\lambda}}_n) > M_n(\hat{\boldsymbol{\lambda}}_n) - \delta/2 \tag{4.2}$$

$$M_n(\boldsymbol{\lambda}_0) > M(\boldsymbol{\lambda}_0) - \delta/2 \tag{4.3}$$

Since $\hat{\boldsymbol{\lambda}}_n$ is the optimal value which satisfies $M_n(\hat{\boldsymbol{\lambda}}_n) = \max_{\lambda \in \Lambda} M_n(\boldsymbol{\lambda})$, it follows that $M_n(\hat{\boldsymbol{\lambda}}_n) \geq M_n(\boldsymbol{\lambda}_0)$, hence we can express (4.2) as,

$$M(\hat{\boldsymbol{\lambda}}_n) > M_n(\boldsymbol{\lambda}_0) - \delta/2 \tag{4.4}$$

Adding both sides of (4.3) and (4.4) we obtain the inequality

$$M(\hat{\boldsymbol{\lambda}}_n) > M(\boldsymbol{\lambda}_0) - \delta \tag{4.5}$$

which reduces to

$$M(\hat{\boldsymbol{\lambda}}_n) > \max_{\lambda \in \bar{\mathcal{G}} \cap \Lambda} M(\boldsymbol{\lambda}) \tag{4.6}$$

From equation (4.6) we conclude that $\hat{\boldsymbol{\lambda}}_n \in \mathcal{G}$ which implies $P[\hat{\boldsymbol{\lambda}}_n \in \mathcal{G}] \geq P[E_n]$. Lemma 4.2 shows that $M_n(\boldsymbol{\lambda})$ converges to $M(\boldsymbol{\lambda})$ in probability uniformly, therefore $\lim_{n\to\infty} P[E_n] = 1$ and accordingly $\lim_{n\to\infty} P[\hat{\boldsymbol{\lambda}}_n \in \mathcal{G}] = 1$. The result that $\hat{\boldsymbol{\lambda}}_n \xrightarrow{p} \boldsymbol{\lambda}_0$ follows. □

Corollary 1. *The DBIT estimator* $\boldsymbol{\delta} = [\mathbf{X}' \odot \mathbf{P}(\hat{\boldsymbol{\lambda}}_n)]\mathbf{y}$ *converges in probability to the true parameter* $\boldsymbol{\beta}_0$.

Proof. See the Appendix. □

We have shown that using the maximal value function in an extremum estimator approach, the normed moment formulation of the estimator is consistent. It is straightforward to see from the above theorems that an alternative formulation which uses the condition that $\mathbf{g}(\mathbf{y}, \boldsymbol{\beta}) = 0$ is consistent as well, but not without considerable impact on finite sample properties. The advantage of the extended formulation is that it derives optimal biased estimates that have superior finite sample performance and are consistent. Our estimator is shown to be robust under problems of an ill-conditioned design, because it does not employ traditional inversion or decomposition procedures of an ill-conditioned matrix, but rather uses the information contained in the data by specifying weights on empirical target points which are recovered through minimization of the K-L distance measure.

Having derived the consistency property of our estimator, we now focus on the limiting distribution. In the next section we show that under certain conditions, the consistent root of the gradient of the maximal value function $M_n(\boldsymbol{\lambda})$ is asymptotically normal and consequently the DBIT estimator $\boldsymbol{\delta}$ asymptotically follows a normal distribution as well.

4.2 Asymptotic Normality

The following set of regularity conditions are used to show that the DBIT estimators for $\boldsymbol{\lambda}_0$ and $\boldsymbol{\beta}_0$ have asymptotically normal distributions.

Assumption 2.

1. *The expectation* $\mathbb{E}(\mathbf{X}'\mathbf{e})$ *equals zero.*

2. *The conditions for the Lindeberg-Feller CLT are satisfied;*

   Let $\mathbf{z}_i = \mathbf{x}_i e_i$ *for* $i = 1, \ldots, n$ *be a sequence of* $(k \times 1)$ *independent random vectors with* $\mathbb{E}(\mathbf{z}_i) = 0$ *and* $Cov(\mathbf{z}_i) = \sigma^2 \mathbf{x}_i \mathbf{x}_i'$ *and distribution function* F_i *such that,*

   (a) $\lim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^{n} \sigma^2 \mathbf{x}_i \mathbf{x}_i') = \sigma^2 \mathbf{Q}$

   (b) $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \int_{\xi} \|\boldsymbol{z}\|^2 dF_i(\boldsymbol{z}) = 0$

where $\xi = \{\|\boldsymbol{z}\| > \epsilon\sqrt{n}\}$ *for each* $\epsilon > 0$, *then*

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{z}_i = n^{-1/2}\mathbf{X}'\mathbf{e} \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{Q})$$

Assumption 2.1 states that in a repeated sampling context, on average $\mathbf{X}'\mathbf{e}$ equals to zero. However in contrast to traditional unbiased methods, within a single finite sample, $\mathbf{X}'\boldsymbol{\gamma}(\mathbf{y})$ is not restricted to equal zero. Assumption 2.2 also represents a common regularity condition used to derive limiting results of method of moments and likelihood estimators. It is clear that the usual Lindeberg-Levy CLT does not apply in this case since $\mathbf{x}_k'\mathbf{e}$ represents the sum of n independent but not *iid* random variables. A stronger condition (Liapounov) sometimes applied to obtain the asymptotic distribution of $n^{-1/2}\mathbf{X}'\mathbf{e}$ is the existence of a third moment of e_i (Judge, 1985; Greenberg and Webster, 1983), yet the weaker Lindeberg-Feller conditions are both necessary and sufficient. Amemiya (1985) shows the derivation of the asymptotic distribution using characteristic functions for the single parameter case. For more general results, see Theil (1971), Schmidt (1976).

Theorem 2. *Let the regularity conditions in Assumptions 1 and 2 hold. The DBIT estimator* $\hat{\boldsymbol{\lambda}}_n$ *defined by* $M_n(\hat{\boldsymbol{\lambda}}_n) = \max_{\lambda\in\Lambda} M_n(\boldsymbol{\lambda})$ *has a limiting normal distribution*

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\lambda}_0})$$

where the asymptotic covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}_0}$ *is appropriately defined.*

Proof. By Taylor expansion of the maximal value function around $\boldsymbol{\lambda}_0$ we have

$$\left.\frac{\partial M_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\right|_{\hat{\boldsymbol{\lambda}}_n} = \left.\frac{\partial M_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}_0} + \left.\frac{\partial^2 M_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\right|_{\boldsymbol{\lambda}^*}(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_0) \tag{4.7}$$

where $\boldsymbol{\lambda}^*$ lies between $\hat{\boldsymbol{\lambda}}_n$ and $\boldsymbol{\lambda}_0$. Since (4.7) equals to $\mathbf{0}$ by definition of $\hat{\boldsymbol{\lambda}}_n$, we rewrite the above equation as

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_0) = -\left[\frac{\partial^2 M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} \bigg|_{\boldsymbol{\lambda}^*} \right]^{-1} \sqrt{n}\, \frac{\partial M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \bigg|_{\boldsymbol{\lambda}_0} \tag{4.8}$$

The asymptotic distribution of $\hat{\boldsymbol{\lambda}}_n$ is derived in several steps. As will be shown, the first term inside the brackets to the right of the equality sign converges to a positive definite matrix and is nonsingular. Addressing the second term, we obtain from Lemma 4.1 the gradient of the maximal value function.

$$\begin{aligned} \sqrt{n}\, \frac{\partial M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \bigg|_{\boldsymbol{\lambda}_0} &= \sqrt{n}\Big(\frac{\mathbf{X}'\mathbf{y}}{n} - \frac{\mathbf{X}'\mathbf{X}}{n}[\mathbf{X}' \odot \mathbf{P}(\boldsymbol{\lambda}_0)]\mathbf{y} - \frac{1}{n}\mathbf{X}'(\mathbf{I}_n \otimes \mathbf{y}')\mathbf{w}(\boldsymbol{\lambda}_0) \Big) \\ &= \frac{1}{\sqrt{n}}\mathbf{X}'\Big(\mathbf{y} - \mathbf{X}[\mathbf{X}' \odot \mathbf{P}(\boldsymbol{\lambda}_0)]\mathbf{y}\Big) - \frac{1}{\sqrt{n}}\mathbf{X}'(\mathbf{I}_n \otimes \mathbf{y}')\mathbf{w}(\boldsymbol{\lambda}_0) \end{aligned} \tag{4.9}$$

In order to show that the second term of (4.9) converges in probability to $\mathbf{0}$, we use Chebychev's inequality with the additivity constraint on $\mathbf{w}(\boldsymbol{\lambda}_0)$ and finite first and second moments of the centered $\mathbf{y}$ to find a number M_1 such that for each k

$$P\big[n^{-1/2}\big|\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ik}y_j w_{ij}(\boldsymbol{\lambda}_0)\big| \geq M_2\big] \leq \frac{\mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ik}y_j w_{ij}(\boldsymbol{\lambda}_0)]^2}{nM_2^2} \leq \frac{M_1}{nM_2^2} \tag{4.10}$$

for any $M_2 > 0$. This result may also be obtained from recognizing that the weight $w_{ij}(\boldsymbol{\lambda}_0)$ converges in probability uniformly to u_{ij}. Referring to equation (3.17) we have,

$$\lim_{n\to\infty} P\Big(\sup_{\lambda_0 \in \Lambda}\Big| \frac{u_{ij}\exp\{\frac{1}{n}y_j\mathbf{x}_i'\boldsymbol{\lambda}_0\}}{\sum_{j=1}^{n} u_{ij}\exp\{\frac{1}{n}y_j\mathbf{x}_i'\boldsymbol{\lambda}_0\}} - \frac{u_{ij}\exp\{0\}}{\sum_{j=1}^{\infty} u_{ij}\exp\{0\}}\Big| < \epsilon\Big) = 1 \tag{4.11}$$

for every $\epsilon > 0$, or simply

$$\lim_{n\to\infty} P(\sup_{\lambda_0 \in \Lambda} |w_{ij}(\boldsymbol{\lambda}_0) - u_{ij}| < \epsilon) = 1 \tag{4.12}$$

which implies $\text{plim}\{\sum_{j=1}^{n} y_j w_{ij}(\boldsymbol{\lambda}_0)\} = 0$ since the elements of $\mathbf{y}$ are deviations about the mean. Applying Slutsky's Theorem, we can conclude that since convergence in probability implies convergence in distribution, the limiting distribution of (4.9) is therefore the same as that of

$$\frac{1}{\sqrt{n}}\mathbf{X}'\Big(\mathbf{y} - \mathbf{X}\Big\{\text{plim}\,[\mathbf{X}' \odot \mathbf{P}(\boldsymbol{\lambda}_0)]\mathbf{y}\Big\}\Big) = \frac{1}{\sqrt{n}}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) \tag{4.13}$$

$$= \frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{e} \tag{4.14}$$

Given that the regularity conditions in Assumption 2 hold, the Lindeberg-Feller CLT can be used to show

$$\sqrt{n}\left.\frac{\partial M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}_0} \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{Q}) \tag{4.15}$$

In order to find the probability limit of the Hessian matrix, we refer to Appendix A for the specific functional form. By the definition of $\boldsymbol{\lambda}^*$ and the consistency of $\hat{\boldsymbol{\lambda}}_n$ it follows that

$$\text{plim}\left.\frac{\partial^2 M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\lambda}'}\right|_{\boldsymbol{\lambda}^*} = \text{plim}\Big\{-\Big(\frac{\mathbf{X}'\mathbf{X}}{n}\Big)\Phi(\boldsymbol{\lambda}_0)\Big(\frac{\mathbf{X}'\mathbf{X}}{n}\Big)' - \Upsilon(\boldsymbol{\lambda}_0)\Big\} \tag{4.16}$$

It is straightforward to see that plim $\Upsilon(\boldsymbol{\lambda}_0)$ converges in probability to the zero matrix since $\frac{1}{n}x_{is}y_j$ is clearly $o_p(n)$ and hence all elements of $\Upsilon(\boldsymbol{\lambda}_0)$ converge to zero in probability. By assumption 1.3, we have $\lim_{n\to\infty}\frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{Q}$ and let plim $\Phi(\boldsymbol{\lambda}_0) \xrightarrow{p} \Xi_{\lambda_0}$ which is positive definite since $p_{km} \in (0, 1)$. Combining the appropriate parts, equation (4.16) becomes

$$\text{plim}\left.\frac{\partial^2 M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\lambda}'}\right|_{\boldsymbol{\lambda}^*} = -\mathbf{Q}\Xi_{\boldsymbol{\lambda}_0}\mathbf{Q}' \tag{4.17}$$

accordingly, we can derive the distribution of (4.8) as

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_0) \xrightarrow{d} N(\mathbf{0}, (\mathbf{Q}\,\Xi_{\boldsymbol{\lambda}_0}\mathbf{Q}')^{-1}\sigma^2\mathbf{Q}\,(\mathbf{Q}\,\Xi_{\boldsymbol{\lambda}_0}\mathbf{Q}')^{-1}) \tag{4.18}$$

$$\equiv N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\lambda}_0}) \tag{4.19}$$

which gives the result proposed in the theorem. □

Now that we have derived the limiting distribution of $\hat{\boldsymbol{\lambda}}_n$, it is straightforward to extend our results to the DBIT estimator of the unknown vector of coefficients. It is clear from the functional form (3.14) that $\boldsymbol{\delta} = \boldsymbol{\delta}(\mathbf{p}(\boldsymbol{\lambda}))$ is a continuous and monotonic function of $\boldsymbol{\lambda}$ and consequently we use the gradient of our estimator to approximate its limiting distribution. This technique is sometimes referred to as the "delta-method", and we summarize the conclusions of its application in the next theorem.

Corollary 2. *Let the appropriate assumptions hold. The DBIT estimator $\boldsymbol{\delta}$ has the following limiting distribution*

$$\sqrt{n}(\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_n) - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_\delta)$$

where the asymptotic covariance matrix $\boldsymbol{\Sigma}_\delta$ is equal to $\sigma^2\mathbf{Q}^{-1}$.

Proof. See the Appendix. □

In terms of limiting values, there is first order equivalence among the EL, EE and DBIT estimators.

4.3 Asymptotic Tests

The issue of hypothesis testing is carried through based on the consistency and asymptotic normality results derived above for the wald (wald, 1943), lagrange multiplier (rao, 1947), and the pseudo likelihood ratio (PLR) test. All test statistics share the same chi-squared distribution with degrees of freedom under the null hypothesis equal to the number of restrictions. As is well known, the three tests are identical in limiting properties. However, they differ in finite samples and are difficult to evaluate because of the functional forms.

Although all three tests are applicable, in the interest of space, we investigate the properties of the restricted DBIT estimator $\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_r)$ and the lagrange multiplier test. Estimates of the confidence interval for the unknown coefficients are then based on the duality property of the test. Consider the following hypothesis,

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r} \tag{4.20}$$

or

$$\mathbf{R}[\mathbf{X}' \odot \mathbf{P}(\boldsymbol{\lambda})]\mathbf{y} \equiv \mathbf{R}\boldsymbol{\delta}(\boldsymbol{\lambda}) = \mathbf{r} \tag{4.21}$$

where $\mathbf{R}$ is a j-dimensional vector of known constants, $\mathbf{r}$ is a $(j \times k)$ matrix of rank $j \leq k$ that restricts linear combinations of the unknown coefficients to equal to some scalar and $\mathbf{p}(\boldsymbol{\lambda})$ is both continuous and monotonic in $\boldsymbol{\lambda}$. we denote the restricted dbit estimator as $\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_r)$ where $\hat{\boldsymbol{\lambda}}_r$ is the value which satisfies

$$M_n(\hat{\boldsymbol{\lambda}}_r) = \sup_{\lambda \in \lambda} \{ M_n(\boldsymbol{\lambda}) \ \mid \ \mathbf{R}\boldsymbol{\delta}(\boldsymbol{\lambda}) = \mathbf{r} \} \tag{4.22}$$

and utilize the asymptotic results developed in sections 4.1 and 4.2. the main results are summarized in the following theorem.

Theorem 3. *Let the appropriate regularity conditions hold. Given the proposed restrictions are correct, the restricted DBIT estimator* $\hat{\boldsymbol{\lambda}}_r$ *defined by (4.22) has the following limiting* *distribution.*

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_r)$$

with asymptotic covariance matrix equal to

$$\boldsymbol{\Sigma}_r = [\mathbf{A}^{-1}(\mathbf{I} - \mathbf{B}'(\mathbf{B}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}\mathbf{A}^{-1})]\sigma^2\mathbf{Q}[\mathbf{A}^{-1}(\mathbf{I} - \mathbf{B}'(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{A}^{-1})]'$$

where $\mathbf{A} = \mathbf{Q}\Xi_{\lambda_0}\mathbf{Q}'$ *and* $\mathbf{B} = \mathbf{R}\mathbf{Q}\Xi_{\lambda_0}$*. Then the corresponding Langrange Multiplier test is of the form*

$$LM = n\boldsymbol{\mu}_r'(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')[\mathbf{B}\mathbf{A}^{-1}\sigma^2\mathbf{I}\mathbf{A}^{-1}\mathbf{B}']^{-1}(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')\boldsymbol{\mu}_r \xrightarrow{d} \chi^2_{j,0}$$

where the unknown quantities $\mathbf{Q}$ *and* σ^2 *can be replaced with* $\frac{1}{n}(\mathbf{X}'\mathbf{X})$ *and* $\hat{\sigma}^2$ *without changing the limiting distribution.*

Proof. See the Appendix. □

The reasoning behind the LM test described above is that given the restrictions are valid, $\hat{\boldsymbol{\lambda}}_r$ should approach the value that maximizes $M_n(\boldsymbol{\lambda})$. Consequently, the slope of the maximal value function should nearly equal $\mathbf{0}$ at this value. In fact, from a constrained optimization interpretation, the Lagrange multiplier measures the rate at which the maximal value function is increased when the restriction $\mathbf{R}\boldsymbol{\delta}(\boldsymbol{\lambda}) = \mathbf{r}$ is relaxed. Therefore, the larger is the discrepancy of $\boldsymbol{\mu}_r$ from $\mathbf{0}$, the less plausible the null hypothesis becomes.

A slightly modified version of the LM test is suggested by Rao (1948) which avoids computation of the Lagrange multiplier in the test statistic. This approach recognizes that under the null hypothesis, $H_0 : \mathbf{R}\boldsymbol{\delta}(\boldsymbol{\lambda}) = \mathbf{r}$, the FOC equation (A.25) of Appendix A implies that

$$\left.\frac{\partial M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\hat{\boldsymbol{\lambda}}_r} = \left[\mathbf{R}\left.\frac{\partial \boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\hat{\boldsymbol{\lambda}}_r}\right]' \boldsymbol{\mu}_r \tag{4.23}$$

Replacing the finite sample version of the quantity $(\mathbf{Q}\Xi_{\lambda_0}\boldsymbol{\mu}_r) \equiv \mathbf{B}\boldsymbol{\mu}_r$ with the term on the left hand side of equation (4.23) then gives us the alternative score form of the Lagrange Multiplier test which is readily computed and has an identical asymptotic distribution.

From the results presented by Theorem 1 we can also conclude that the restricted DBIT estimator of the unknown vector of coefficients follows a normal distribution as well. This conclusion follows directly from application of the delta-method.

Corollary 3. *Given the results of Theorem 3 hold, the restricted DBIT estimator $\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_r)$ has the limiting distribution,*

$$\sqrt{n}(\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_r) - \boldsymbol{\delta}(\boldsymbol{\lambda}_0)) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\delta_r})$$

with the asymptotic covariance matrix equal to $\mathbf{Q}\Xi_{\lambda_0}\boldsymbol{\Sigma}_r(\mathbf{Q}\Xi_{\lambda_0})'$.

Proof. See the Appendix. □

The pseudo-likelihood ratio test which is also applicable differs from the aforementioned asymptotic tests in that it uses both restricted and unrestricted DBIT estimates of $\boldsymbol{\lambda}$. It is based on the difference between the restricted and unrestricted values of the maximal value function $M_n(\boldsymbol{\lambda})$ and can be shown to have identical limiting properties as the Wald and the Lagrange Multiplier tests. As mentioned before, the finite sample properties of the DBIT estimator are difficult to analyze. Monte Carlo simulation does represent a practical method to evaluate the finite sample properties of the estimator and the powers of the discussed tests, and this is the topic to which we turn.

5 Finite Sample Results and Monte Carlo Evidence

We have noted that our estimator cannot be expressed in closed form and the optimal solutions to the proposed estimator formulations are highly non-linear functions of the data. As a consequence, finite sample properties of our estimator are difficult to obtain analytically. In this section we investigate these finite sample properties through the use of Monte Carlo sampling experiments and compare the performance of our estimator to the LS, EL, and a MOR-ridge estimator under the squared error loss measure for various levels of ill-conditioning of the design matrix.

In terms of the sampling experiment, a design matrix $\mathbf{X}$ is created for levels of ill-conditioning ranging from $k(\mathbf{X}'\mathbf{X}) = 1$ to 100 for a linear model with $\kappa = 4$ parameters and $n = 10$ observations. In order to create the (10×4) design matrix, coefficients are drawn from an *iid* $N(0, 1)$ pseudo-random number generator. This matrix is then appropriately transformed according to the specific condition number, $k(\mathbf{X}'\mathbf{X}) = \mu$, by replacing each characteristic root obtained through singular value decomposition of $\mathbf{X} = \mathbf{ULV}'$ by

$$\mathbf{a} = \left[\sqrt{\frac{2}{1+\mu}}, 1, 1, \sqrt{\frac{2\mu}{1+\mu}}\,\right]' \tag{5.1}$$

such that $\|\mathbf{a}\|^2 = 4$. The resulting design matrix $\mathbf{X}_n = \mathbf{UL_aV}'$ exhibits the property that $k(\mathbf{X}_n'\mathbf{X}_n) = \mu$, Belsley (1991). The vector of parameters is arbitrarily chosen to be $\boldsymbol{\beta}_0 = [2, 1, -3, 2]'$ following the experimental design of Golan, Judge, and Miller (1996). The noise component, e_i, is generated from an *iid* $N(0, 1)$ distribution from which the dependent values are calculated by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{e}$.

5.1 Sampling Experiments

In order to evaluate the sampling properties of the DBIT estimator, we make use of the quadratic loss function (1.2). In particular, the performance of the DBIT estimator will be compared to the LS, EL, and MOR-ridge estimator, where ridge estimates are computed using the optimal tuning parameter,k^*, that minimizes an unbiased estimator of the risk $\rho(\boldsymbol{\delta}(\mathbf{y};k)|\boldsymbol{\beta}) = \mathbb{E}_y\|\mathbf{X}\boldsymbol{\delta}(\mathbf{y};k) - \mathbf{X}\boldsymbol{\beta}\|^2$. Various procedures exist for selecting the optimal tuning parameter such as the simple, iterative, and generalized ridge regressor, but for illustrative purposes we follow Beran (1998) and use the optimal k^* which minimizes the criterion

$$\hat{R} = \frac{1}{\kappa}\sum_{i=1}^{\kappa}[\hat{\sigma}^2 f_i^2 + (z_i^2 - \hat{\sigma}^2)(1-f_i)^2] \tag{5.2}$$

where z_i is the ith element of $\mathbf{z} = \mathbf{U}'\mathbf{y}$ and $\mathbf{U}$ is obtained through singular value decomposition of $\mathbf{X}$. The element f_i is derived from $\mathbf{F} = diag[l_i^2/(l_i^2+k)]$ where l_i is the ith singular value of $\mathbf{X}$.

In addition, we investigate the performance of an alternative formulation where we replace the model consistency constraint (3.5) with

$$\mathbf{y} = \mathbf{X}(\mathbf{X}' \odot \mathbf{P})\mathbf{y} + (\mathbf{I}_n \otimes \mathbf{y}')\mathbf{w} \tag{5.3}$$

which we refer to as the DBIT(D) estimator and consequently increase the number of constraints from k to n. The results of an illustrative sampling experiment that compares the sampling performance of the DBIT estimator with its competitors are summarized in Table 5.1. A graphical representation of the table is given in Figure 5.1. These results are obtained using the GAMS optimization software for 5000 Monte Carlo trials and the above various levels of ill-conditioning. Visual inspection of the results indicates that in accordance with theory, the ML-LS estimator displays relatively poor sampling performance. The unbounded risk of the ML-LS estimator is virtually an increasing affine function of the degree of ill-conditioning ranging from a value of 3.96 for $k(\mathbf{X}'\mathbf{X}) = 1$, close to its theoretical value of 4, to a maximum of 52.46 for $k(\mathbf{X}'\mathbf{X}) = 100$. The evidence suggests that although LS es-

Table 5.1: Empirical SEL: results based on 5000 Monte Carlo trials

$k(\mathbf{X}'\mathbf{X})$	LS,EL	MOR-ridge	DBIT(D)	DBIT(M)
1	3.96279	4.18893	4.17322	3.96279
10	8.22531	5.28254	5.32828	4.87369
20	12.75875	6.77197	6.03561	5.38283
30	17.73376	9.13747	6.05134	5.87445
40	22.83802	10.28170	4.53039	4.07715
50	27.36695	12.59009	6.68919	5.91275
60	33.56641	13.11964	5.29602	4.83144
70	39.40164	15.99187	4.23836	3.86507
80	43.18114	16.66875	4.10232	3.77469
90	47.47539	19.44977	5.90707	5.76781
100	52.46118	20.88587	5.66741	5.02102

timates are on average unbiased, the increase in variance of these estimates due to problems of an ill-conditioned design matrix causes the overall sampling performance to be poor.

Observing the performance of the ridge regressor from Figure 5.1, we see that adding a small scalar k to each diagonal element of the matrix $(\mathbf{X}'\mathbf{X})$ causes a significant improvement in performance for $k(\mathbf{X}'\mathbf{X}) > 1$. The reason is that it dampens the effect of inverting eigenvalues close to zero by adding this small amount and hence reducing the value of the diagonal elements of the inverted matrix. Consequently the risk is still unbounded as $R(\boldsymbol{\delta}^r(\mathbf{y};k)|\boldsymbol{\beta}_0) \in [1,\infty)$ but increases at a lower constant rate than the risk of the LS estimator. Comparing variants of our estimator with the performance of other conventional estimators, we observe that both the moment and data formulation perform quite well compared to LS and ridge estimation.

While not a proof of the small sample superiority, these empirical risk outcomes suggest that the DBIT(M) estimator strictly dominates the LS, ridge, and DBIT(D) estimators. While both DBIT(M) and DBIT(D) perform well, we note the moment formulation produces more biased but more efficient estimates than the data formulation. The empirical risk under

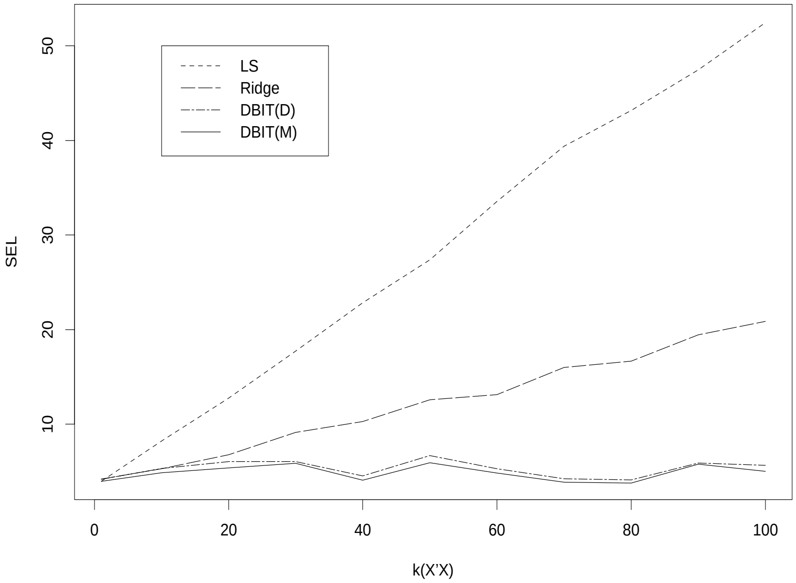

Figure 5.1: Comparison of performance in terms of SEL of selected estimators with DBIT regressors under various levels of ill-conditioning.

SEL for the moment variant ranges from a minimum of 3.96 for $k(\mathbf{X}'\mathbf{X}) = 1$ to a maximum around 6 for $k(\mathbf{X}'\mathbf{X}) \in [1, 100]$. The risk for DBIT(M) under condition number 1 is identical to the LS risk and suggests that for the orthonormal linear model, our estimates are identical to LS.

As an additional measure for estimator comparison, we use the squared error prediction loss (SEPL) of the form

$$L(\boldsymbol{\delta}(\mathbf{y})|\boldsymbol{\beta}) = ||\mathbf{X}\boldsymbol{\delta}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}||^2 \tag{5.4}$$

which is a weighted loss measure of (1.2). Comparing our sampling results we observe from Figure 5.2 that the DBIT(M) estimator strictly dominates the LS, EL and DBIT(D) estimator for the range of ill-conditioning. We note that the empirical SEPL for $k(\mathbf{X}'\mathbf{X}) = 1$

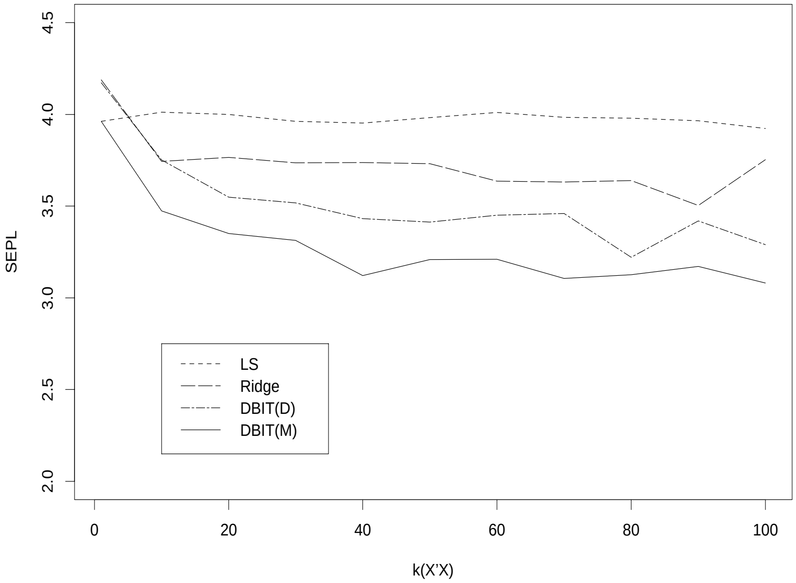

Figure 5.2: Comparison of performance in terms of SEPL of selected estimators with DBIT regressors under various levels of ill-conditioning.

are identical to those listed in Table 5.1 clearly because both measures are identical since $\mathbf{X}'\mathbf{X} = \mathbf{I}_k$. One surprising result however is that the empirical SEPL functions for the DBIT(D) and LS cross, indicating that LS performs better for the orthonormal linear model.

5.2 Empirical Distribution of $\boldsymbol{\delta}_2$

Under the framework of linear model (1.1), we do not state assumptions about the underlying data generating process but we can still refer to the results of Section 4 to derive limiting properties of DBIT estimates. Figure 5.3 displays the empirical density DBIT(M) estimate δ_2 of $\beta_2 = 1$ for the selected estimators when $k(\mathbf{X}'\mathbf{X}) = 50$ where the Gaussian kernel is selected with a bandwidth $b = 1.3$. We observe that the LS estimator is unbiased as its density is

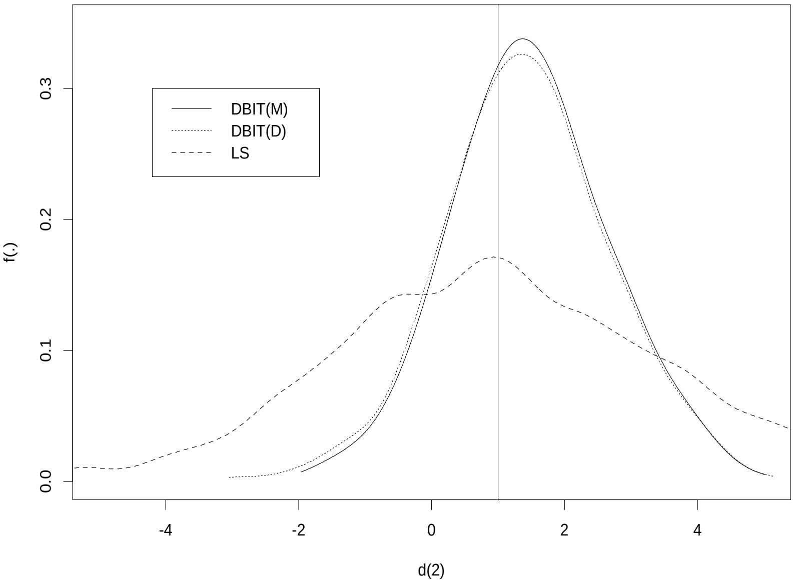

Figure 5.3: Empirical distribution of δ_2 when $k(\mathbf{X}'\mathbf{X}) = 50$

centered on the true coefficient $\beta_2 = 1$ but at the expense of a large variance. Furthermore the mean and variance are 1.02883 and 6.48985, compared in contrast to DBIT(M) values of 1.46458 and 1.33245 respectively. The bias of our DBIT estimator is clearly observed from the center of its distribution to the left of 1, while the smaller variance allows it to outperform the LS and EL under SEL.

As the degree of ill-conditioning increases, the variance of the LS estimator increases dramatically. The LS estimate is centered on the true value of 1, while the DBIT estimates are clearly biased but yield a considerable reduction in variance. Specifically, DBIT(M) has a mean value of 1.46458 compared to the LS estimate of 1.02833, yet in contrast the variances are 1.33245 compared to 6.48985 for LS. For a condition number 90, these effects are increased as the variance of the LS estimate explodes while the DBIT(M) estimate remains fairly accurate.

5.3 Empirical Distribution of the PLR Test Statistic

Figure 5.4 displays the empirical distribution of the pseudo-likelihood ratio (PLR) test statistic for the simple hypothesis $H_0 : \sum_{j=1}^{k} \beta_j = 2$. The test statistic is computed for 1000 Monte Carlo trials following the experimental design described in the introduction to this section and shows the empirical distribution for selected sample sizes $n = 10, 20$, and 30. In comparison to the chi-square(1) probability distribution, it is evident that the test statistic performs reasonably well in small samples. There appears to be a bias towards accepting the null hypothesis indicated by larger mass of the distributions concentrated around the origin which becomes smaller as the sample size increases, a result in agreement with the limiting results. The tails of the distributions also thicken to account for the increase in mass. In addition, it should be noted that the distortions observed in the left tails of the three empirical

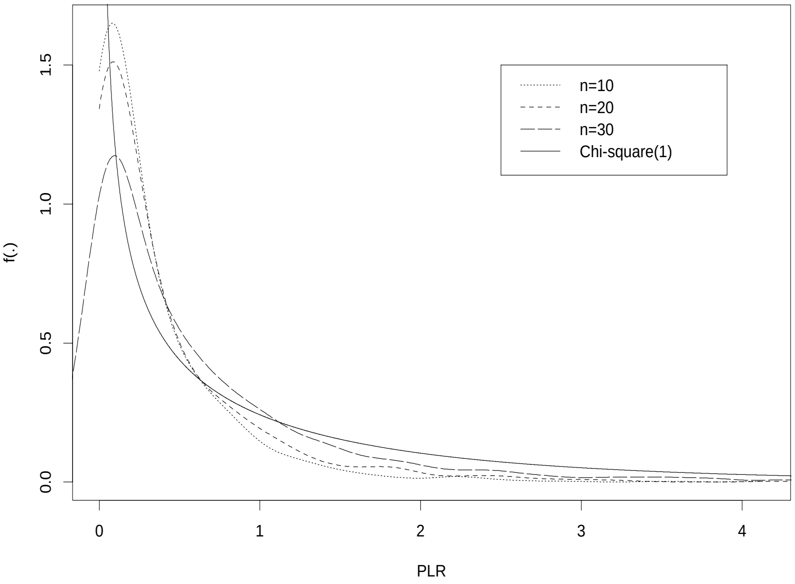

Figure 5.4: Empirical distribution of the pseudo-likelihood ratio (PLR) test statistic for sample sizes n=10,20, and 30.

distributions are due to selection of the Gaussian kernel in the density estimates.

5.4 Non-normal Disturbances

In order to gauge the robustness of the DBIT estimator under non-normal disturbances, we present sampling results for two alternative probability distributions. We investigate the sampling performance for the before examined set of estimators under the $\chi^2(4)$ and $t(3)$ distributions. The experimental design remains identical where $n = 10$, $k = 4$ and the data is generated according to the previous methods for 5000 Monte Carlo trials. It should be noted that the errors are drawn from an *iid* $N(0, 1)$ pseudo random number generator and then transformed to the appropriate distribution since the GAMS optimization package does not provide a pseudo random number generator for the chi-square and t-distribution. The chi-square distributed errors are centered by subtracting the mean of 4 from each term so that they may have negative values. Also, they are standardized to have unit variances by dividing values with the standard deviation $\sqrt{8}$ so that the sampling results are comparable to the experiments performed in the preceding sections. The t-distributed errors are also standardized to have unit variances but in this case the standard deviation is the square root of the degrees of freedom parameter or simply $\sqrt{3}$.

Figures 5.5 and 5.6 graphically summarize the information presented in Tables 5.2 and 5.3. Estimator performance is measured in terms of the SEL criterion and the results are not unlike those produced using standard normal errors. The SEL of the LS estimator under $\chi^2(4)$ errors increases with degree of ill-conditioning at a fairly constant rate of increase and is very similar to the results displayed in Table 5.1. The ridge regressor performs slightly worse under $\chi^2(4)$ errors than before while the rate of increase is nearly identical. Inspecting Figure 5.6 more closely, we can observe the rate of increase of the LS empirical SEL function is not constant as with standard normal errors but varies at several points over the range of ill-conditioning due to outlying values in thicker tails produced with t-distributed errors.

Table 5.2: Comparative performance of DBIT and selected regressors under SEL with standardized and centered $\chi^2(4)$ disturbances

$k(\mathbf{X}'\mathbf{X})$	LS,EL	Ridge	DBIT(M)
1	3.89360	4.46875	3.89360
10	7.96059	7.28096	4.72086
20	12.77414	9.50738	5.17080
30	17.75777	10.77814	5.84282
40	23.03823	11.98120	4.11402
50	27.42179	12.75590	5.51387
60	33.05799	13.80714	5.01966
70	36.81202	14.82771	3.83637
80	43.49648	16.09723	3.77920
90	47.34990	17.01336	5.72556
100	52.26859	17.41067	5.01862

Table 5.3: Comparative performance of DBIT and selected regressors under SEL with standardized and centered $t(3)$ disturbances

$k(\mathbf{X}'\mathbf{X})$	LS,EL	Ridge	DBIT(M)
1	4.20994	3.96302	4.20994
10	8.88392	6.58930	5.42143
20	13.09763	9.09495	5.24437
30	22.51305	14.13076	6.22545
40	22.07754	12.24956	4.13267
50	30.95490	13.29359	5.73645
60	36.09374	15.82557	4.92293
70	43.58555	17.02269	4.14288
80	47.60230	15.11180	4.08517
90	56.03513	16.63191	6.22665
100	55.23973	19.01172	5.18704

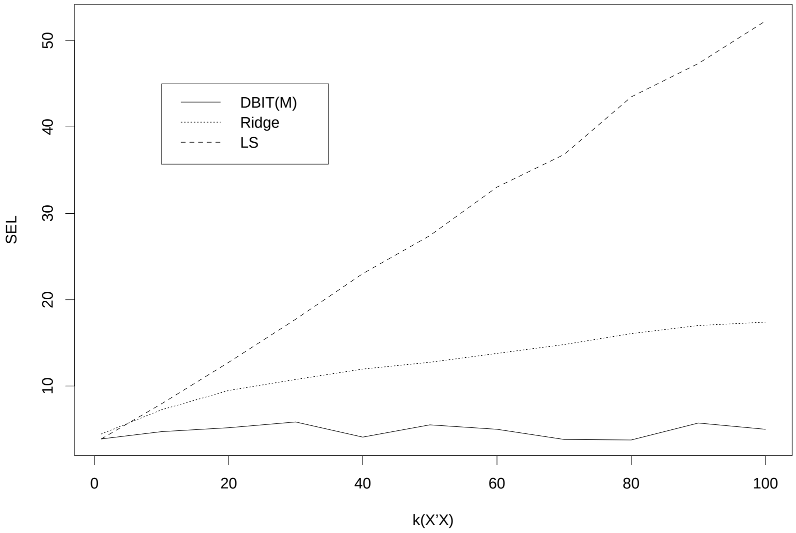

Figure 5.5: Comparison of performance of selected estimators with DBIT regressor under $\chi^2(4)$ distributed errors.

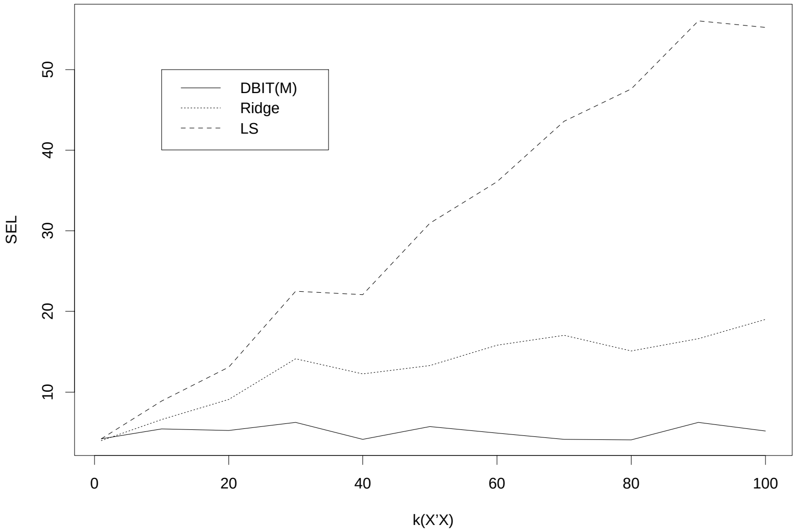

Figure 5.6: Comparison of performance of selected estimators with DBIT regressor under $t(3)$ distributed errors.

When outcome data are not well behaved and there are for example outliers, perhaps it would be better to use as a moment constraint, say in the location and scale case, the median $\sum_{i-1}^{n} p_i \text{sgn}(y_i - \beta) = 0$, where sgn(.) is the sign function. One reason DBIT estimation works well is that under high condition numbers we don't have the outliers that LS has to deal with because of a near singular $(\mathbf{X}'\mathbf{X})$. The ridge estimator does have a smooth empirical SEL function. These results suggest that the DBIT method produces robust estimates under at least the alternative error distributions examined in this section.

6 Concluding Remarks

In this paper we have been concerned in the case of the general linear sampling model with the question of "How to reason with a sample of Data?" In this context, we have demonstrated an extremum non-linear inversion procedure that permits us to avoid making some of the model assumptions that we may not wish to make and to cope with the problem of an ill-conditioned design matrix. Since in practice statistical probability models are often ill-posed-ill conditioned, and information about the sample is often represented in terms of fragmented moment relations, we have presented a basis for specifying the linear model where Kullback-Leibler information theoretic procedures may be used to recover the unknowns. As a result we demonstrate information theoretic estimation methods that (i) permit weak distributional assumptions, (ii) are flexible in terms of introducing sample and nonsample information, (iii) are appropriate for both well and ill-conditioned problems, (iv) focuses on both estimation precision and prediction objectives, (v) reflects a multiple shrinkage alternative that is data based, (vi) presents data based alternatives to formulations where the supports for the non-observables are based on nonsample information, and (vii) appears to be robust relative to underlying sampling processes that involve severe outliers.

When sample information is reflected in a structural constraint in the form (3.3) and the design matrix for the sampling model is ill-conditioned, the DBIT estimators appear to offer a robust alternative to ML, EL and LS procedures that are based on unbiased estimating functions. The DBIT estimators also offer a data based alternative to penalized likelihood, MOR-ridge, and principal components type estimators. In addition, the DBIT estimator solves the problem of bounds, spacing, and number of support spaces for $\mathbf{p}$ and $\mathbf{w}$ in generalized maximum entropy (GME) estimators (Golan, Judge, and Miller, 1996) for the case of an ill-posed inverse problem with noise. The proposed formulation and the resulting estimator achieves asymptotic results analogous to those used with parametric likelihoods while also demonstrating finite sample efficiency gains. Applicability of the formulation to a range of linear model sampling processes raises the possibility of a generalized moment based approach to estimation and inference. On a final note, we can give the DBIT estimator a multiple shrinkage interpretation (George, 1986) in that the support for each unknown observable contains n target points for which we seek an optimal linear combination under the K-L distance. Consequently we may find ways to scale the dispersion of these empirical target points based on some objective criterion which could result in improved finite sample performance.

Appendix A

Proof of Lemma 4.1

The linear statistical model (1.1) can be expressed as,

$$\frac{\mathbf{X}'\mathbf{y}}{n} = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)\boldsymbol{\beta}_o + \frac{\mathbf{X}'\mathbf{e}}{n} \tag{A.1}$$

By assumption 1.2 we have,

$$\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) \longrightarrow \mathbf{Q} \tag{A.2}$$

and from assumption 1.1 it follows that for all k, $\mathbb{E}[\mathbf{x}_k'\mathbf{e}] = 0$ and $\mathbb{E}[(\mathbf{x}_k'\mathbf{e})^2] = \sigma^2(\mathbf{x}_k'\mathbf{x}_k)$. Because $\mathbf{Q}$ is finite valued, there exists a number M_1 such that

$$P[n^{-1}|\mathbf{x}_k'\mathbf{e}| \geq M_2] \leq \frac{\sigma^2(\mathbf{x}_k'\mathbf{x}_k)}{n^2 M_2^2} \leq \frac{\sigma^2 M_1}{n M_2^2} \tag{A.3}$$

for any $M_2 > 0$ by Chebychev's inequality. Hence we obtain the well-known result that $\mathbf{X}'\mathbf{e} = o_p(n)$, (Judge et al, 1985). It follows,

$$\frac{\mathbf{X}'\mathbf{y}}{n} \xrightarrow{p} \mathbf{Q}\boldsymbol{\beta}_0 \tag{A.4}$$

and

$$\left(\frac{\mathbf{y}'\mathbf{X}}{n}\right)\boldsymbol{\lambda} \xrightarrow{p} \boldsymbol{\beta}_0'\mathbf{Q}'\boldsymbol{\lambda} \tag{A.5}$$

Proof of Lemma 4.2

The second term of the maximal value function $M_n(\boldsymbol{\lambda})$ in Lemma 4.1 is $\sum_{k=1}^{\kappa}\log(\Omega_k(\boldsymbol{\lambda}))$ where

$$\Omega_k(\boldsymbol{\lambda}) = \sum_{m=1}^{n} q_{km}\exp\left\{\sum_{i=1}^{n}\sum_{l=1}^{\kappa}\lambda_l\left(\frac{x_{il}x_{ik}}{n}\right)x_{mk}y_m\right\} \tag{A.6}$$

Since $\frac{1}{n}(\mathbf{X}'\mathbf{X}) \to \mathbf{Q}$, the sum $\sum_{i=1}^{n}\sum_{l=1}^{\kappa}\lambda_l\frac{1}{n}(x_{il}x_{ik})$ approaches the limit $\sum_{l=1}^{\kappa}\mathbf{Q}_{lk}\lambda_l$ as n goes to ∞. Taking the limit of the above expression we have,

$$\lim_{n\to\infty}\Omega_k(\boldsymbol{\lambda}) = \sum_{m=1}^{\infty} q_{km}\exp\left\{x_{mk}y_m\sum_{l=1}^{\kappa}\mathbf{Q}_{lk}\lambda_l\right\} \tag{A.7}$$

Likewise, the third term of $M_n(\boldsymbol{\lambda})$ is $\sum_{i=1}^{n}\log(\Psi_i(\boldsymbol{\lambda}))$ where the normalization factor in its explicit form is formulated as

$$\Psi_i(\boldsymbol{\lambda}) = \sum_{j=1}^{n} u_{ij}\exp\left\{\frac{1}{n}\sum_{l=1}^{\kappa}\lambda_l x_{il}y_j\right\} \tag{A.8}$$

which by application of Slutsky's theorem in the limit becomes

$$\text{plim}\ \Psi_i(\boldsymbol{\lambda}) = \sum_{j=1}^{\infty} u_{ij}\exp\{0\} = 1 \tag{A.9}$$

and

$$\text{plim}\sum_{i=1}^{\infty}\log(\Psi_i(\boldsymbol{\lambda})) = 0 \tag{A.10}$$

Combining the results of Lemma 4.2 with equations (A.7) and (A.10) we can conclude

$$M_n(\boldsymbol{\lambda}) \xrightarrow{p} \boldsymbol{\beta}_0'\mathbf{Q}'\boldsymbol{\lambda} - \sum_{k=1}^{\kappa} \log\left(\sum_{m=1}^{\infty} q_{km} \exp\left\{x_{mk} y_m \sum_{l=1}^{\kappa} \mathbf{Q}_{lk}\lambda_l\right\}\right) \equiv M(\boldsymbol{\lambda}) \qquad \text{(A.11)}$$

Next we shall show that $M(\boldsymbol{\lambda})$ attains a unique global maximum at $\boldsymbol{\lambda}_0$. It can be demonstrated that $M_n(\boldsymbol{\lambda})$ is strictly concave which remains true for the its limiting function and implies that the maximum will be uniqe and global. We can observe that $M(\boldsymbol{\lambda})$ is continuous and twice differentiable with the gradient function,

$$\begin{aligned}
\frac{\partial M(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} &= \boldsymbol{\beta}_0'\mathbf{Q}' - \sum_{k=1}^{\kappa}\left\{\frac{1}{\sum_{m=1}^{\infty} q_{km}\exp(.)} \sum_{m=1}^{\infty} q_{km}\exp(.)x_{mk}y_m\mathbf{Q}_k'\right\} \\
&= \boldsymbol{\beta}_0'\mathbf{Q}' - \sum_{k=1}^{\kappa}\sum_{m=1}^{\infty}\left\{\frac{q_{km}\exp(.)}{\sum_{m=1}^{\infty} q_{km}\exp(.)}\right\}x_{mk}y_m\mathbf{Q}_k' \\
&\equiv \boldsymbol{\beta}_0'\mathbf{Q}' - \sum_{k=1}^{\kappa}\sum_{m=1}^{\infty} p_{km}(\boldsymbol{\lambda})x_{mk}y_m\mathbf{Q}_k' \qquad \text{(A.12)}
\end{aligned}$$

For notation purposes, we revert to $\mathbf{Q}_k$ and $\mathbf{x}_i$ to indicate the kth column vector of $\mathbf{Q}$ and the ith row vector of the design matrix. Evaluating the Jacobian at the true parameter $\boldsymbol{\lambda}_0$ it becomes evident that the limiting function is globally maximized.

$$\begin{aligned}
\frac{\partial M(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\Big|_{\boldsymbol{\lambda}_0} &= \boldsymbol{\beta}_0'\mathbf{Q}' - \sum_{k=1}^{\kappa}\sum_{m=1}^{\infty} p_{km}(\boldsymbol{\lambda}_0)x_{mk}y_m\mathbf{Q}_k' \\
&\equiv \boldsymbol{\beta}_0'\mathbf{Q}' - \sum_{k=1}^{\kappa} \beta_0(k)\mathbf{Q}_k' \\
&\equiv \boldsymbol{\beta}_0'\mathbf{Q}' - \boldsymbol{\beta}_0'\mathbf{Q}' = 0 \qquad \text{(A.13)}
\end{aligned}$$

which gives us the desired result.

Proof of Corollary 1

For $\boldsymbol{\lambda} \in \Lambda$ we define $f_n(\boldsymbol{\lambda})$ dependent on the sample size as

$$f_n(\boldsymbol{\lambda}) = \sum_{m=1}^{n} \gamma x_{mk} y_m p_{km}(\boldsymbol{\lambda}) \tag{A.14}$$

then, as $n \to \infty$, let

$$\lim_{n\to\infty} f_n(\boldsymbol{\lambda}) = \sum_{m=1}^{\infty} \gamma x_{mk} y_m p_{km}(\boldsymbol{\lambda}) \equiv f(\boldsymbol{\lambda}) \tag{A.15}$$

Equation (3.14) shows that f is uniformly continuous for all $\boldsymbol{\lambda}$. By definition of continuity, for every $\epsilon > 0$ we can find η such that

$$\|\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_0\| < \eta \Longrightarrow \|f(\hat{\boldsymbol{\lambda}}_n) - f(\boldsymbol{\lambda}_0)\| < \epsilon \tag{A.16}$$

and it follows that

$$P[\|\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_0\| < \eta] \leq P[\|f(\hat{\boldsymbol{\lambda}}_n) - f(\boldsymbol{\lambda}_0)\| < \epsilon] \tag{A.17}$$

By Theorem 1, we have consistency of $\hat{\boldsymbol{\lambda}}_n$ and $\lim_{n\to\infty} P[\|\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_0\| < \eta] = 1$ which implies that $\lim_{n\to\infty} P[|f(\hat{\boldsymbol{\lambda}}_n) - f(\boldsymbol{\lambda}_0)| < \epsilon] = 1$ or equivalently $\lim_{n\to\infty} P[|\boldsymbol{\delta}_k - \boldsymbol{\beta}_0(k)| < \epsilon] = 1$ for all k.

Proof of Corollary 2

Referring to Section A.1.2 of van Akkeren (1999), we can express the sth partial derivative of the kth estimate as

$$\frac{\partial \boldsymbol{\delta}_k(\boldsymbol{\lambda})}{\partial \lambda_s} = \sum_{m=1}^{n} x_{km} y_m \frac{\partial p_{km}}{\partial \lambda_s} \tag{A.18}$$

$$= \sum_{i=1}^{n} \left(\frac{x_{is} x_{ik}}{n} \right) \left\{ \sum_{m=1}^{n} (x_{km} y_m)^2 p_{km} - (\sum_{m=1}^{n} x_{km} y_m p_{km})^2 \right\} \qquad \forall\, k, s \tag{A.19}$$

In matrix notation, the matrix of partial derivatives of $\boldsymbol{\delta}(\boldsymbol{\lambda})$ becomes

$$\frac{\partial \boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right) \Phi(\boldsymbol{\lambda}) \tag{A.20}$$

where

$$\text{plim} \left. \frac{\partial \boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \right|_{\hat{\boldsymbol{\lambda}}_n} = \mathbf{Q}\ \Xi_{\lambda_0} \equiv \Gamma. \tag{A.21}$$

Then we combine all parts to show that the asymptotic distribution of $\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_n)$ defined by equations (3.4)-(3.7) is characterized by,

$$\sqrt{n}(\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_n) - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \Gamma \boldsymbol{\Sigma}_{\boldsymbol{\lambda}_0} \Gamma') \tag{A.22}$$

$$\equiv N(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}) \tag{A.23}$$

which gives us the required result of the theorem.

Proof of Theorem 3

The constrained optimization problem formulated in equation (4.22) is solved using the usual method of Lagrange multipliers. Accordingly, the Lagrangian function

$$\mathcal{L} = M_n(\boldsymbol{\lambda}) + \boldsymbol{\mu}'(\mathbf{r} - \mathbf{R}\boldsymbol{\delta}(\boldsymbol{\lambda})) \tag{A.24}$$

yields the following optimality conditions

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = \frac{\partial M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} - \left[\mathbf{R}\frac{\partial \boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right]' \boldsymbol{\mu} = \mathbf{0} \tag{A.25}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = \mathbf{r} - \mathbf{R}\boldsymbol{\delta}(\boldsymbol{\lambda}) = \mathbf{0}. \tag{A.26}$$

Assuming $H_0 : \mathbf{R}\boldsymbol{\delta}(\boldsymbol{\lambda}) = \mathbf{r}$ is true, we next evaluate the FOCs at solutions $(\hat{\boldsymbol{\lambda}}_r, \boldsymbol{\mu}_r)$ and expand $\frac{\partial M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}$ around $\boldsymbol{\lambda}_0$ to obtain

$$\left.\frac{\partial M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\hat{\boldsymbol{\lambda}}_r} = \left.\frac{\partial M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}_0} + \left.\frac{\partial^2 M_n(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'}\right|_{\boldsymbol{\lambda}^*} (\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_0). \tag{A.27}$$

For equation (A.27) to hold with equality, the mean-value theorem states that the $(k \times 1)$ vector $\boldsymbol{\lambda}^*$ is between $\hat{\boldsymbol{\lambda}}_r$ and $\boldsymbol{\lambda}_0$. Also we have

$$\mathbf{R}[\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_r) - \boldsymbol{\delta}(\boldsymbol{\lambda}_0)] = \mathbf{0} \tag{A.28}$$

where because of continuity of $\boldsymbol{\delta}(\boldsymbol{\lambda})$, the mean-value theorem allows us to express,

$$\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_r) - \boldsymbol{\delta}(\boldsymbol{\lambda}_0) = \left.\frac{\partial \boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}^*} (\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_0). \tag{A.29}$$

Consequently, we rewrite the optimality conditions (A.25) and (A.26) as the following set of equations.

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\lambda}} = \frac{\partial M_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\bigg|_{\boldsymbol{\lambda}_0} + \frac{\partial^2 M_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\bigg|_{\boldsymbol{\lambda}^*}(\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_0) - \left[\mathbf{R}\frac{\partial\boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\bigg|_{\hat{\boldsymbol{\lambda}}_r}\right]'\boldsymbol{\mu}_r = \mathbf{0} \tag{A.30}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\mu}} = \mathbf{R}\frac{\partial\boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\bigg|_{\boldsymbol{\lambda}^*}(\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_0) = \mathbf{0} \tag{A.31}$$

Multiplying both equations (A.30) and (A.31) by $\sqrt{n}$, we can arrange the above system of equations in the following partitioned matrix form

$$\begin{bmatrix} -\frac{\partial^2 M_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\Big|_{\boldsymbol{\lambda}^*} & \left(\mathbf{R}\frac{\partial\boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\Big|_{\hat{\boldsymbol{\lambda}}_r}\right)' \\ \mathbf{R}\frac{\partial\boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\Big|_{\boldsymbol{\lambda}^*} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \sqrt{n}(\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_0) \\ \sqrt{n}\boldsymbol{\mu}_r \end{bmatrix} = \begin{bmatrix} \sqrt{n}\frac{\partial M_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\Big|_{\boldsymbol{\lambda}_0} \\ \mathbf{0} \end{bmatrix}. \tag{A.32}$$

Under the null hypothesis $H_0 : \mathbf{R}\boldsymbol{\delta}(\boldsymbol{\lambda}) = \mathbf{r}$, the estimate $\hat{\boldsymbol{\lambda}}_r$ converges in probability to $\boldsymbol{\lambda}_0$ since we have shown that $\hat{\boldsymbol{\lambda}}_n \xrightarrow{p} \boldsymbol{\lambda}_0$ by Theorem 1, therefore $\frac{\partial\boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\Big|_{\hat{\boldsymbol{\lambda}}_r} \xrightarrow{p} \mathbf{Q}\Xi_{\boldsymbol{\lambda}_0}$ by equation (A.21) of the proof to Corollary 2. Furthermore, we have derived in the previous section the following limiting results; i)plim $-\frac{\partial^2 M_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\Big|_{\boldsymbol{\lambda}^*} = \mathbf{Q}\Xi_{\boldsymbol{\lambda}_0}\mathbf{Q}'$, see result (4.17); ii)$\sqrt{n}\frac{\partial M_n(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\Big|_{\boldsymbol{\lambda}_0} \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{Q})$ from result (4.15). Then since convergence in probability implies convergence in distribution, we can represent the limiting properties of (A.32) using the following system of partitioned matrices.

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}' \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \sqrt{n}(\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_0) \\ \sqrt{n}\boldsymbol{\mu}_r \end{bmatrix} = \begin{bmatrix} \mathbf{Z} \\ \mathbf{0} \end{bmatrix} \tag{A.33}$$

For notation purposes we let $\mathbf{A} \equiv \mathbf{Q}\Xi_{\boldsymbol{\lambda}_0}\mathbf{Q}'$ and $\mathbf{B} = \mathbf{R}\mathbf{Q}\Xi_{\lambda_0}$ let $\mathbf{Z}$ be a multivariate normal random variable with mean $\mathbf{0}$ and covariances equal to $\sigma^2\mathbf{Q}$. To solve for the variables of interest we refer to the rules of partitioned matrix inversion (Rao and Toutenburg, 1995) which allows us to obtain our next result.

$$\begin{bmatrix} \sqrt{n}(\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_0) \\ \sqrt{n}\boldsymbol{\mu}_r \end{bmatrix} =$$
$$\begin{bmatrix} \mathbf{A}^{-1}(\mathbf{I} - \mathbf{B}'(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{A}^{-1}), & \mathbf{A}^{-1}\mathbf{B}'(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1} \\ (\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{A}^{-1}, & -(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{Z} \\ \mathbf{0} \end{bmatrix} \tag{A.34}$$

Two conclusions can be immediately drawn from the above manipulation. First we can observe that,

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_0) \overset{\text{LD}}{=} \mathbf{A}^{-1}(\mathbf{I} - \mathbf{B}'(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{A}^{-1}]\mathbf{Z}. \tag{A.35}$$

which implies that

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_r) - \boldsymbol{\lambda}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_r) \tag{A.36}$$

where

$$\boldsymbol{\Sigma}_r = [\mathbf{A}^{-1}(\mathbf{I} - \mathbf{B}'(\mathbf{B}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}\mathbf{A}^{-1})]\sigma^2\mathbf{Q}[\mathbf{A}^{-1}(\mathbf{I} - \mathbf{B}'(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{A}^{-1})]' \tag{A.37}$$

In addition, we observe from equation (A.34) that

$$\sqrt{n}\boldsymbol{\mu}_r \overset{\text{LD}}{=} (\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{Z} \tag{A.38}$$

which represents linear combinations of *iid* random variables, hence

$$\sqrt{n}\boldsymbol{\mu}_r \overset{A}{\sim} N(\mathbf{0}, (\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{A}^{-1}\sigma^2\mathbf{Q}\mathbf{A}^{-1}\mathbf{B}'(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}) \tag{A.39}$$

which can be expressed as

$$[(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{A}^{-1}\sigma^2\mathbf{Q}\mathbf{A}^{-1}\mathbf{B}(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}]^{-1/2}\sqrt{n}\boldsymbol{\mu}_r \overset{A}{\sim} N(\mathbf{0},\mathbf{I}_j) \tag{A.40}$$

and finally it follows,

$$n\boldsymbol{\mu}_r'[(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}\mathbf{B}\mathbf{A}^{-1}\sigma^2\mathbf{Q}\mathbf{A}^{-1}\mathbf{B}(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-1}]^{-1}\boldsymbol{\mu}_r \overset{A}{\sim} \chi^2_{j,0} \tag{A.41}$$

which gives the distribution of the LM test proposed by the theorem.

Proof of Corollary 3

From the proof to Corollary 2, equation (A.21) states

$$\frac{\partial\boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}} = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)\Phi(\boldsymbol{\lambda}) \tag{A.42}$$

such that given the restrictions are correct, we have

$$\text{plim}\left.\frac{\partial\boldsymbol{\delta}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\right|_{\hat{\boldsymbol{\lambda}}_r} = \mathbf{Q}\Xi_{\lambda_0}. \tag{A.43}$$

Then, straightforward application of the delta-method gives

$$\sqrt{n}(\boldsymbol{\delta}(\hat{\boldsymbol{\lambda}}_r) - \boldsymbol{\delta}(\boldsymbol{\lambda}_0)) \xrightarrow{d} N(\mathbf{0},(\mathbf{Q}\Xi_{\lambda_0})\boldsymbol{\Sigma}_r(\mathbf{Q}\Xi_{\lambda_0})') \tag{A.44}$$

References

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

Belsley, D. A. (1991). *Conditioning Diagnostics*, John Wiley and Sons, New York.

Csiszar, I. (1991). Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems, *The Annals of Statistics* **19**: 1032–2066.

Dey, D. K., Ghosh, M. and Strawderman, W. E. (1999). On estimation with balanced loss functions, *working paper* .

DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is bartlett-correctable, *The Annals of Statistics* **19**: 1053–1061.

DiCiccio, T. and Romano, J. (1990). Nonparametric confidence limits by resampling methods and least favorable families, *International Statistical Review* **58**: 59–76.

George, E. I. (1986). Minimax multiple shrinkage estimation, *The Annals of Statistics* **14**: 188–205.

Godambe, V. (1960). An optimum property of regular maximum likelihood estimation, *Annals of Mathematical Statistics* **31**: 1208–1212.

Gokhale, D. and Kullback, S. (1978). *The Information in Contingency Tables*, Marcel Dekker.

Golan, A., Judge, G. G. and Miller, D. (1996). *Maximum Entropy Econometrics*, John Wiley and Sons, New York.

Greenberg, E. and Webster, C. (1983). *Advanced Econometrics: A Bridge to the Literature*, John Wiley and Sons, New York.

Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood, *The Annals of Statistics* **18**: 121–140.

Heyde, C. and Morton, R. (1998). Multiple roots in general estimating equations, *Biometrika* **85**(4): 954–959.

Hoerl, A. and Kennard, R. (1970a). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**: 55–67.

Hoerl, A. and Kennard, R. (1970b). Ridge regression: Iterative estimation of the biasing parameter, *Technometrics* **12**: 69–82.

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

James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proceedings of the Fourth Berkeley Symposium on Math. Statist. Prob.* **1**: 361–379.

Jaynes, E. (1957a). Information theory and statistical mechanics, *Physics Review* **106**: 620–630.

Jaynes, E. (1957b). Information theory and statistical mechanics ii, *Physics Review* **108**: 171–190.

Jing, B. and Wood, A. (1996). Exponential empirical likelihodd is not bartlett correctable, *The Annals of Statistics* **24**: 365–369.

Judge, G. G. and Bock, M. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*, North-Holland, New York.

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

Kolaczyk, E. D. (1994). Empirical likelihood for generalized linear models, *Statistica Sinica* **4**: 199–218.

Mittelhammer, R., Judge, G. and Miller, D. (1999). *Econometric Foundations*, Cambridge University Press, New York.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing, *Handbook of Econometrics* **4**: 2111–2241.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems, *Statistical Science* **1**: 502–527.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika* **75**: 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions, *The Annals of Statistics* **18**: 90–120.

Owen, A. (1991). Empirical likelihood for linear models, *The Annals of Statistics* **19**(4): 1725–1747.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations, *The Annals of Statistics* **22**(1): 300–325.

Rao, C. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, *Proceedings of the Cambridge Philosophical Society* **44**: 50–57.

Schmidt, P. (1976). *Econometrics*, Marcel Dekker, New York.

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

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

Skilling, J. (1989). The axioms of maximum entropy, *in* J. Skilling (ed.), *Maximum Entropy and Bayesian Methods in Science and Engineering*, Kluwer Academic, Dordrecht, pp. 173–187.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* pp. 197–206.

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

Titterington, D. (1985). Common structures of smoothing techniques in statistics, *International Statistical Review* **53**: 141–170.

van Akkeren, M. (1999). *Data Based Information Theoretic Estimation*, PhD thesis, University of California, Berkeley, Berkeley, CA 94720.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society* **54**: 426–482.

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**: 1–16.

White, H. (1993). *Estimation, Inference, and Specification Analysis*, Cambridge University Press, New York.

Zellner, A. (1994a). Bayesian and non-bayesian estimation using balanced loss functions, *in* S. Gupta and J. Berger (eds), *Statistical Decision Theory and Related Topics*, Springer Verlag, New York.

Zellner, A. (1994b). Bayesian method of moments/instrumental variable (bmom/iv) analysis of mean and regression models, *Proceedings of the American Statistical Association* .