

Article

An Entropic Estimator for Linear Inverse Problems

Amos Golan ¹ and Henryk Gzyl ^{2,*}

¹ Department of Economics, Info-Metrics Institute, American University, 4400 Massachusetts Ave., Washington, DC 20016, USA; E-Mail: agolan@american.edu

² Centro de Finanzas, IESA, Caracas 1010, Venezuela

* Author to whom correspondence should be addressed; E-Mail: henryk.gzyl@iesa.edu.ve; Tel.: +58-212-555-4385; Fax: +58-212-555-4437.

Received: 29 February 2012; in revised form: 2 April 2012 / Accepted: 17 April 2012 /

Published: 10 May 2012

Abstract: In this paper we examine an Information-Theoretic method for solving noisy linear inverse estimation problems which encompasses under a single framework a whole class of estimation methods. Under this framework, the prior information about the unknown parameters (when such information exists), and constraints on the parameters can be incorporated in the statement of the problem. The method builds on the basics of the maximum entropy principle and consists of transforming the original problem into an estimation of a probability density on an appropriate space naturally associated with the statement of the problem. This estimation method is generic in the sense that it provides a framework for analyzing non-normal models, it is easy to implement and is suitable for all types of inverse problems such as small and or ill-conditioned, noisy data. First order approximation, large sample properties and convergence in distribution are developed as well. Analytical examples, statistics for model comparisons and evaluations, that are inherent to this method, are discussed and complemented with explicit examples.

Keywords: maximum entropy method; generalized entropy estimator; information-theoretic methods; parameter estimation; inverse problems

PACS Codes: 02.50.Ga; 02.50.Tt; 02.70.Rr; 83.85.Ns

1. Introduction

Researchers in all disciplines are often faced with small and/or ill-conditioned data. Unless much is known, or assumed, about the underlying process generating these data (the signal and the noise) these types of data lead to ill-posed noisy (inverse) problems. Traditionally, these types of problems are solved by using parametric and semi-parametric estimators such as the least squares, regularization and non-likelihood methods. In this work, we propose a semi-parametric information theoretic method for solving these problems while allowing the researcher to impose prior knowledge in a non-Bayesian way. The model developed here provides a major extension of the Generalized Maximum Entropy model of Golan, Judge and Miller [1] and provides new statistical results of estimators discussed in Gzyl and Velásquez [2].

The overall purpose of this paper is fourfold. First, we develop a generic information theoretic method for solving linear, noisy inverse problems that uses minimal distributional assumptions. This method is generic in the sense that it provides a framework for analyzing non-normal models and it allows the user to incorporate prior knowledge in a non-Bayesian way. Second, we provide detailed analytic solutions for a number of possible priors. Third, using the concentrated (unconstrained) model, we are able to compare our estimator to other estimators, such as the Least Squares, regularization and Bayesian methods. Our proposed model is easy to apply and suitable for analyzing a whole class of linear inverse problems across the natural and social sciences. Fourth, we provide the large sample properties of our estimator.

To achieve our goals, we build on the current Information-Theoretic (IT) literature that is founded on the basis of the Maximum Entropy (ME) principle (Jaynes [3,4]) and on Shannon's [5] information measure (entropy) as well as other generalized entropy measures. To understand the relationship between the familiar linear statistical model and the approach we take here, we now briefly define our basic problem, discuss its traditional solution and provide the basic logic and related literature we use here in order to solve that problem such that our objectives are achieved.

Consider the basic (linear) problem of estimating the K -dimensional location parameter vector (signal, input) β given an N -dimensional observed sample (response) vector y and an $N \times K$ design (transfer) matrix X such that $y = X\beta + \epsilon$ and ϵ is an N -dimensional random vector such that $E[\epsilon] = 0$ and with some positive definite covariance matrix with a scale parameter σ^2 . The statistical nature of the unobserved noise term is supposed to be known, and we suppose that the second moments of the noise are finite. The researcher's objective is to estimate the unknown vector β with minimal assumptions on ϵ . Recall that under the traditional regularity conditions for the linear model (and for X of rank K), the least squares, (LS), unconstrained, estimator is $\hat{\beta}_{LS} = (X'X)^{-1}X'y$ and $\hat{\beta}_{LS} \sim N\left(\beta, \sigma^2(X'X)^{-1}\right)$ where " $'$ " stands for transpose.

Consider now the problem of estimating β and ϵ simultaneously while imposing minimal assumptions on the likelihood structure and while incorporating certain constraints on the signal and perhaps on the noise. Further, rather than following the tradition of employing point estimators, consider estimating the empirical distribution of the unknown quantities β_k and ϵ_n with the joint objectives of maximizing the in-and-out of sample prediction.

With these objectives, the problem is inherently under-determined and cannot be solved with the traditional least squares or likelihood approaches. Therefore, one must resort to a different principle. In the work done here, we follow the Maximum Entropy (ME) principle that was developed by Jaynes [3,4] for similar problems. The classical ME method consists of using a variational method to choose a probability distribution from a class of probability distributions having pre-assigned generalized moments.

In more general terms, consider the problem of estimating an unknown discrete probability distribution from a finite and possibly noisy set of observed generalized (sample) moments, that is, arbitrary functions of the data. These moments (and the fact that the distribution is proper) are supposed to be the only available information. Regardless of the level of noise in these observed moments, if the dimension of the unknown distribution is larger than the number of observed moments, there are infinitely many proper probability distributions satisfying this information. Such a problem is called an under-determined problem. Which one of the infinitely many solutions that satisfy the data should one choose? Within the class of information-theoretic (IT) methods, the chosen solution is the one that maximizes an information criterion-entropy. Procedure that we propose below to solve the estimation problem described above, fits in that framework.

We construct our proposed estimator for solving the noisy, inverse, linear problem in two basic steps. In our first step, each unknown parameter (β_k and ε_n) is constructed as the expected value of a certain random variable. That is, we view the possible values of the unknown parameters as values of random variables whose distributions are to be determined. We will assume that the range of each such random variable contains the true unknown value of β_k and ε_n respectively. This step actually involves two specifications. The first one is the pre-specified support space for the two sets of parameters (finite/infinite and/or bounded/unbounded). At the outset of section two we shall do this as part of the mathematical statement of the problem. Any further information we may have about the parameters is incorporated into the choice of a prior (reference) measure on these supports. Since usually a model for the noise is supposed to be known, the statistical nature of the noise is incorporated at this stage. As far as the signal goes, this is an auxiliary construction. This constitutes our second specification.

In our second step, because minimal assumptions on the likelihood implies that such a problem is under-determined, we resort to the ME principle. This means that we need to convert this under-determined problem to a well-posed, constrained optimization. Similar to the classical ME method the objective function in that constrained optimization problem is composed of $N \times K$ entropy functions: one for each one of the $N \times K$ proper probability distributions (one for each signal β_k and one for each noise component ε_n). The constraints are just the observed information (data) and the requirement that all probability distributions are proper. Maximizing (simultaneously) the $N \times K$ entropies subject to the constraints yields the desired solution. This optimization yields a unique solution in terms of a unique set of proper probability distribution which in turn yields the desired point estimates β_k and ε_n . Once the constrained model is solved, we construct the concentrated (unconstrained) model. In the method proposed here, we also allow introduction of different priors corresponding to one's beliefs about the data generating process and the structure of the unknown β 's.

Our proposed estimator is a member of the IT family of estimators. The members of this family of estimators include the Empirical Likelihood (EL), the Generalized EL (GEL), the Generalized Method

of Moments (GMM), the Bayesian Method of Moments, (BMOM), the Generalized Maximum Entropy (GME), and the Maximum Entropy in the Mean (MEM), and are all related to the classical Maximum Entropy, ME. (e.g., Owen [6,7]; Qin and Lawless [8]; Smit, [9]; Newey and Smith [10]; Kitamura and Stutzer [11]; Imbens *et al.* [12]; Zellner [13,14]; Zellner and Tobias [15]; Golan, Judge and Miller [1]; Gamboa and Gassiat [16]; Gzyl [17]; Golan and Gzyl [18]). See also Gzyl and Velásquez [2], which builds upon Golan and Gzyl [18] where the synthesis was first proposed. If, in addition, the data are ill-conditioned, one often has to resort to the class of regularization methods (e.g., Hoerl and Kennard [19] O’Sullivan [20], Breiman [21], Tibshirani [22], Titterton [23], Donoho *et al.* [24]; Besnerais *et al.* [25]. A reference for regularization in statistics is Bickel and Li [26]. If some prior information on the data generation process or on the model is available, Bayesian methods are often used. For a detailed review and synthesis of the IT family of estimators, historical perspective and synthesis, see Golan [27]. For other background and related entropy and IT methods of estimation see the special volume of *Advances in Econometrics* (Fomby and Hill [28]) and the two special issues of the *Journal of Econometrics* [29,30]. For additional mathematical background see Mynbaev [31] and Asher, Borchers and Thurber [32].

Our proposed generic IT method will provide us with an estimator for the parameters of the linear statistical model that reconciles some of the objectives achieved by each one of the above methods. Like the philosophy behind the EL, we do not assume a pre-specified likelihood, but rather recover the (natural) weight of each observation via the optimization procedure (e.g., Owen [7]; Qin and Lawless [8]). Similar to regularization methods used for ill-behaved data, we follow the GME logic and use here the pre-specified support space for each one of the unknown parameters as a form of regularization (e.g., Golan, Judge and Miller [1]). The estimated parameters must fall within that space. However, unlike the GME, our method allows for infinitely large support spaces and continuous prior distributions. Like Bayesian approaches, we do use prior information. But we use these priors in a different way—in a way consistent with the basics of information theory and in line with the Kullback–Liebler entropy discrepancy measure. In that way, we are able to combine ideas from the different methods described above that together yield an efficient and consistent IT estimator that is statistically and computationally efficient and easy to apply.

In Section 2, we lay out the basic formulation and then develop our basic model. In Section 3, we provide a detailed closed form examples of the normal priors’ case and other priors. In Section 4 we develop the basic statistical properties of our estimator including first order approximation. In Section 5, we compare our method with Least Squares, regularization and Bayesian methods, including the Bayesian Method of Moments. The comparisons are done under the normal priors. An additional set of analytical examples, providing the formulation and solution of four basic priors (bounded, unbounded and a combination of both) is developed in Section 6. In Section 7, we comment on model comparison. We provide detailed closed form formulations for that section in an Appendix. We conclude in Section 8. The Appendices provide the proofs and detailed analytical formulations.

2. Problem Statement and Solution

2.1. Notation and Problem Statement

Consider the linear statistical model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\beta} \in C_s \quad (1)$$

where $\boldsymbol{\beta} \in \mathbb{R}^K$ is an unknown K -dimensional signal vector that cannot be directly measured but is required to satisfy some convex constraints expressed as $\boldsymbol{\beta} \in C_s$ where C_s is a closed convex set. For example, $C_s = \{\boldsymbol{\beta} \in \mathbb{R}^K \mid \beta_k \in [\underline{z}_k, \bar{z}_k]; k=1, \dots, K\}$ with constants $\underline{z}_k < \bar{z}_k$. (These constraints may come from constraints on $\beta_k = \frac{\partial E[y]}{\partial x_k}$, and may have a natural reason for being imposed). $\mathbf{X}$ is an $N \times K$ known linear operator (design matrix) that can be either fixed or stochastic, $\mathbf{y} \in \mathbb{R}^N$ is a vector of noisy observations, and $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ is a noise vector. Throughout this paper we assume that the components of the noise vector $\boldsymbol{\varepsilon}$ are i.i.d. random variables with zero mean and a variance σ^2 with respect to a probability law $dQ_n(\mathbf{v})$ on $\mathbb{R}^N$. We denote by Q_s and Q_n the prior probability measures reflecting our knowledge about $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ respectively.

Given the indirect noisy observations $\mathbf{y}$, our objective is to simultaneously recover $\boldsymbol{\beta}^* \in \mathbb{R}^K$ and the residuals $\boldsymbol{\varepsilon}^* \in \mathbb{R}^N$ so that Equation (1) holds. For that, we convert problem (1) into a generalized moment problem and consider the estimated $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ as expected values of random variables $\mathbf{z}$ and $\mathbf{v}$ with respect to an unknown probability law P . Note that $\mathbf{z}$ is an auxiliary random variable whereas $\mathbf{v}$ is the actual model for the noise perturbing the measurements. Formally:

Assumption 2.1. The range of $\mathbf{z}$ is the constraint set C_s embodying the constraints that the unknown $\boldsymbol{\beta}$ is to satisfy. Similarly, we assume that the range of $\mathbf{v}$ is a closed convex set C_n where “s” and “n” stand for signal and noise respectively. Unless otherwise specified, and in line with tradition, it is assumed that $\mathbf{v}$ is symmetric about zero.

Comment. It is reasonable to assume that C_n is convex and symmetric in $\mathbb{R}^N$. Further, in some cases the researcher may know the statistical model of the noise. In that case, this model should be used. As stated earlier, Q_s and Q_n are the prior probability measures for $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ respectively. To ensure that the expected values of $\mathbf{z}$ and $\mathbf{v}$ fall in $C = C_s \times C_n$ we need the following assumption.

Assumption 2.2. The closures of the convex hulls of the supports of Q_s and Q_n are respectively C_s and C_n and we set $dQ = dQ_s \times dQ_n$.

Comment. This assumption implies that for any strictly positive density $\rho(\mathbf{z}, \mathbf{v})$ we have:

$$\int \mathbf{z} \rho(\mathbf{z}, \mathbf{v}) dQ_s(\mathbf{z}) dQ_n(\mathbf{v}) \in C_s \quad \text{and} \quad \int \mathbf{v} \rho(\mathbf{z}, \mathbf{v}) dQ_s(\mathbf{z}) dQ_n(\mathbf{v}) \in C_n$$

To solve problems like (1) with minimal assumptions one has to (i) incorporate some prior knowledge, or constraints, on the solution, or (ii) specify a certain criterion to choose among the infinitely many solutions, or (iii) use both approaches. The different criteria used within the IT methods are all directly related to the Shannon’s information (entropy) criterion (Golan [33]). The criterion used in the

method developed and discussed here is the Shannon's entropy. For a detailed discussion and further background see for example the two special issues of the *Journal of Econometrics* [29,30].

2.2. The Solution

In what follows we explain how to transform the original linear problem into a generalized moment problem, or how to transform any constrained linear model like (1) into a problem consisting of finding an unknown density.

Instead of searching directly for the point estimates $(\beta, \varepsilon)^t$ we view it as the expected value of auxiliary random variables $(z, v)^t$ that take values in the convex set $C_s \times C_n$ distributed according to some unknown auxiliary probability law $dP(z, v)$. Thus:

$$\begin{pmatrix} \beta \\ \varepsilon \end{pmatrix} = E_P \left[\begin{pmatrix} z \\ v \end{pmatrix} \right] \quad (2)$$

where E_P denotes the expected value with respect to P .

To obtain P , we introduce the reference measure $dQ(z, v) = dQ_s(z) dQ_n(v)$ on the Borel subsets of the product space $C = C_s \times C_n$. Again, note that while C is binding, Q_s describes one's own belief/knowledge on the unknown β , whereas Q_n describes the actual model for ε . With the above specification, problem (1) becomes:

Problem (1) restated: We search for a density $\rho(z, v)$ such that $dP = \rho dQ$ is a probability law on C and the linear relations:

$$y = X E_P[z] + E_P[v] \quad (3)$$

are satisfied, where:

$$E_P[z] = \int_C z \rho(z, v) dQ(z, v) \text{ and } E_P[v] = \int_C v \rho(z, v) dQ(z, v).$$

Under this construction, $\beta^* = E_{P^*}[z]$ is a random estimator of the unknown parameter vector β and $\varepsilon^* = E_{P^*}[v]$ is an estimator of the noise.

Comment. Using $dQ(z, v) = dQ_s(z) dQ_n(v)$ amounts to assuming an a priori independence of signal and noise. This is a natural assumption as the signal part is a mathematical artifact and the noise part is the actual model of the randomness/noise.

There are potentially many candidates ρ 's that satisfy (3). To find one (the least informative one given the data), we set up the following variational problem: Find $\rho^*(z, v)$ that maximizes the entropy functional, $S_Q(\rho)$ defined by:

$$S_Q(\rho) = - \int_C \rho(z, v) \ln \rho(z, v) dQ(z, v) \quad (4)$$

on the following admissible class of densities:

$$P(C) = \{\rho: C \rightarrow [0, \infty) \mid dP = \rho dQ \text{ is a proper probability satisfying (3)}\} \quad (5)$$

where “ $\ln$ ” stands for the natural logarithm. As usual we extend $x \ln x$ as 0 to $x = 0$. If the maximization problem has a solution, the estimates satisfy the constraints and Equations (1) or (3). The familiar and classical answer to the problem of finding such a ρ^* is expressed in the following lemma.

Lemma 2.1. Assume that ρ is any positive density with respect to dQ and that $\ln \rho$ is integrable with respect to $dP = \rho dQ$, then $S_Q(P) < 0$.

Proof. By the concavity of the logarithm and Jensen’s inequality it is immediate to verify that:

$$S_Q(P) = -E_P \left[\ln \left(\frac{dP}{dQ} \right) \right] = E_P \left[\ln \left(\frac{dQ}{dP} \right) \right] \leq \ln E_P \left[\left(\frac{dQ}{dP} \right) \right] = 0 \quad (6)$$

Before applying this result to our model, we define $\mathbf{A} = [\mathbf{X} \ \mathbf{I}]$ as an $N \times (K + N)$ matrix obtained from juxtaposing $\mathbf{X}$ and the $N \times N$ identity matrix $\mathbf{I}$. We now work with the matrix $\mathbf{A}$ which allows us to consider the larger space rather than just the more traditional moment space. This is shown and discussed explicitly in the examples and derivations of Sections 4–6. For practical purposes, when facing a relatively small sample, the researcher may prefer working with $\mathbf{A}$, rather than with the sample moments. This is because for finite sample the total information captured by using $\mathbf{A}$ is larger than when using the sample’s moments.

To apply lemma (1) to our model, let ρ be any member of the exponential (parametric) family:

$$\rho(\boldsymbol{\lambda}, \mathbf{z}, \mathbf{v}) = \Omega(\boldsymbol{\lambda})^{-1} \exp(-\langle \boldsymbol{\lambda}, \mathbf{A}\boldsymbol{\xi} \rangle) \quad (7)$$

where $\boldsymbol{\xi}' = (\mathbf{z}, \mathbf{v})'$, $\langle \mathbf{a}, \mathbf{b} \rangle$ denotes the Euclidean scalar (inner) product of vectors $\mathbf{a}$ and $\mathbf{b}$, and $\boldsymbol{\lambda} \in \mathbb{R}^N$ are N free parameters that will play the role of Lagrange multipliers (one multiplier for each observation). The quantity $\Omega(\boldsymbol{\lambda})$ is the normalization function:

$$\Omega(\boldsymbol{\lambda}) = \int_C \exp(-\langle \boldsymbol{\lambda}, \mathbf{A}\boldsymbol{\xi} \rangle) dQ(\boldsymbol{\xi}) = \omega(\mathbf{A}'\boldsymbol{\lambda}) \quad (8)$$

where:

$$\omega(\boldsymbol{\tau}) = \int_C \exp(-\langle \boldsymbol{\tau}, \boldsymbol{\xi} \rangle) dQ(\boldsymbol{\xi}) = \int_{C_s} \exp(-\langle \boldsymbol{\tau}_s, \mathbf{z} \rangle) dQ_s(\mathbf{z}) \int_{C_n} \exp(-\langle \boldsymbol{\tau}_n, \mathbf{v} \rangle) dQ_n(\mathbf{v}) = \omega_s(\boldsymbol{\tau}_s) \omega_n(\boldsymbol{\tau}_n)$$

is the Laplace transform of Q . Next taking logs in (7) and defining:

$$\sum(\boldsymbol{\lambda}) = \ln \Omega(\boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \mathbf{y} \rangle \quad (9)$$

Lemma 2.1 implies that $\sum(\boldsymbol{\lambda}) \geq S_Q(\rho)$ for any $\boldsymbol{\lambda} \in \mathbb{R}^N$ and for any ρ in the class of probability laws $P(C)$ defined in (5). However, the problem is that we do not know whether the solution $\rho^*(\boldsymbol{\lambda}, \mathbf{z}, \mathbf{v})$ is a member of $P(C)$ for some $\boldsymbol{\lambda}$. Therefore, we search for $\boldsymbol{\lambda}^*$ such that $\rho^* = \rho(\boldsymbol{\lambda}^*)$ is in $P(C)$ and $\boldsymbol{\lambda}^*$ is a minimum. If such a $\boldsymbol{\lambda}^*$ is found, then we would have found a density (the unique one, for S_Q is strictly convex in ρ) that maximizes the entropy, and by using the fact that $\boldsymbol{\beta}^* = E_{\rho^*}[\mathbf{z}]$ and $\boldsymbol{\varepsilon}^* = E_{\rho^*}[\mathbf{v}]$, the solution to (1), which is consistent with the data (3), is found. Formally, the result is contained in the following theorem. (Note that the Kullback’s measure (Kullback [34]), is a particular case of $S_Q(P)$, with a sign change and when both P and Q have densities).

Theorem 2.1. Assume that $D(\Omega) = \{\lambda \in \mathbb{R}^N \mid \Omega(\lambda) < \infty\}$ has a non-empty interior and that the minimum of the (convex) function $\sum(\lambda)$ is achieved at λ^* . Then, $dP^*(\xi) = \rho(\lambda^*, \xi) dQ(\xi)$ satisfies the set of constraints (3) or (1) and maximizes the entropy.

Proof. Consider the gradient of $\sum(\lambda)$ at λ^* . The equation to be solved to determine λ^* is $-\nabla_{\lambda} \ln \Omega(\lambda) = y$, which coincides with Equation (3) when the gradient is written out explicitly.

Note that this is equivalent to minimizing (9) which is the concentrated likelihood-entropy function. Notice as well that $\sum(\lambda^*) = S_Q(\rho^*)$.

Comment. This theorem is practically equivalent to representing the estimator in terms of the estimating equations. Estimation equations (or functions) are the underlying equations from which the roots or solutions are derived. The logic for using these equations is (i) they have simpler form (e.g., a linear form for the LS estimator) than their roots, and (ii) they preserve the sampling properties of their roots (Durbin, [35]). To see the direct relationship between estimation equations and the dual/concentrated model (extremum estimator), note that the estimation equations are the first order conditions of the respective extremum problem. The choice of estimation equations is appropriate whenever the first order conditions characterize the global solution to the (extremum) optimization problem, which is the case in the model discussed here.

Theorem 2.1 can be summarized as follows: in order to determine β and ε from (1), it is easier to transform the algebraic problem into the problem of obtaining a minimum of the convex function $\sum(\lambda)$, and then use $\beta^* = E_{P^*}[z]$ and $\varepsilon^* = E_{P^*}[v]$ to compute the estimates β^* and ε^* . The above procedure is designed in such a way that $\beta^* \in C_s$ is automatically satisfied. Since the actual measurement noise is unknown, it is treated as a quantity to be determined, and treated (mathematically) as if both β and ε were unknown. The interpretations of the reconstructed residual ε^* and the reconstructed β^* , are different. The latter is the unknown parameter vector we are after while the former is the residual (reconstructed error) such that the linear Equation (1), $y = X\beta^* + \varepsilon^*$, is satisfied. With that background, we now discuss the basic properties of our model. For a detailed comparison of a large number of IT estimation methods see Golan ([27–33]) and the nice text of Mittelhammer, Judge and Miller [36]

3. Closed Form Examples

With the above formulation, we now turn to a number of relatively simple analytical examples. These examples demonstrate the advantages of our method and its simplicity. In Section 6 we provide additional closed form examples.

3.1. Normal Priors

In this example the index d takes the possible values (dimensions) K , N , or $K+N$ depending if it relates to C_s (or z), to C_n (or v) or to both. Assume the reference prior dQ is a normal random vector with $d \times d$ [i.e., $K \times K$, $N \times N$ or $(N+K) \times (N+K)$] covariance matrix D , the law of which has

density $(2\pi)^{-d/2} (\det \mathbf{D})^{-1/2} \exp\left(-\frac{1}{2} \langle (\mathbf{c} - \mathbf{c}^0), \mathbf{D}^{-1} (\mathbf{c} - \mathbf{c}^0) \rangle\right)$ where $\mathbf{c}^0 = \begin{pmatrix} \mathbf{z}^0 \\ \mathbf{v}^0 \end{pmatrix}$ is the vector of prior means

and is specified by the researcher. Next, we define the Laplace transform, $\omega(\boldsymbol{\tau})$, of the normal prior. This transform involves the diagonal covariance matrix for the noise and signal models:

$$\omega(\boldsymbol{\tau}) = \exp\left\{\frac{1}{2} \langle \boldsymbol{\tau}, \mathbf{D} \boldsymbol{\tau} \rangle + \langle \boldsymbol{\tau}, \mathbf{c}^0 \rangle\right\}. \quad (10)$$

Since $\ln \omega(\boldsymbol{\tau}) = \frac{1}{2} \langle \boldsymbol{\tau}, \mathbf{D} \boldsymbol{\tau} \rangle + \langle \boldsymbol{\tau}, \mathbf{c}^0 \rangle$, then replacing $\boldsymbol{\tau}$ by either $X^t \boldsymbol{\lambda}$ or by $\boldsymbol{\lambda}$, (for the noise vector) verifies that $\Omega(\boldsymbol{\lambda})$ turns out to be of a quadratic form, and therefore the problem of minimizing $\Sigma(\boldsymbol{\lambda})$ is just a quadratic minimization problem. In this case, no bounds are specified on the parameters. Instead, normal priors are used.

From (10) we get the concentrated model:

$$\sum(\boldsymbol{\lambda}) = \ln \Omega(\boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \mathbf{y} \rangle = \langle \boldsymbol{\lambda}, \mathbf{A} \mathbf{D} \mathbf{A}^t \boldsymbol{\lambda} \rangle / 2 + \langle \boldsymbol{\lambda}, (\mathbf{y} - \mathbf{A} \mathbf{c}^0) \rangle \quad (11)$$

with a minimum at $\boldsymbol{\lambda}^*$, satisfying:

$$\mathbf{M} \boldsymbol{\lambda}^* \equiv \mathbf{A} \mathbf{D} \mathbf{A}^t \boldsymbol{\lambda}^* = -(\mathbf{y} - \mathbf{A} \mathbf{c}^0) \quad (12)$$

If $\mathbf{M}^\#$ denotes the generalized inverse of $\mathbf{M} = \mathbf{A} \mathbf{D} \mathbf{A}^t$, then $\boldsymbol{\lambda}^* = -\mathbf{M}^\# (\mathbf{y} - \mathbf{A} \mathbf{c}^0)$ and therefore:

$$\boldsymbol{\beta}^* = \mathbf{c}^0 + \mathbf{D} \mathbf{A}^t \boldsymbol{\lambda}^* = \mathbf{c}^0 - \mathbf{D} \mathbf{A}^t \mathbf{M}^\# (\mathbf{y} - \mathbf{A} \mathbf{c}^0). \quad (13)$$

For the general case $\mathbf{A} = [\mathbf{X} \ \mathbf{I}]$ and:

$$\mathbf{D} = \begin{pmatrix} \text{Cov}(\mathcal{Q}_s) & \mathbf{0} \\ \mathbf{0} & \text{Cov}(\mathcal{Q}_n) \end{pmatrix} \equiv \begin{pmatrix} \mathbf{D}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_n \end{pmatrix} \equiv \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix},$$

the generalized entropy solution for the traditional linear model is:

$$\mathbf{A} \mathbf{D} \mathbf{A}^t = \mathbf{X} \mathbf{D}_s \mathbf{X}^t + \mathbf{D}_n$$

so:

$$E_{P^*} \begin{pmatrix} \mathbf{z} \\ \mathbf{v} \end{pmatrix} = E_{P^*}(\boldsymbol{\xi}) = \begin{pmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\varepsilon}^* \end{pmatrix} = \mathbf{c}_0 + \mathbf{D} \mathbf{A}^t \mathbf{M}^{-1} (\mathbf{y} - \mathbf{A} \mathbf{c}_0) = \boldsymbol{\xi}^*,$$

and finally:

$$\begin{aligned} \boldsymbol{\beta}^* &= \mathbf{z}^0 + \mathbf{D}_1 \mathbf{X}^t \mathbf{M}^{-1} \mathbf{B} \\ \boldsymbol{\varepsilon}^* &= \mathbf{v}^0 + \mathbf{D}_2 \mathbf{M}^{-1} \mathbf{B}. \end{aligned} \quad (14)$$

Here $\mathbf{B} = (\mathbf{y} - \mathbf{A} \mathbf{c}_0)$. See Appendix 2 for a detailed derivation.

3.2. Discrete Uniform Priors — A GME Model

Consider now the uniform priors, which is basically the GME method (Golan, Judge and Miller [1]). Jaynes's classical ME estimator (Jaynes [3,4]) is a special case of the GME. Let the components of $\mathbf{z}$ take discrete values, and let $C_s^k = \{z_{k1}, \dots, z_{kM(k)}\}$ for $1 \leq k \leq K$. Note that we allow for the cardinality of

each of these sets to vary. Next, define $C_s = C_s^1 \times \dots \times C_s^K$. A similar construction may be proposed for the noise terms, namely we put $C_n = C_n^1 \times \dots \times C_n^N$. Since the spaces are discrete, the information is described by the obvious σ -algebras and both the prior and post-data measures will be discrete. As a prior on the signal space, we may consider:

$$Q_s(z_{1,k_1}, \dots, z_{K,M(K)}) = Q_s^1(z_{1,k_1}) \dots Q_s^K(z_{K,M(K)}) = q_{1,k_1}^s \dots q_{K,M(K)}^s,$$

where a similar expression may be specified for the priors on C_n . Finally, we get:

$$\omega_s(\tau) = \prod_{j=1}^K \left(\sum_{k=1}^{M(j)} e^{-\tau_j z_{jk}} q_{jk} \right),$$

together with a similar expression for the Laplace transform of the noise prior. Notice that since the noise and signal are independent in the priors, this is also true for the post-data, so:

$$P(\mathbf{z}, \mathbf{v}) = \rho(\mathbf{z}, \mathbf{v}) Q(\mathbf{z}, \mathbf{v}) = e^{-\langle \lambda^*, \mathbf{xz} \rangle - \langle \lambda^*, \mathbf{v} \rangle} Q(\mathbf{z}, \mathbf{v}) = P_s(\mathbf{z}) P_n(\mathbf{v}) = \prod_{jn} p_{jn}^s \prod_{lm} p_{lm}^n.$$

Finally, $\beta = E_{P_s^*}[\mathbf{z}]$ and $\varepsilon = E_{P_n^*}[\mathbf{v}]$. For detailed derivations and discussion of the GME see Golan, Judge and Miller [1].

3.3. Signal and Noise Bounded Above and Below

Consider the case in which both β and ε are bounded above and below. This time we place a Bernoulli measure on the constraint space C_s and the noise space C_n . Let $C_s = \prod_{j=1}^K [a_j, b_j]$ and $C_n = \prod_{l=1}^N [-e, e]$ for the signal and noise bounds a_j, b_j and e respectively. The Bernoulli a priori measure on $C = C_s \times C_n$ is:

$$dQ(\xi) = \prod_{j=1}^K dQ_j(z_j) \prod_{l=1}^N dQ_l(v_l) = \prod_{j=1}^K (p_j \delta_{a_j}(dz_j) + q_j \delta_{b_j}(dz_j)) \prod_{l=1}^N \left(\frac{1}{2} \delta_{-e}(dv_l) + \frac{1}{2} \delta_e(dv_l) \right),$$

where $\delta_c(dz)$ denotes the (Dirac) unit point mass at some point c . Recalling that $\mathbf{A} = [\mathbf{X}, \mathbf{I}]$ we now compute the Laplace transform $\omega(\mathbf{t})$ of Q , which in turn yields $\Omega(\lambda) = \omega(\mathbf{A}^T \lambda)$:

$$\Omega(\lambda) = \prod_{j=1}^K (p_j e^{-(\mathbf{x}^T \lambda)_j a_j} + q_j e^{-(\mathbf{x}^T \lambda)_j b_j}) \prod_{j=1}^N \frac{1}{2} (e^{\lambda_j e} + e^{-\lambda_j e}).$$

The concentrated entropy function is:

$$\sum(\lambda) = \ln \Omega(\lambda) + \langle \lambda, \mathbf{y} \rangle = \sum_{j=1}^K \ln (p_j e^{-(\mathbf{x}^T \lambda)_j a_j} + q_j e^{-(\mathbf{x}^T \lambda)_j b_j}) + \sum_{j=1}^N \ln \frac{1}{2} (e^{\lambda_j e} + e^{-\lambda_j e}) + \langle \lambda, \mathbf{y} \rangle.$$

The minimizer of this function is the Lagrange multiplier vector λ^* . Once it has been found, then $\beta^* = -\nabla_{\tau} \ln \omega_s(\tau) \Big|_{\tau=X^T \lambda^*}$ and $\varepsilon^* = -\nabla_{\tau} \ln \omega_n(\tau) \Big|_{\tau=\lambda^*}$. Explicitly:

$$\beta_j^* = \frac{\partial}{\partial t_j} \sum \ln(p_l e^{-t_l a_l} + q_l e^{-t_l b_l}) = a_j P_j + b_j Q_j$$

where:

$$P_j = p_j e^{-t_j a_j} / (p_j e^{-t_j a_j} + q_j e^{-t_j b_j}) \text{ and } Q_j = q_j e^{-t_j b_j} / (p_j e^{-t_j a_j} + q_j e^{-t_j b_j})$$

and $\tau = (X^T \lambda^*)$. Similarly:

$$\varepsilon_l^* = -e \tilde{P}_l + e_l \tilde{Q}_l$$

where:

$$\tilde{P}_l = e^{-\lambda_l^* e} / (e^{-\lambda_l^* e} + e^{\lambda_l^* e}) \text{ and } \tilde{Q}_l = e^{\lambda_l^* e} / (e^{-\lambda_l^* e} + e^{\lambda_l^* e}).$$

These are respectively the Maximum Entropy probabilities that the auxiliary random variables z_j will attain the values a_j or b_j , or the auxiliary random variables v_l describing the error terms attain the values $\pm e$. These can be also obtained as the expected values of $\mathbf{v}$ and $\mathbf{z}$ with respect to the post-data measure $P^*(\lambda, d\xi)$ given by:

$$P^*(\lambda, d\xi) = \prod_{j=1}^K (P_j \delta_{a_j}(dz_j) + Q_j \delta_{b_j}(dz_j)) \prod_{l=1}^N (\tilde{P}_l \delta_{-e}(dv_l) + \tilde{Q}_l \delta_e(dv_l)).$$

Note that this model is the continuous version of the discrete GME model described earlier.

4. Main Results

4.1. Large Sample Properties

In this section we develop the basic statistical results. In order to develop these results for our generic IT estimator, we needed to employ tools that are different than the standard tools used for developing asymptotic theories (e.g., Mynbaev [31] or in Mittelhammer *et al.* [36]).

4.1.1. Notations and First Order Approximation

Denote by β_N^* the estimator of the true β when the sample size is N . Throughout this section we add a subscript N to all quantities introduced in Section 2 to remind us that the size of the data set is N . We want to show that $\beta_N^* \rightarrow \beta$ and $\sqrt{N}(\beta_N^* - \beta) \rightarrow N(\mathbf{0}, \mathbf{V})$ as $N \rightarrow \infty$ in some appropriate way (for some covariance $\mathbf{V}$). We state here the basic notations, assumptions and results and leave the details to the Appendix. The problem is that when N varies, we are dealing with problems of different sizes (recall λ is of dimension N in our generic model). To turn all problems to the same size let:

$$\tilde{\mathbf{y}}_N = \frac{1}{N} X_N' \mathbf{y}_N = \frac{1}{N} X_N' X_N \beta + \frac{1}{N} X_N' \varepsilon_N := W_N \beta + \frac{1}{N} X_N' \varepsilon_N. \quad (15)$$

The modified data vector and the modified error terms are K -dimensional (moment) vectors, and the modified design matrix is a $K \times K$ -matrix. Problem (15), call it the moment, or the stochastic moment, problem, can be solved using the above generic IT approach which reduces to minimizing the modified concentrated (dual) entropy function:

$$\tilde{\Sigma}_N(\boldsymbol{\mu}) = \ln \tilde{\Omega}_N(\boldsymbol{\mu}) + \langle \tilde{\mathbf{y}}_N, \boldsymbol{\mu} \rangle$$

where $\boldsymbol{\mu} \in \mathbb{R}^K$ and $\tilde{\Omega}_N(\boldsymbol{\mu}) = \Omega_N(\frac{1}{N} X_N' \boldsymbol{\mu})$.

Assumption 4.1. Assume that there exists an invertible $K \times K$ symmetric and positive definite matrix W such that $W_N := \frac{1}{N} X_N' X_N \rightarrow W$. More precisely, assume that $\|W_N - W\| = o(1/N)$ as $N \rightarrow \infty$.

Assume as well that for any N -vector $\mathbf{v}$, as $N \rightarrow \infty$ $\frac{1}{N} \|X_N' \mathbf{v}\| = o(1/N)$.

Recall that in finite dimensions all norms are equivalent so convergence in any norm is equivalent to component wise convergence. This implies that under Assumption 4.1, the vectors $\frac{1}{N} X_N' \boldsymbol{\varepsilon}_N$ converge to $\mathbf{0}$ in L_2 , therefore in probability. To see the logic for that statement, recall that the vector $\boldsymbol{\varepsilon}_N$ has covariance matrix $\sigma^2 I_N$. Therefore, $\text{Var}(\frac{1}{N} X_N' \boldsymbol{\varepsilon}_N) = \frac{\sigma^2}{N} W_N$ and assumption 4.1 yields the above conclusion. (To keep notations simple, and without loss of generality, we discuss here the case of $\sigma^2 I_N$.)

Corollary 4.1. By Equation (15) $\tilde{\mathbf{y}}_N = W_N \boldsymbol{\beta} + \frac{1}{N} X_N' \boldsymbol{\varepsilon}_N$. Let $\tilde{\mathbf{y}}_\infty = W \boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is the true but unknown vector of parameters. Then, $E[\|\tilde{\mathbf{y}}_N - \tilde{\mathbf{y}}_\infty\|^2] \rightarrow 0$ as $N \rightarrow \infty$ (the proof is immediate).

Lemma 4.1. Under Assumption 4.1 and assume that for real a $\int_{C_n} \exp(a\|v\|) dQ_{n,N}(v) < \infty$. Then, for

$\boldsymbol{\mu} \in \mathbb{R}^K$, $\tilde{\Omega}_{n,N}(\boldsymbol{\mu}) = \int_{C_n} \exp(-\langle \boldsymbol{\mu}, \frac{1}{N} X_N' \mathbf{v} \rangle) dQ_{n,N}(\mathbf{v}) \rightarrow 1$ as $N \rightarrow \infty$. Equivalently, $\frac{1}{N} X_N' \boldsymbol{\varepsilon}_N \rightarrow 0$ as $N \rightarrow \infty$ weakly in $\mathbb{R}^K$ with respect to the appropriate induced measure.

Proof of lemma 4.1. Note that for $\boldsymbol{\mu} \in \mathbb{R}^K$:

$$\tilde{\Omega}_{n,N}(\boldsymbol{\mu}) = \int_{C_n} \exp\left(-\left\langle \boldsymbol{\mu}, \frac{1}{N} X_N' \mathbf{v} \right\rangle\right) dQ_{n,N}(\mathbf{v}) \rightarrow 1 \text{ as } N \rightarrow \infty.$$

This is equivalent to the assertion of the lemma.

Lemma 4.2. Let $\tilde{\mathbf{y}}_\infty = W \boldsymbol{\beta}$. Then, under Assumption 4.1:

$$\tilde{\Omega}_N(\boldsymbol{\mu}) \rightarrow \tilde{\Omega}_\infty(\boldsymbol{\mu}) := \int_{C_s} \exp(-\langle \boldsymbol{\mu}, W \mathbf{z} \rangle) dQ_s(\mathbf{z}).$$

Comment. Observe that the $\boldsymbol{\mu}^*$ that minimizes $\tilde{\Sigma}_\infty(\boldsymbol{\mu}) = \ln \tilde{\Omega}_\infty(\boldsymbol{\mu}) + \langle \tilde{\mathbf{y}}_\infty, \boldsymbol{\mu} \rangle$ satisfies:

$$\tilde{\mathbf{y}}_\infty = W \int_{C_s} \frac{\mathbf{z}}{\tilde{\Omega}_\infty(\boldsymbol{\mu}^*)} \exp\left(-\langle \boldsymbol{\mu}^*, W\mathbf{z} \rangle\right) dQ_s(\mathbf{z}).$$

Or since W is invertible, $\boldsymbol{\beta}$ admits the representation

$$\boldsymbol{\beta} = \int_{C_s} \frac{\mathbf{z}}{\tilde{\Omega}_\infty(\boldsymbol{\mu}^*)} \exp\left(-\langle \boldsymbol{\mu}^*, W\mathbf{z} \rangle\right) dQ_s(\mathbf{z}).$$

Note that this last identity can be written as

$$\boldsymbol{\beta} = -\nabla_{\boldsymbol{\tau}} \ln \omega_s(\boldsymbol{\tau})|_{\boldsymbol{\tau}=W\boldsymbol{\mu}^*}$$

Next, we define the function:

$$\boldsymbol{\tau} \in \mathbb{R}^K \rightarrow \boldsymbol{\theta}(\boldsymbol{\tau}) = -\nabla_{\boldsymbol{\tau}} \ln \omega_s(\boldsymbol{\tau}) \in \mathbb{R}^K.$$

Assumption 4.2. The function $\boldsymbol{\theta}(\boldsymbol{\tau})$ is invertible and continuously differentiable.

Observe that we also have $\boldsymbol{\beta} = \boldsymbol{\theta}(W\boldsymbol{\mu}^*)$. To relate the solution to problem (1) to that of problem (15), observe that $\tilde{\Omega}_N(\boldsymbol{\mu}) = \Omega_N\left(\frac{1}{N}X_N\boldsymbol{\mu}\right)$ as well as $\tilde{\Sigma}_N(\boldsymbol{\mu}) = \Sigma_N\left(\frac{1}{N}X_N\boldsymbol{\mu}\right)$ where Ω_N and Σ_N are the functions introduced in Section 2 for a problem of size N . To relate the solution of the problem of size K to that of the problem of size N , we have:

Lemma 4.3. If $\boldsymbol{\mu}_N^*$ denotes the minimizer of $\tilde{\Sigma}_N(\boldsymbol{\mu})$ then $\boldsymbol{\lambda}_N^* = \frac{1}{N}X_N\boldsymbol{\mu}_N^*$ is the minimizer of $\Sigma_N(\boldsymbol{\lambda})$.

Proof of Lemma 4.3. Recall that:

$$\tilde{\Omega}_N(\boldsymbol{\mu}) = \int_C e^{-\left\langle \boldsymbol{\mu}, \frac{1}{N}X_N' A \boldsymbol{\xi} \right\rangle} dQ(\boldsymbol{\xi}) = \int_{C_s} e^{-\left\langle \boldsymbol{\mu}, \frac{1}{N}X_N' X_N \mathbf{z} \right\rangle} dQ(\mathbf{z}) \int_{C_n} e^{-\left\langle \boldsymbol{\mu}, \frac{1}{N}X_N' \mathbf{v} \right\rangle} dQ_n(\mathbf{v})$$

From this, the desired result follows after a simple computation.

We write the post data probability that solves problem (15) (or (1)) as:

$$dP_N^o(\boldsymbol{\xi}) = \frac{e^{-\left\langle \boldsymbol{\mu}_N^*, \frac{1}{N}X_N' A \boldsymbol{\xi} \right\rangle}}{\tilde{\Omega}(\boldsymbol{\mu}_N^*)} dQ(\boldsymbol{\xi})$$

Recalling that $dP_N^*(\boldsymbol{\xi})$ is the solution for the N -dimensional (data) problem and $dP_N^0(\boldsymbol{\xi})$ is the solution for the moment problem, we have the following result:

Corollary 4.2. With the notations introduced above and by Lemma 4.3 we have

$$E_{P_N^o}[\boldsymbol{\xi}] = E_{P_N^*}[\boldsymbol{\xi}] = \begin{pmatrix} \boldsymbol{\beta}_N^* \\ \boldsymbol{\varepsilon}_N^* \end{pmatrix}.$$

To state Lemma 4.4 we must consider the functions $\mathbb{R}^K \rightarrow \mathbb{R}^K$ defined by:

$$\boldsymbol{\mu} \rightarrow \varphi_N(\boldsymbol{\mu}) := -\nabla_{\boldsymbol{\mu}} \ln \tilde{\Omega}_N(\boldsymbol{\mu}) \quad \text{and} \quad \boldsymbol{\mu} \rightarrow \varphi_\infty(\boldsymbol{\mu}) := -\nabla_{\boldsymbol{\mu}} \ln \tilde{\Omega}_\infty(\boldsymbol{\mu}).$$

Denote by P_N^μ the measure with density $e^{-\left\langle \boldsymbol{\mu}, \frac{1}{N}X_N' A \boldsymbol{\xi} \right\rangle} / \tilde{\Omega}_N(\boldsymbol{\mu})$ with respect to Q . The invertibility of the functions defined above is related to the non-singularity of their Jacobian matrices, which are the P_N^μ -covariances of $\boldsymbol{\xi}$. These functions will be invertible as long as these quantities are positive definite. The relationship among the above quantities is expressed in the following lemma:

Lemma 4.4. With the notations introduced above and in (8), and recall that we suppose that $D_2 = \sigma^2 I_N$, we have:

$$C_\lambda(\xi, \xi) = \nabla_\tau \nabla_\tau \ln \omega(\tau) \big|_{\tau=A'\lambda} = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}; \quad C_\mu(\xi, \xi) = \nabla_\tau \nabla_\tau \ln \omega(\tau) \big|_{\tau=A'X_N'\lambda/N} = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \text{ and}$$

$$\nabla_\lambda \nabla_\lambda \ln \Omega_N(\lambda) = \mathbf{A} C_\lambda(\xi, \xi) \mathbf{A} = X_N D_1 X_N' + \sigma^2 I_N; \quad \nabla_\mu \nabla_\mu \ln \tilde{\Omega}_N(\mu) = W_N \tilde{D}_1 W_N' + \sigma^2 W_N / N$$

where $\tilde{D}_1 = \theta'(W_N \mu) = -\nabla_\tau \nabla_\tau \ln \omega_s(\tau) \big|_{\tau=W_N \mu}$, $D_1 = \theta'(X_N' \lambda) = -\nabla_\tau \nabla_\tau \ln \omega_s(\tau) \big|_{\tau=X_N' \lambda}$ and θ' is the first derivative of θ .

Comment. The block structure of the covariance matrix results from the independence of the signal and the noise components in both in the prior measure dQ and the post data (maximum entropy) probability measure dP^* .

Following the above, we assume:

Assumption 4.3. The eigenvalues of the Hessian matrix $\nabla_\mu \nabla_\mu \ln \tilde{\Omega}_N(\mu)$ are uniformly (with respect to N and μ) bounded below away from zero.

Proposition 4.1. Let $\psi_N(\mathbf{y}), \psi_\infty(\mathbf{y})$ respectively denote the compositional inverses of $\varphi_N(\mu), \varphi_\infty(\mu)$. Then, as $N \rightarrow \infty$, (i) $\varphi_N(\mu) \rightarrow \varphi_\infty(\mu)$ and (ii) $\psi_N(\mathbf{y}) \rightarrow \psi_\infty(\mathbf{y})$.

The proof is presented in the Appendix.

4.1.2. First Order Unbiasedness

Lemma 4.5. (First Order Unbiasedness). With the notations introduced above and under Assumptions 4.1–4.3, assume furthermore that $\|\psi_N - \psi_\infty\|_\infty = o(1/N)$ as $N \rightarrow \infty$. Then up to $o(1/N)$, β_N^* is an unbiased estimator of β .

The proof is presented in the Appendix.

4.1.3. Consistency

The following lemma and proposition provide results related to the large sample behavior of our generalized entropy estimator. For simplicity of the proof and without loss of generality, we suppose here that $\mathbf{D}_2 = \sigma_2^2 \mathbf{I}_N$.

Lemma 4.6. (Consistency in squared mean). Under the same assumptions of Lemma 4.5, since $E[\varepsilon] = \mathbf{0}$ and the ε are homoskedastic, then $\beta_N^* \rightarrow \beta$ in square mean as $N \rightarrow \infty$.

Next, we provide our main result of convergence in distribution.

Proposition 4.2. (Convergence in distribution). Under the same assumptions as in Lemma 4.5 we have

$$(a) \beta_N^* \xrightarrow{D} \beta \text{ as } N \rightarrow \infty,$$

$$(b) \sqrt{N}(\beta_N^* - \beta) \xrightarrow{D} N(\mathbf{0}, \sigma^2 \mathbf{W}^{-1}) \text{ as } N \rightarrow \infty,$$

where $\xrightarrow{D}$ stands for convergence in distribution (or law).

Both proofs are presented in the Appendix.

4.2. Forecasting

Once the Generalized Entropy (GE) estimated vector β^* has been found, we can use it to predict future (yet) “unobserved” values. If additive noise (ϵ or $\mathbf{v}$) is distributed according to the same prior Q_n , and if future observations are determined by the design matrix $\mathbf{X}_f$, then the possible future observations are described by a random variable $\mathbf{y}_f$ given by $\mathbf{y}_f = \mathbf{X}_f \beta^* + \mathbf{v}_f$. For example, if $\mathbf{v}_f$ is centered (on $\mathbf{0}$), then $E_{Q_n}[\mathbf{y}_f] = \mathbf{X}_f \beta^*$ and:

$$\text{Var}(\mathbf{y}_f) = \text{tr} \left\{ E_{Q_n} (\mathbf{y}_f - \mathbf{X}_f \beta^*) (\mathbf{y}_f - \mathbf{X}_f \beta^*)^t \right\} = \text{tr} E_{Q_n} \mathbf{v}_f \mathbf{v}_f^t = \text{Var}(\mathbf{v}_f).$$

In the next section we contrast our estimator with other estimators. Then, in Section 6 we provide more analytic solutions for different priors.

5. Method Comparison

In this Section we contrast our IT estimator with other estimators that are often used for estimating the location vector β in the noisy, inverse linear problem. We start with the least squares (LS) model, continue with the generalized LS (GLS) and then discuss the regularization method often used for ill-posed problems. We then contrast our estimator with a Bayesian one and with the Bayesian Method of Moments (BMOM). We also show that exact correspondence between our estimator and the other estimators under normal priors.

5.1. The Least Squares Methods

5.1.1. The General Case

We first consider the purely geometric/algebraic approach for solving the linear model (1). A traditional method consists of solving the variational problem:

$$\text{Min}_{\beta} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|^2 \mid \beta \in \mathbb{R}^K \right\} \quad (16)$$

The rationale here is that because of the noise ϵ , the data $\mathbf{y} = \mathbf{X}\beta_{\text{true}} + \epsilon$ may fall outside the range $\mathbf{X}(\mathbb{R}^K) \equiv \{\mathbf{X}\beta : \beta \in \mathbb{R}^K\}$ of $\mathbf{X}$, so the objective is to minimize that discrepancy. The minimizer β_{LS}^* of (16) provides us with the LS estimates that minimize the errors sum of square distance from the data to $\mathbf{X}(\mathbb{R}^K)$. When $(\mathbf{X}'\mathbf{X})^{-1}$ exists, then $\beta_{LS}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$. The reconstruction error $\epsilon_{LS}^* = \mathbf{y} - \mathbf{X}\beta_{LS}^*$ can be thought of as our estimate of the “minimal error in quadratic norm” of the measurement errors, or of the noise present in the measurements.

The optimization (16) can be carried out with respect to different norms. In particular, we could have considered $\|\eta\|_{D_2}^2 = \langle \eta, \mathbf{D}_2^{-1} \eta \rangle$. In this case we get the GLS solution $\beta_{GLS}^* = (\mathbf{X}'\mathbf{D}_2^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}_2^{-1}\mathbf{y}$ for any general (covariance) matrix $\mathbf{D}$ with blocks $\mathbf{D}_1$ and $\mathbf{D}_2$.

If, on the other hand, our objective is to reconstruct simultaneously both the signal and the noise, we can rewrite (1) as:

$$\mathbf{y} = \mathbf{A}\xi \quad (17)$$

where $\mathbf{A}$ and ξ are as defined in Section 2. Since $\mathbf{y} \in \mathbb{R}^N$, $\xi \in \mathbb{R}^{N+K}$ and the matrix $\mathbf{A}$ is of dimension $N \times (N+K)$, there are infinitely many solutions that satisfy the observed data in (1) (or (17)). To choose a single solution we solve the following model:

$$\underset{\xi}{\text{Min}} \left\{ \frac{1}{2} \|\xi\|^2 \mid \mathbf{A}\xi = \mathbf{y} \right\} \quad (18)$$

In the more general case we can incorporate the covariance matrix to weigh the different components of ξ :

$$\underset{\xi}{\text{Min}} \left\{ \frac{1}{2} \|\xi\|_D^2 \mid \mathbf{A}\xi = \mathbf{y} \right\} \quad (19)$$

where $\|\xi\|_D^2 = \langle \xi, \mathbf{D}^{-1}\xi \rangle$ is a weighted norm in the extended signal-noise space ($C = C_s \times C_n$) and $\mathbf{D}$ can be taken to be the full covariance matrix composed of both $\mathbf{D}_1$ and $\mathbf{D}_2$ defined in Section 3.1. Under the assumption that $\mathbf{M} \equiv (\mathbf{A}\mathbf{D}\mathbf{A}^t)^{-1}$ is invertible, the solution to the variational problem (19) is given by $\xi_{GE}^* = \mathbf{D}\mathbf{A}^t (\mathbf{A}\mathbf{D}\mathbf{A}^t)^{-1} \mathbf{y} = \mathbf{D}\mathbf{A}^t \mathbf{M}^{-1} \mathbf{y}$. This solution coincides with our Generalized Entropy formulation when *normal priors* are imposed and are centered about zero ($\mathbf{c}_0 = \mathbf{0}$) as is developed explicitly in Equation (14).

If, on the other hand, the problem is ill-posed (e.g., $\mathbf{X}$ is not invertible), then the solution is not unique, and a combination of the above two methods (16 and 18) can be used. This yields the regularization method consisting of finding β such that:

$$\underset{\beta}{\text{Min}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \frac{\alpha}{2} \|\beta\|^2 \mid \beta \in \mathbb{R}^K \right\} \quad (20)$$

is achieved (see for example, Donoho *et al.* [25] for a nice discussion of regularization within the ME formulation.) Traditionally, the positive penalization parameter α is specified to favor small sized reconstructions, meaning that out of all possible reconstructions with a given discrepancy, those with the smallest norms are chosen. The norms in (20) can be chosen to be weighted, so that the model can be generalized to:

$$\underset{\beta}{\text{Min}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_{D_2}^2 + \frac{\alpha}{2} \|\beta\|_{D_1}^2 \mid \beta \in \mathbb{R}^K \right\} \quad (21)$$

The solution is:

$$\beta_p^* = \left(\mathbf{X}^t \mathbf{D}_2^{-1} \mathbf{X} + \alpha \mathbf{D}_1^{-1} \right)^{-1} \mathbf{X}^t \mathbf{D}_2^{-1} \mathbf{y} \quad (22)$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ can be substituted for any weight matrix of interest. Using the first component of $\xi_{GE}^* = \begin{pmatrix} \beta_{GE}^* \\ \epsilon_{GE}^* \end{pmatrix}$, we can state the following.

Lemma 5.1. With the above notations, $\beta_p^* = \beta_{GE}^*$ for $\alpha = 1$.

Proof of Lemma 5.1. The condition $\beta_P^* = \beta_{GE}^*$ amounts to:

$$\left(\mathbf{X}' \mathbf{D}_2^{-1} \mathbf{X} + \alpha \mathbf{D}_1^{-1} \right)^{-1} \mathbf{X}' \mathbf{D}_2^{-1} \mathbf{y} = \mathbf{D}_1 \mathbf{X}' \mathbf{M}^{-1} \mathbf{y} = \mathbf{D}_1 \mathbf{X}' \left\{ \mathbf{X} \mathbf{D}_1 \mathbf{X}' + \mathbf{D}_2 \right\}^{-1}$$

independently of $\mathbf{y}$. For this equality to hold, $\alpha=1$.

The above result shows that if we weigh the discrepancy between the observed data ($\mathbf{y}$) and its true value ($\mathbf{X}\beta$) by the prior covariance matrix $\mathbf{D}_2$, the penalized GLS and our entropy solutions coincide for $\alpha=1$ and for normal priors.

The comparison of β_{GE}^* with β_{LS}^* is stated below.

Lemma 5.2. With the above notations, $\beta_{GE}^* = \beta_{LS}^*$ when the constraints are in terms of pure moments (zero moments).

Proof of Lemma 5.2. If $\beta_{GE}^* = \beta_{LS}^*$, then $\mathbf{D}_1 \mathbf{X}' \mathbf{M}^{-1} \mathbf{y} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$ for all $\mathbf{y}$, which implies the following chain of identities:

$$\begin{aligned} \mathbf{D}_1 \mathbf{X}' \mathbf{M}^{-1} &= (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \Leftrightarrow \mathbf{X}' \mathbf{X} \mathbf{D}_1 \mathbf{X}' = \mathbf{X}' \mathbf{M} \\ \Leftrightarrow \mathbf{X}' \mathbf{X} \mathbf{D}_1 \mathbf{X}' &= \mathbf{X}' (\mathbf{X} \mathbf{D}_1 \mathbf{X}' + \mathbf{D}_2) \Leftrightarrow \mathbf{X}' \mathbf{D}_2 = 0. \end{aligned}$$

Clearly there are only two possibilities. First, if the noise components are not constant, $\mathbf{D}_2$ is invertible and therefore $\mathbf{X}'$ must vanish (trivial but an uninteresting case). Second, if the variance of the noise component is zero, (1) becomes a pure linear inverse problem (*i.e.*, we solve $\mathbf{y} = \mathbf{X}\beta$).

5.1.2. The Moments' Case

Up to now, the comparison was done where the Generalized Entropy, GE, estimator was optimized under a larger space ($\mathbf{A}$) than the other LS or GLS estimators. In other words, the constraints in the GE estimator are the data points rather than the moments. The comparison is easier if one performs the above comparisons under similar spaces, namely using the sample's moments. This can easily be done if $\mathbf{X}' \mathbf{X}$ is invertible, and where we re-specify $\mathbf{A}$ to be the generic matrix $\mathbf{A} = [\mathbf{X}' \mathbf{X} \ \mathbf{X}']$, rather than $\mathbf{A} = [\mathbf{X} \ \mathbf{I}]$. Now, let $\mathbf{y}' \equiv \mathbf{X}' \mathbf{y}$, $\mathbf{X}' \equiv \mathbf{X}' \mathbf{X}$, and $\boldsymbol{\varepsilon}' \equiv \mathbf{X}' \boldsymbol{\varepsilon}$, then the problem is represented as $\mathbf{y}' = \mathbf{X}' \beta + \boldsymbol{\varepsilon}'$. In that case the conditions for $\beta_{GE}^* = \beta_{LS}^*$ is the trivial condition $\mathbf{X}' \mathbf{D}_2 \mathbf{X} = \mathbf{0}$.

In general, when $\mathbf{X}' \mathbf{X}$ is invertible, it is easy to verify that the solutions to variational problems of the type $\mathbf{y}' \equiv \mathbf{X}' \mathbf{y} = \mathbf{X}' \mathbf{X} \beta$ are of the form $(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$. In one case, the problem is to find:

$$\beta^* = \arg \min \left\{ \frac{1}{2} \|\mathbf{y}' - \mathbf{X}' \mathbf{X} \beta\|^2 \mid \beta \in \mathbb{R}^K \right\} \quad (23)$$

while in the other case the solution consists of finding:

$$\beta^* = \arg \min \left\{ \frac{1}{2} \|\beta\|^2 \mid \beta \in \mathbb{R}^K ; \mathbf{y}' = \mathbf{X}' \mathbf{X} \beta \right\}. \quad (24)$$

Under this “moment” specification, the solutions to the three different methods described above (16, 23 and 24) coincide.

5.2. The Basic Bayesian Method

Under the Bayesian approach we may think of our problem in the following way. Assume, as before, that C_n and C_s are closed convex subsets of $\mathbb{R}^N$ and $\mathbb{R}^K$ respectively and that $Q_n(d\mathbf{v}) = g_n(\mathbf{v})d\mathbf{v}$ and $Q_s(d\mathbf{z}) = g_s(\mathbf{z})d\mathbf{z}$. For the rest of this section, the priors $g_s(\mathbf{z})$, $g_n(\mathbf{v})$ will have their usual Bayesian interpretation. For a given $\mathbf{z}$, we think of $\mathbf{y} = \mathbf{X}\mathbf{z} + \mathbf{v}$ as a realization of the random variable $Y = \mathbf{X}\mathbf{Z} + \mathbf{V}$. Then, $g_{y|z}(\mathbf{y}|\mathbf{z}) = g_n(\mathbf{y} - \mathbf{X}\mathbf{z})$. The joint density $g_{y,z}(\mathbf{y}, \mathbf{z})$ of Y and Z , where Z is distributed according to the prior Q_s is:

$$g_{y,z}(\mathbf{y}, \mathbf{z}) = g_{y|z}(\mathbf{y}|\mathbf{z})g_s(\mathbf{z}) = g_n(\mathbf{y} - \mathbf{X}\mathbf{z})g_s(\mathbf{z})$$

The marginal distribution of $\mathbf{y}$ is $\int_{C_s} g_{y,z}(\mathbf{y}, \mathbf{z})d\mathbf{z}$ and therefore by Bayes Theorem the posterior (post-data) conditional $g_{z|y}(\mathbf{z}|\mathbf{y})$ is $g_{y,z}(\mathbf{y}, \mathbf{z})/g(\mathbf{y})$ from which:

$$\beta_B^* = E[Z|Y=y] = \frac{\int_{C_s} \mathbf{z}g_n(\mathbf{y} - \mathbf{X}\mathbf{z})g_s(\mathbf{z})d\mathbf{z}}{\int_{C_s} g_n(\mathbf{y} - \mathbf{X}\mathbf{z})g_s(\mathbf{z})d\mathbf{z}} \quad (25)$$

As usual β_B^* minimizes $E(\|\mathbf{Z} - \hat{\beta}(\mathbf{y})\|^2)$ where Z and Y are distributed according to $g_{y,z}(\mathbf{y}, \mathbf{z})$. The conditional covariance matrix:

$$E\left[(\mathbf{Z} - \beta_B^*)(\mathbf{Z} - \beta_B^*)^t | Y = y\right] = \int_{C_s} (\mathbf{z} - \beta_B^*)(\mathbf{z} - \beta_B^*)^t g_{z|y}(\mathbf{z}|\mathbf{y})d\mathbf{z}$$

is such that:

$$\text{tr}\left\{E\left[(\mathbf{Z} - \beta_B^*)(\mathbf{Z} - \beta_B^*)^t | Y = y\right]\right\} = \text{Var}(\mathbf{Z}|\mathbf{y}),$$

where $\text{Var}(\mathbf{Z}|\mathbf{y})$ is the total variance of the K random variates $\mathbf{z}$ in Z . Finally, it is important to emphasize here that the Bayesian approach provides us with a whole range of tools for inference, forecasting, model averaging, posterior intervals, etc. In this paper, however, the focus is on estimation and on the basic comparison of our GE method with other methods under the notations and formulations developed here. Extensions to testing and inference are left for future work.

5.2.1. A Standard Example: Normal Priors

As before, we view β and ε as realizations of random variables $\mathbf{Z}$ and $\mathbf{V}$ having the informative normal “a priori” (priors for signal and noise) distributions:

$$f_Z(\mathbf{z}) = \frac{\exp\left(-\frac{1}{2}\langle \mathbf{z}, \mathbf{D}_1^{-1}\mathbf{z} \rangle\right)}{\left((2\pi)^K \det \mathbf{D}_1\right)^{1/2}}$$

and:

$$f_{\mathbf{V}}(\mathbf{v}) = \frac{\exp\left(-\frac{1}{2}\langle\mathbf{v}, \mathbf{D}_2^{-1}\mathbf{v}\rangle\right)}{\left((2\pi)^N \det \mathbf{D}_2\right)^{1/2}}.$$

For notational convenience we assume that both $\mathbf{Z}$ and $\mathbf{V}$ are centered on zero and independent, and both covariance matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ are strictly positive definite. For comparison purposes, we are using the same notation as in Section 3. The randomness is propagated to the data $\mathbf{Y}$ such that the conditional density (or the conditional priors on $\mathbf{y}$) of $\mathbf{Y}$ is:

$$f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) = \frac{\exp\left(-\frac{1}{2}\langle(\mathbf{y} - \mathbf{X}\mathbf{z}), \mathbf{D}_2^{-1}(\mathbf{y} - \mathbf{X}\mathbf{z})\rangle\right)}{\left((2\pi)^N \det \mathbf{D}_2\right)^{1/2}} \quad (26)$$

Then, the marginal distribution of $\mathbf{Y}$ is $f_{\mathbf{Y}}(\mathbf{y}) = \iint f_{\mathbf{Y}|\mathbf{Z},\mathbf{V}}(\mathbf{y}|\mathbf{z},\mathbf{v})f_{\mathbf{Z}}(\mathbf{z})f_{\mathbf{V}}(\mathbf{v})d\mathbf{z}d\mathbf{v}$. The conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$ is easy to obtain under the normal setup. Thus, the post-data distribution of the signal, β , given the data $\mathbf{y}$ is:

$$f_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y}) = \frac{\exp\left(-\frac{1}{2}\langle(\mathbf{z} - \mathbf{L}\mathbf{y}), \mathbf{C}^{-1}(\mathbf{z} - \mathbf{L}\mathbf{y})\rangle\right)}{\left((2\pi)^K \det \mathbf{C}\right)^{1/2}} \quad (27)$$

where $\mathbf{L} = \mathbf{C}\mathbf{X}^t\mathbf{D}_2^{-1}$ and $\mathbf{C}^{-1} := (\mathbf{D}_1^{-1} + \mathbf{X}^t\mathbf{D}_2^{-1}\mathbf{X})$. That is, “the posterior (post-data)” distribution of $\mathbf{Z}$ has changed (relative to the prior) by the data. Finally, the post-data expected value of $\mathbf{Z}$ is given by:

$$\beta_B^* = E_{\mathbf{Z}|\mathbf{Y}}[\mathbf{Z}] = \mathbf{C}\mathbf{X}^t\mathbf{D}_2^{-1}\mathbf{y} \quad (28)$$

This is the traditional Bayesian solution for the linear regression using the support spaces for both signal and noise within the framework developed here. As before, one can compare this Bayesian solution with our Generalized Entropy solution. Equation (28) is comparable with our solution (14) for $\mathbf{z}^0 = \mathbf{0}$ which is the Generalized Entropy method with normal priors and center of supports equal zero. In addition, it is easy to see that the Bayesian solution (28) coincides with the penalized GLS (model (24)) for $\alpha = I$.

A few comments on these brief comparisons are in place. First, under both approaches the complete posterior (or post-data) density is estimated and not only the posterior mean, though under the GE estimator the post-data is related to the pre-specified spaces and priors. (Recall that the Bayesian posterior means are specific to a particular loss function.) Second, the agreement between the Bayesian result and the minimizer of (24) with $\alpha = I$ assumes a known value of σ_v^2 , which is contained in $\mathbf{D}_2$. In the Bayesian result σ_v^2 is marginalized, so it is not conditional on that parameter. Therefore, with a known value of σ_v^2 , both estimators are the same.

There are two reasons for the equivalence of the three methods (GE, Bayes and Penalized GLS). The first is that there are no binding constraints imposed on the signal and the noise. The second is the choice of imposing the normal densities as informative priors for both signal and noise. In fact, this result is standard in inverse problem theory where $\mathbf{L}$ is known as the Wiener filter (see for example Bertero and Boccacci [37]). In that sense, the Bayesian technique and the GE technique have some

procedural ingredients in common, but the distinguishing factor is the way the posterior (post-data) is obtained. (Note that “posterior” for the entropy method, means the “post data” distribution which is based on both the priors and the data, obtained via the optimization process). In one case it is obtained by maximizing the entropy functional while in the Bayesian approach it is obtained by a direct application of Bayes theorem. For more background and related derivation of the ME and Bayes rule see Zellner [38,39].

5.3. Comparison with the Bayesian Method of Moments (BMOM)

The basic idea behind Zellner’s BMOM is to avoid a likelihood function. This is done by maximizing the continuous (Shannon) entropy subject to the empirical moments of the data. This yields the most conservative (closest to uniform) post data density (Zellner [14,40–43]; Zellner and Tobias [15]). In that way the BMOM uses only assumptions on the realized error terms which are used to derive the post data density.

Building on the above references, assume $(\mathbf{X}'\mathbf{X})^{-1}$ exists, then the LS solution to (1) is $\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ which is assumed to be the post data mean with respect to (yet) unknown distribution (likelihood). This is equivalent to assuming $\mathbf{X}'E[\mathbf{V} | \text{Data}] = \mathbf{0}$ (the columns of $\mathbf{X}$ are orthogonal to the $N \times 1$ vector $E[\mathbf{V} | \text{Data}]$). To find $g(\mathbf{z} | \text{Data})$, or in Zellener’s notation $g(\boldsymbol{\beta} | \text{Data})$, one applies the classical ME with the following constraints (information):

$$E[\mathbf{Z} | \text{Data}] = \hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

and:

$$\text{Var}[\mathbf{Z} | \text{Data}] = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2$$

where $\text{Var}[\mathbf{Z} | \text{Data}]$ is based on the assumption that $\text{Var}[\mathbf{V} | \text{Data}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\sigma^2$, or similarly under Zellner’s notation $\text{Var}[\boldsymbol{\varepsilon} | \text{Data}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\sigma^2$, and σ^2 is a positive parameter. Then, the maximum entropy density satisfying these two constraints (and the requirement that it is a proper density) is:

$$g(\boldsymbol{\beta} | \text{Data}) = g(\mathbf{z} | \text{Data}) \sim N\left(\hat{\boldsymbol{\beta}}_{LS}, (\mathbf{X}'\mathbf{X})^{-1} \sigma^2\right).$$

This is the BMOM post data density for the parameter vector with mean $\hat{\boldsymbol{\beta}}_{LS}$ under the two side conditions used here. If more side conditions are used, the density function g will not be normal. Information other than moments can also be incorporated within the BMOM.

Comparing Zellner’s BMOM with our Generalized Entropy method we note that the BMOM produces the post data density from which one can compute the vector of point estimates $\hat{\boldsymbol{\beta}}_{LS}$ of the unconstrained problem (1). Under the GE model, the solution $\boldsymbol{\beta}_{GE}^*$ satisfies the data/constraints within the joint support space C . Further, for the GE construction there is no need to impose exact moment constraints, meaning it provides a more flexible post data density. Finally, under both methods one can use the post data densities to calculate the uncertainties around future (unobserved) observations.

6. More Closed Form Examples

In Section 3, we formulated three relatively simple closed form cases. In the current section, we extend our analytic solutions to a host of potential priors. This section demonstrates the capabilities of our proposed estimator. We do not intend here to formulate our model under all possible priors. We present our examples in such a way that the priors can be assigned for either the signal or the noise components. The different priors we discuss correspond to different prior beliefs: unbounded (unconstrained), bounded below, bounded above, or bounded below and above. The following set of examples, together with those in Section 3, represents different cases of commonly used prior distributions, and their corresponding partition functions. Specifically, the different cases are the Laplace (bilateral exponential) which is symmetric but with heavy tails, the Gamma distribution that is bounded below and is non-symmetric, the continuous and discrete uniform distributions, and the Bernoulli distribution that allows an easy specification of a prior mean that is not at the center of the pre-specified supports. In all the examples below, the index d takes the possible values (dimensions) K , N , or $K+N$ depending if it relates to C_s (or $\mathbf{z}$), to C_n (or $\mathbf{v}$) or to both.

We note that information theoretic procedures were also used for producing priors (e.g., Jeffreys', Berger and Bernardo's, Zellner, *etc.*). In future work we will try to relate them to the procedure developed here. In these approaches, β is not always viewed as the mean, as given in Equation (2). For example, Jeffreys, Zellner (e.g., Zellner [41]) and others have used Cauchy priors and unbounded measure priors, for which the mean does not exist.

6.1. The Basic Formulation

Case 1. Bilateral Exponential—Laplace Distribution. Like the normal distribution, another possible unconstrained model is obtained if we take as reference measure a bilateral exponential, or Laplace distribution. This is useful for modeling distributions with tails heavier than the normal. The following derivation holds for both our generic model (captured via the generic matrix $\mathbf{A}$) and for just the signal or noise parts separately. We only provide here the solution for the signal part.

In this case, the density of dQ is $\prod_j (\sigma_j/2) \exp[-(\sigma_j |z_j - z_j^0|)]$. The parameters z_j^0 is the set of prior means and $1/2\sigma_j$ is the variance of each component. The Laplace transform of dQ is:

$$\omega(\boldsymbol{\tau}) = \exp(-\langle \boldsymbol{\tau}, \mathbf{z}_0 \rangle) \prod_{j=1}^d \frac{\sigma_j^2}{\sigma_j^2 - \tau_j^2} \quad (29)$$

Next, we use the relationship $\boldsymbol{\tau} = \mathbf{X}'\boldsymbol{\lambda}$. (Note that under the generic formulation, instead of $\mathbf{X}'$, we can work with $\mathbf{X}^*$ which stands for either $\mathbf{X}'$, $\mathbf{I}$ or $\mathbf{A}'$.) We compute $\Omega(\boldsymbol{\lambda})$ via the Laplace transformation (8). It then follows that $D(\Omega) = \left\{ \boldsymbol{\lambda} \in \mathbb{R}^N \mid -\sigma_j < (\mathbf{X}'\boldsymbol{\lambda})_j < \sigma_j \right\}$ where $\omega(\mathbf{t})$ is always finite and positive. For this relationship to be satisfied, $|\tau_j| < \sigma_j$ for all $j = 1, 2, \dots, d$. Finally, replacing $\boldsymbol{\tau}$ by $\mathbf{X}'\boldsymbol{\lambda}$ yields $D(\Omega)$.

Next, minimizing the concentrated entropy function:

$$\sum(\lambda) = \ln \Omega(\lambda) + \langle \lambda, \mathbf{y} \rangle = \sum \ln \left(\frac{\sigma_j^2}{\sigma_j^2 - (\mathbf{X}'\lambda)_j} \right) + \langle \lambda, \mathbf{y} - \mathbf{X}\mathbf{z}^0 \rangle$$

by equating its gradient with respect to λ to $\mathbf{0}$, we obtain that at the minimum:

$$-\nabla_{\lambda} \ln \Omega(\lambda) = -X \left(\nabla_{\tau} \ln \omega(\tau) \right) \Big|_{\tau = X'\lambda} = \mathbf{y}. \quad (30)$$

Explicitly:

$$2 \sum_j \frac{X_{ij} (\mathbf{X}'\lambda)_j}{\sigma_j^2 - (\mathbf{X}'\lambda)_j^2} = (\mathbf{X}\mathbf{z}^0 - \mathbf{y})_i \quad (31)$$

Finally, having solved for the optimal vector λ^* that minimizes $\sum(\lambda)$, and such that the previous identity holds, we can rewrite our model as:

$$\sum_j X_{ij} \left(c_j^0 - \frac{2(\mathbf{X}'\lambda^*)_j}{\sigma_j^2 - (\mathbf{X}'\lambda^*)_j^2} \right) = \sum_j X_{ij} \beta_j^* = y_i \quad (32)$$

where rather than solve (31) directly, we make use of λ^* , that minimizes $\sum(\lambda)$ and satisfies (30).

As expected, the post-data has a well-defined Laplace distribution (Kotz *et al.* [44]) but this distribution is not symmetrical anymore, and the decay rate is modified by the data. Specifically:

$$dP(\lambda^*, \mathbf{z}) = \frac{e^{-\langle \lambda^*, \mathbf{X}\mathbf{z} \rangle}}{\Omega(\lambda^*)} dQ(\mathbf{z}) = e^{-\langle \lambda^*, \mathbf{X}(\mathbf{z} - \mathbf{z}^0) \rangle} \prod_j \left(\frac{\sigma_j^2 - (\mathbf{X}'\lambda^*)_j}{2\sigma_j} \right) \exp \left[-(\sigma_j |z_j - z_j^0|) \right] dz_j.$$

Case 2. Lower Bounds—Gamma Distribution. Suppose that the β 's are all bounded below by theory. Then, we can specify a random vector $\mathbf{Z}$ with values in the positive orthant translated to the lower bound K -dimensional vector $\mathbf{l}$, so $C_s = [l_1, \infty) \times \dots \times [l_d, \infty)$, where $[l_j, \infty) = [\mathbf{z}_j, \infty)$. Like related methods, we assume that each component $\mathbf{z}_j$ of $\mathbf{Z}$ is distributed in $[l_j, \infty)$ according to a translated $\Gamma(a_j, b_j)$. With this in mind, a direct calculation yields:

$$\omega(\tau) = \prod_{j=1}^d \left(\frac{a_j}{a_j + \tau_j} \right)^{b_j+1} e^{-\tau_j l_j} \quad (33)$$

where the particular case of $b_j = 0$ corresponds to the standard exponential distribution defined on $[\mathbf{l}_j, \infty)$. Finally, when τ is replaced by $\mathbf{X}'\lambda$, we get:

$$D(\Omega) = \left\{ \lambda \in \mathbb{R}^N \mid (\mathbf{X}'\lambda)_j > a_j, j = 1, \dots, K \right\}. \quad (34)$$

Case 3. Bounds on Signal and Noise. Consider the case where each component $\mathbf{Z}$ and $\mathbf{V}$ takes values in some bounded interval $[a_j, b_j]$. A common choice for the bounds of the errors supports in that case are the three-sigma rule (Pukelsheim [44]) where “sigma” is the empirical standard deviation of the sample analyzed (see for example Golan, Judge and Miller, [1] for a detailed discussion). In this

situation we provide two simple (and extreme) choices for the reference measure. The first is a uniform measure on $[a_j, b_j]$, and the second is a Bernoulli distribution supported on a_j and b_j .

6.1.2. Uniform Reference Measure

In this case the reference (prior) measure $dQ(z)$ is distributed according to the uniform density $\prod_j (b_j - a_j)^{-1}$ and the Laplace transform of this density is:

$$\omega(\boldsymbol{\tau}) = \prod_{j=1}^d \frac{e^{-\tau_j a_j} - e^{-\tau_j b_j}}{\tau_j (b_j - a_j)} \quad (35)$$

and $\omega(\boldsymbol{\tau})$ is finite for every vector $\boldsymbol{\tau}$.

6.1.3. Bernoulli Reference Measure

In this case the reference measure is singular (with respect to the volume measure) and is given by $dQ(\mathbf{z}) = \prod_j [p_j \delta_{a_j}(dz_j) + q_j \delta_{b_j}(dz_j)]$, where $\delta_c(dz)$ denotes the (Dirac) unit point mass at some point c , and where p_j and q_j do not have to sum up to one, yet they determine the weight within the bounded interval $[a_j, b_j]$. The Laplace transform of dQ is:

$$\omega(\boldsymbol{\tau}) = \prod_j [p_j e^{-\tau_j a_j} + q_j e^{-\tau_j b_j}] \quad (36)$$

where again, $\omega(\boldsymbol{\tau})$ is finite for all $\boldsymbol{\tau}$.

In this case, there is no common criterion that can be used to decide which a priori reference measure to choose. In many specific cases, we have noticed that a reconstruction with the discrete Bernoulli prior of $p = q = 1/2$ yields estimates that are very similar to the continuous uniform prior.

Case 4. Symmetric Bounds. This is a special case of Case 3 above for $a_j = -c_j$ and $b_j = c_j$ for positive c_j 's. The corresponding versions of (35) and (36), the uniform and Bernoulli, are respectively:

$$\omega(\boldsymbol{\tau}) = \prod_{j=1}^d \frac{e^{\tau_j c_j} - e^{-\tau_j c_j}}{2\tau_j c_j} \quad (37)$$

and:

$$\omega(\boldsymbol{\tau}) = \prod_j [p_j e^{\tau_j c_j} + q_j e^{-\tau_j c_j}] \quad (38)$$

6.2. The Full Model

Having developed the basic formulations and building blocks of our model, we note that the list can be amplified considerably and these building blocks can be assembled into a variety of combinations. We already demonstrated such a case in Section 3.3. We now provide such an example.

6.2.1. Bounded Parameters and Normally Distributed Errors

Consider the common case of naturally bounded signal and normally distributed errors. This case combines *Case 2* of Section 6.1 together with the normal case discussed in Section 3.1. Let $\beta \in [l_1, \infty) \times \dots \times [l_K, \infty)$ but we impose no constraints on the ε . From Section 2,

$dQ(\mathbf{z}, \mathbf{v}) = dQ_s(\mathbf{z})dQ_n(\mathbf{v})$ with $dQ_s(\mathbf{z}) = \prod_{j=1}^K \frac{a_j^{b_j} z_j^{b_j-1} e^{-z_j a_j}}{\Gamma(b_j)} dz_j$, and $dQ_n(\mathbf{v}) = \frac{e^{-\langle \mathbf{v}, \mathbf{D}_2^{-1} \mathbf{v} \rangle / 2}}{\sqrt{(2\pi)^N \det \mathbf{D}_2}} d\mathbf{v}$. The

signal component is formulated earlier, while $\omega_n(\mathbf{t}) = \exp\langle \mathbf{t}, \mathbf{D}_2 \mathbf{t} \rangle / 2$. Using $\mathbf{A} = [\mathbf{X} \ \mathbf{I}]$ we have

$\mathbf{A}^t \boldsymbol{\lambda} = \begin{pmatrix} \mathbf{X}^t \boldsymbol{\lambda} \\ \boldsymbol{\lambda} \end{pmatrix}$ for the N -dimensional vector $\boldsymbol{\lambda}$, and therefore, $\Omega(\boldsymbol{\lambda}) = \omega_s(\mathbf{X}^t \boldsymbol{\lambda}) \omega_n(\boldsymbol{\lambda}) = \Omega_s(\boldsymbol{\lambda}) \Omega_n(\boldsymbol{\lambda})$.

The maximal entropy probability measures (post-data) are:

$$\begin{aligned} dP^*(\mathbf{z}, \mathbf{v}) &= \frac{\exp\langle \boldsymbol{\lambda}^*, \mathbf{X}^t \mathbf{z} \rangle \exp\langle \boldsymbol{\lambda}^*, \mathbf{v} \rangle}{\Omega_s(\boldsymbol{\lambda}^*) \Omega_n(\boldsymbol{\lambda}^*)} dQ(\mathbf{z}, \mathbf{v}) \\ &= \prod_{j=1}^K \frac{\left(a_j + (\mathbf{X}^t \boldsymbol{\lambda})_j\right)^{b_j} z_j^{b_j-1} e^{-((\mathbf{X}^t \boldsymbol{\lambda})_j + a_j)(z_j - l_j)}}{\Gamma(b_j)} d\xi_j \frac{e^{\langle \mathbf{v} + \mathbf{D}_2 \boldsymbol{\lambda}, \mathbf{D}_2^{-1} (\mathbf{v} + \mathbf{D}_2 \boldsymbol{\lambda}) \rangle / 2}}{\sqrt{(2\pi)^N \det D_2}} d\mathbf{v} \end{aligned}$$

where $\boldsymbol{\lambda}^*$ is found by minimizing the concentrated entropy function:

$$\Sigma(\boldsymbol{\lambda}) = \frac{1}{2} \langle \boldsymbol{\lambda}, \mathbf{D}_2 \boldsymbol{\lambda} \rangle + \sum_{j=1}^K b_j \ln \left(\frac{a_j}{(\mathbf{X}^T \boldsymbol{\lambda})_j + a_j} \right) + \langle \boldsymbol{\lambda}, \mathbf{y} - \mathbf{X} \mathbf{l} \rangle$$

$\mathbf{l} = (l_1, l_2, \dots, l_K)$ and $\mathbf{l}$ determines the “shift” for each coordinate. (For example, if $\mathbf{D}_2 = \sigma^2 \mathbf{I}_N$, then $\frac{1}{2} \langle \boldsymbol{\lambda}, \mathbf{D}_2 \boldsymbol{\lambda} \rangle = \frac{1}{2} \sum_i \lambda_i^2 \sigma^2$, or for the simple heteroscedastic case, we have $\frac{1}{2} \langle \boldsymbol{\lambda}, \mathbf{D}_2 \boldsymbol{\lambda} \rangle = \frac{1}{2} \sum_i \lambda_i^2 \sigma_i^2$). Finally, once $\boldsymbol{\lambda}^*$ is found, we get:

$$\begin{aligned} E_{P_s^*}(z_j) &= \beta_j^* = l_j + \frac{b_j}{(\mathbf{X}^T \boldsymbol{\lambda}^*)_j + a_j}, \\ E_{P_n^*}(v_i) &= \varepsilon_i^* = -(\mathbf{D}_2 \boldsymbol{\lambda}^*)_i. \end{aligned}$$

For example, if $\mathbf{D}_2 = \sigma^2 \mathbf{I}_N$, then $\frac{1}{2} \langle \boldsymbol{\lambda}, \mathbf{D}_2 \boldsymbol{\lambda} \rangle = \frac{1}{2} \sum_i \lambda_i^2 \sigma^2$, or for the simple heteroscedastic case, we have

$$\frac{1}{2} \langle \boldsymbol{\lambda}, \mathbf{D}_2 \boldsymbol{\lambda} \rangle = \frac{1}{2} \sum_i \lambda_i^2 \sigma_i^2.$$

7. A Comment on Model Comparison

So far we have described our model, its properties and specified some closed form examples. The next question facing the researcher is how to decide on the most appropriate prior/model to use for a given set of data. In this section, we briefly comment on a few possible model comparison techniques.

A possible criterion for comparing estimations (reconstructions) resulting from different priors should be based on a comparison of the post-data entropies associated with the proposed setup.

Implicit in the choice of priors is the choice of supports (Z and V), which in turn is dictated by the constraints imposed on the β 's. Assumption 2.1 means that these constraints are properly specified, namely there is no arbitrariness in the choice of C_s . The choice of a specific model for the noise involves two assumptions. The first one is about the support that reflects the actual range of the errors. The second is the choice of a prior describing the distribution of the noise within that support. To contrast two possible priors, we want to compare the reconstructions provided by the different models for the signal and noise variables. Within the information theoretic approach taken here, comparing the post-data entropies seems a reasonable choice.

From a practical point of view, the post-data entropies depend on the priors and the data in an explicit but nonlinear way. All we can say for certain is that for all models (or priors) the optimal solution is:

$$S_Q(P^*) = S_Q(P_s^*) + S_Q(P_n^*) = \Sigma(\lambda^*) = \ln \Omega_s(\lambda^*) + \ln \Omega_n(\lambda^*) + \langle \lambda^*, y \rangle \quad (39)$$

where λ^* has to be computed by minimizing the concentrated entropy function $\Sigma(\lambda)$, and it is clear that the total entropy difference between the post-data and the priors is just the entropy difference for the signal plus the entropy difference for the noise. Note that $2S_Q(P^*) = 2\Sigma(\lambda^*) \rightarrow \chi^2_{(d)}$ as $N \rightarrow \infty$ where d is the dimension of λ . This is the entropy ratio statistics which is similar in nature to the empirical likelihood ratio statistic (e.g., Golan [27]). Rather than discussing this statistic here, we provide in Appendix 3 analytic formulations of Equation (39) for a large number of prior distributions. These formulations are based on the examples of earlier sections. Last, we note that in some cases, where the competing models are of different dimensions, a normalization of both statistics is necessary.

8. Conclusions

In this paper we developed a generic information theoretic method for solving a noisy, linear inverse problem. This method uses minimal a-priori assumptions, and allows us to incorporate constraints and priors in a natural way for a whole class of linear inverse problems across the natural and social sciences. This inversion method is generic in the sense that it provides a framework for analyzing non-normal models and it performs well also for data that are not of full rank.

We provided detailed analytic solutions for a large class of priors. We developed the first order properties as well as the large sample properties of that estimator. In addition, we compared our model to other methods such as the Least Squares, Penalized LS, Bayesian and the Bayesian Method of Moments.

The proposed model main advantage over other LS and ML methods is that it has better performance (more stable and lower variances) for (possibly small) finite samples. The smaller the sample and/or the more ill-behaved (e.g., collinear) is the sample, the better this method performs. However, if one knows the underlying distribution, the sample is well behaved and large enough the traditional ML is the correct model to use. The other advantages of our proposed model (relative to the GME and other IT estimators) are that (i) we can impose different priors (discrete or continuous) for the signal and the noise, (ii) we estimate the full distribution of each one of the two sets of unknowns (signal and noise), and (iii) our model is based on minimal assumptions.

In future research, we plan to study the small sample properties as well as develop statistics to evaluate the performance of the competing priors and models. We conclude by noting that the same

framework developed here can be easily extended for nonlinear estimation problems. This is because all the available information enters as stochastic constraints within the constrained optimization problem.

Acknowledgments

We thank Bertrand Clarke, George Judge, Doug Miller, Arnold Zellner and Ed Greenberg as well as participants of numerous conferences and seminars for their comments and suggestions on earlier versions of that paper. Golan thanks the Edwin T. Jaynes International Center for Bayesian Methods and Maximum Entropy for partially supporting this project. We also want to thank the referees for their comments, for their careful reading of the manuscript and for pointing out reference [26] to us. Their comments improved the presentation considerably.

References

1. Golan, A.; Judge, G.G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*; John Wiley & Sons: New York, NY, USA, 1996.
2. Gzyl, H.; Velásquez, Y. *Linear Inverse Problems: The Maximum Entropy Connection*; World Scientific Publishers: Singapore, 2011.
3. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
4. Jaynes, E.T. Information theory and statistical mechanics II. *Phys. Rev.* **1957**, *108*, 171–190.
5. Shannon, C. A mathematical theory of communication. *Bell System Technical. J.* **1948**, *27*, 379–423, 623–656.
6. Owen, A. Empirical likelihood for linear models. *Ann. Stat.* **1991**, *19*, 1725–1747.
7. Owen, A. *Empirical Likelihood*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2001.
8. Qin, J.; Lawless, J. Empirical likelihood and general estimating equations. *Ann. Stat.* **1994**, *22*, 300–325.
9. Smith, R.J. Alternative semi parametric likelihood approaches to GMM estimations. *Econ. J.* **1997**, *107*, 503–510.
10. Newey, W.K.; Smith, R.J. Higher order properties of GMM and Generalized empirical likelihood estimators. Department of Economics, MIT, Boston, MA, USA. Unpublished work, 2002.
11. Kitamura, Y.; Stutzer, M. An information-theoretic alternative to generalized method of moment estimation. *Econometrica* **1997**, *66*, 861–874.
12. Imbens, G.W.; Johnson, P.; Spady, R.H. Information-theoretic approaches to inference in moment condition models. *Econometrica* **1998**, *66*, 333–357.
13. Zellner, A. Bayesian Method of Moments/Instrumental Variables (BMOM/IV) analysis of mean and regression models. In *Prediction and Modeling Honoring Seymour Geisser*; Lee, J.C., Zellner, A., Johnson, W.O., Eds.; Springer Verlag: New York, NY, USA, 1996.
14. Zellner, A. The Bayesian Method of Moments (BMOM): Theory and applications. In *Advances in Econometrics*; Fomby, T., Hill, R., Eds.; JAI Press: Greenwich, CT, USA, 1997; Volume 12, pp. 85–105.
15. Zellner, A.; Tobias, J. Further results on the Bayesian method of moments analysis of multiple regression model. *Int. Econ. Rev.* **2001**, *107*, 1–15.

16. Gamboa, F.; Gassiat, E. Bayesian methods and maximum entropy for ill-posed inverse problems. *Ann. Stat.* **1997**, *25*, 328–350.
17. Gzyl, H. Maxentropic reconstruction in the presence of noise. In *Maximum Entropy and Bayesian Studies*; Erickson, G., Ryckert, J., Eds.; Kluwer: Dordrecht, The Netherlands, 1998.
18. Golan, A.; Gzyl, H. A generalized maxentropic inversion procedure for noisy data. *Appl. Math. Comput.* **2002**, *127*, 249–260.
19. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **1970**, *1*, 55–67.
20. O’Sullivan, F. A statistical perspective on ill-posed inverse problems. *Stat. Sci.* **1986**, *1*, 502–527.
21. Breiman, L. Better subset regression using the nonnegative garrote. *Technometrics* **1995**, *37*, 373–384.
22. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288.
23. Titterton, D.M. Common structures of smoothing techniques in statistics. *Int. Stat. Rev.* **1985**, *53*, 141–170.
24. Donoho, D.L.; Johnstone, I.M.; Hoch, J.C.; Stern, A.S. Maximum entropy and the nearly black object. *J. R. Stat. Soc. Ser. B* **1992**, *54*, 41–81.
25. Besnerais, G.L.; Bercher, J.F.; Demoment, G. A new look at entropy for solving linear inverse problems. *IEEE Trans. Inf. Theory* **1999**, *45*, 1565–1578.
26. Bickel, P.; Li, B. Regularization methods in statistics. *Test* **2006**, *15*, 271–344.
27. Golan, A. Information and entropy econometrics—A review and synthesis. *Found. Trends Econometrics* **2008**, *2*, 1–145.
28. Fomby, T.B.; Hill, R.C. *Advances in Econometrics*; JAI Press: Greenwich, CT, USA, 1997.
29. Golan, A., Ed. *Special Issue on Information and Entropy Econometrics (Journal of Econometrics)*; Elsevier: Amsterdam, The Netherlands, 2002; Volume 107, Issues 1–2, pp. 1–376.
30. Golan, A.; Kitamura, Y., Eds. *Special Issue on Information and Entropy Econometrics: A Volume in Honor of Arnold Zellner (Journal of Econometrics)*; Elsevier: Amsterdam, The Netherlands, 2007; Volume 138, Issue 2, pp. 379–586.
31. Mynbayev, K.T. *Short-Memory Linear Processes and Econometric Applications*; John Wiley & Sons: Hoboken, NY, USA, 2011.
32. Asher, R.C.; Borchers, B.; Thurber, C.A. *Parameter Estimation and Inverse Problems*; Elsevier: Amsterdam, Holland, 2003.
33. Golan, A. Information and entropy econometrics—Editor’s view. *J. Econom.* **2002**, *107*, 1–15.
34. Kullback, S. *Information Theory and Statistics*; John Wiley & Sons: New York, NY, USA, 1959.
35. Durbin, J. Estimation of parameters in time-series regression models. *J. R. Stat. Soc. Ser. B* **1960**, *22*, 139–153.
36. Mittelhammer, R.; Judge, G.; Miller, D. *Econometric Foundations*; Cambridge Univ. Press: Cambridge, UK, 2000.
37. Bertero, M.; Boccacci, P. *Introduction to Inverse Problems in Imaging*; CRC Press: Boca Raton, FL, USA, 1998.
38. Zellner, A. Optimal information processing and Bayes theorem. *Am. Stat.* **1988**, *42*, 278–284.
39. Zellner, A. Information processing and Bayesian analysis. *J. Econom.* **2002**, *107*, 41–50.

40. Zellner, A. Bayesian Method of Moments (BMOM) Analysis of Mean and Regression Models. In *Modeling and Prediction*; Lee, J.C., Johnson, W.D., Zellner, A., Eds.; Springer: New York, NY, USA, 1994; pp. 17–31.
41. Zellner, A. Models, prior information, and Bayesian analysis. *J. Econom.* **1996**, *75*, 51–68.
42. Zellner, A. *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*; Edward Elgar Publishing Ltd.: Cheltenham Glos, UK, 1997; pp. 291–304, 308–318.
43. Kotz, S.; Kozubowski, T.; Podgórski, K. *The Laplace Distribution and Generalizations*; Birkhauser: Boston, MA, USA, 2001.
44. Pukelsheim, F. The three sigma rule. *Am. Stat.* **1994**, *48*, 88–91.

Appendix 1: Proofs

Proof of Proposition 4.1. From Assumptions 4.1–4.3 we have:

$$\varphi_N(\boldsymbol{\mu}) = -\frac{1}{N} X_N^T A \int_C \xi \frac{\exp\left(-\left\langle \boldsymbol{\mu}, \frac{1}{N} X_N A \xi \right\rangle\right)}{\tilde{\Omega}_N(\boldsymbol{\mu})} dQ(\xi) \rightarrow -W \int_{C_s} \xi \frac{\exp\left(-\langle \boldsymbol{\mu}, W \xi \rangle\right)}{\tilde{\Omega}_\infty(\boldsymbol{\mu})} dQ_s(\xi) = \varphi_\infty(\boldsymbol{\mu}).$$

Note that if the P_N^* -covariance of the noise component of ξ is $C_n^*(\mathbf{v}, \mathbf{v}) = \sigma^2 \mathbf{I}_N$, then P_N^o -covariance of ξ is an $(N + K) \times (N + K)$ -matrix given by $C_{N+K}(\xi, \xi) = W_N C_{ss}^*(\mathbf{z}, \mathbf{z}) W_N + \sigma^2 W_N / N$. Here $C_s^*(\mathbf{z}, \mathbf{z})$ is the P_N^* -covariance of the signal component of ξ . Again, from Assumptions 4.1–4.3 it follows that $W_N C_s^*(\xi, \xi) W_N \rightarrow W C_s^*(\xi, \xi) W$ which is the covariance of the signal component of ξ with respect to the limit probability $dP_\infty(\mathbf{z}) = \frac{\exp(-\langle \boldsymbol{\mu}, W \mathbf{z} \rangle)}{\tilde{\Omega}_\infty(\boldsymbol{\mu})} dQ(\mathbf{z})$. Therefore, φ_∞ is also invertible and $\varphi_\infty^{-1} = \psi_\infty$. To verify the uniform convergence of $\psi_N(\mathbf{y})$ towards $\psi_\infty(\mathbf{y})$ note that:

$$\|\boldsymbol{\mu} - \psi_N(\varphi_\infty(\boldsymbol{\mu}))\| = \|\psi_N(\varphi_N(\boldsymbol{\mu})) - \psi_N(\varphi_\infty(\boldsymbol{\mu}))\| \leq K_\mu \|\varphi_N(\boldsymbol{\mu}) - \varphi_\infty(\boldsymbol{\mu})\| \rightarrow 0 \text{ as } N \rightarrow \infty.$$

Proof of Lemma 4.5 (First Order Unbiasedness). Observe that for N large, keeping only the first term of the Taylor expansion we have:

$$\boldsymbol{\beta}_N - \boldsymbol{\beta} = \boldsymbol{\theta}'(W\boldsymbol{\mu}^*)\{W_N(\boldsymbol{\mu}_N^* - \boldsymbol{\mu}) + (W_N - W)\boldsymbol{\mu}\} = \boldsymbol{\theta}'(W\boldsymbol{\mu}^*)W_N(\boldsymbol{\mu}_N^* - \boldsymbol{\mu})$$

after we drop the $o(1/N)$ term. Keeping only the first term of the Taylor expansion, and invoking the assumptions of Lemma 4.5:

$$\boldsymbol{\mu}_N^* - \boldsymbol{\mu} = \psi_N(\tilde{\mathbf{y}}_N) - \psi_\infty(\tilde{\mathbf{y}}_\infty) = \psi'(\tilde{\mathbf{y}}_\infty)(\tilde{\mathbf{y}}_N - \tilde{\mathbf{y}}_\infty) + o(1/N).$$

Incorporating the model's equations, we see that under the approximations made so far:

$$\boldsymbol{\beta}_N^* - \boldsymbol{\beta} = \boldsymbol{\theta}'(W\boldsymbol{\mu}^*)W_N\psi'(\tilde{\mathbf{y}}_\infty)\left(\frac{1}{N} X_N^T \boldsymbol{\varepsilon}_N\right) + o(1/N) = W_N^{-1}\left(\frac{1}{N} X_N^T \boldsymbol{\varepsilon}_N\right) + o(1/N).$$

We used the fact that $\boldsymbol{\theta}'(W\boldsymbol{\mu}^*) = \tilde{D}_1$ and $\psi'(\tilde{\mathbf{y}}_\infty) = (W_N \tilde{D}_1 W_N + \sigma^2 W_N / N)^{-1}$ are the respective Jacobian matrices. The first order unbiasedness follows by taking expectations, $E_{Q_n}[\boldsymbol{\varepsilon}_N] = 0$.

Proof of Lemma 4.6 (Consistency in squared mean). With the same notations as above, consider $E_{Q_n}[\|\beta_N^* - \beta\|^2]$. Using the representation of lemma 4.5, $\beta_N^* - \beta = W_N^{-1} \left(\frac{1}{N} X_N^T \epsilon \right)$ and computing the expected square norm indicated above, we obtain $\sigma^2 \text{tr}(W_N^{-1}) / N$ which from Assumption 4.1 tends to 0 as $N \rightarrow \infty$.

Proof of Proposition 4.2.

Part (a) The proof is based on Lemma 4.5. Notice that under Q_n for any $\mathbf{k} \in \mathbb{R}^K$,

$$\langle \mathbf{k}, \beta_N^* - \beta \rangle = \left\langle \mathbf{k}, W_N^{-1} \left(\frac{1}{N} X_N^T \epsilon \right) \right\rangle = \frac{1}{N} \langle \mathbf{b}_N, \epsilon \rangle \text{ where } \mathbf{b}_N = X_N W_N^{-1} \mathbf{k}. \text{ Since the components of } \epsilon \text{ are i.i.d.}$$

random variables, the standard approximations yield:

$$E_{Q_n}[e^{i\langle \mathbf{k}, \beta_N^* - \beta \rangle}] \approx \exp - \frac{\sigma^2}{2N^2} \|\mathbf{b}_N\|^2 = \exp - \frac{\sigma^2}{2N} \text{tr}(W_N^{-1}), \text{ where } i = \sqrt{-1}, \text{ and therefore the law of } \beta_N^* - \beta \text{ concentrates at } \mathbf{0} \text{ asymptotically. This completes Part (a).}$$

Part (b) This part is similar to the previous proof, except that now the $\sqrt{N}$ factor in the exponent changes the result to be $E_{Q_n}[e^{i\langle \mathbf{k}, \sqrt{N}(\beta_N^* - \beta) \rangle}] \approx \exp - \frac{\sigma^2}{2} \text{tr}(W_N^{-1})$ as $N \rightarrow \infty$, from which assertion (b) of the proposition follows by the standard continuity theorem.

Appendix 2: Normal Priors — Derivation of the Basic Linear Model

Consider the linear model $y_i = a + bx_i + \epsilon_i$ where $\mathbf{X} = (\mathbf{1}, \mathbf{x})$, and $\beta = (ab)^T$. We assume that (i) both Q_s and Q_n are normal and that (ii) Q_s and Q_n are independent, meaning the Laplace transform is just (10). Recalling that $\mathbf{t} = \mathbf{A}'\lambda$, and that in the generic model

$\mathbf{A} = [\mathbf{X} \ \mathbf{I}] = [\mathbf{1} \ \mathbf{x} \ \mathbf{I}]$ where $\mathbf{A}$ is an $N \times (N+2)$, or $N \times (N+K)$ for the general model with $K > 2$, dimensional matrix. The log of the normalization factor of the post-data, $\Omega(\lambda)$, is:

$$\ln \Omega(\lambda) = \langle \mathbf{A}'\lambda, \mathbf{D}\mathbf{A}'\lambda \rangle / 2 - \langle \mathbf{A}'\lambda, \mathbf{c}^0 \rangle = \langle \lambda, \mathbf{A}\mathbf{D}\mathbf{A}'\lambda \rangle / 2 - \langle \lambda, \mathbf{A}\mathbf{c}^0 \rangle.$$

Building on (11), the concentrated (dual) entropy function is:

$$\Sigma(\lambda) = \ln \Omega(\lambda) + \langle \lambda, \mathbf{y} \rangle = \langle \lambda, \mathbf{A}\mathbf{D}\mathbf{A}'\lambda \rangle / 2 - \langle \lambda, \mathbf{A}\mathbf{c}^0 \rangle + \langle \lambda, \mathbf{y} \rangle = \langle \lambda, \mathbf{A}\mathbf{D}\mathbf{A}'\lambda \rangle / 2 - \langle \lambda, \mathbf{B} \rangle$$

where $\mathbf{B} \equiv \mathbf{y} - \mathbf{A}\mathbf{c}^0$. Solving for λ^* , $(\nabla_\lambda \Sigma(\lambda) = \mathbf{0})$, yields:

$$\underbrace{\mathbf{A}\mathbf{D}\mathbf{A}'}_{\mathbf{M}} \lambda^* + \mathbf{B} = \mathbf{0}$$

and finally, $\lambda^* = -\mathbf{M}^{-1}\mathbf{B}$. Explicitly, $\mathbf{M}$ is:

$$\begin{aligned}
[\mathbf{X} \mathbf{I}] \mathbf{D} \begin{pmatrix} \mathbf{X}' \\ \mathbf{I} \end{pmatrix} &= [\mathbf{1} \ \mathbf{x} \ \mathbf{I}] \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix} \begin{pmatrix} \mathbf{1}' \\ \mathbf{x}' \end{pmatrix} = [\mathbf{1} \ \mathbf{x} \ \mathbf{I}] \begin{pmatrix} \mathbf{D}_1 \begin{pmatrix} \mathbf{1}' \\ \mathbf{x}' \end{pmatrix} \\ \mathbf{D}_2 \end{pmatrix} \\
&= \left((\mathbf{1} \ \mathbf{x}) \mathbf{D}_1 \begin{pmatrix} \mathbf{1}' \\ \mathbf{x}' \end{pmatrix} + \mathbf{D}_2 \right) = (\mathbf{X} \mathbf{D}_1 \mathbf{X}' + \mathbf{D}_2) = (\mathbf{M}_1 + \mathbf{M}_2) \equiv \mathbf{M}
\end{aligned}$$

where:

$$[\mathbf{1} \ \mathbf{x}] \mathbf{D}_1 \begin{pmatrix} \mathbf{1}' \\ \mathbf{x}' \end{pmatrix} = [\mathbf{1} \ \mathbf{x}] \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{1}' \\ \mathbf{x}' \end{pmatrix} = \mathbf{1} \mathbf{D}_{11} \mathbf{1}' + \mathbf{1} \mathbf{D}_{12} \mathbf{x}' + \mathbf{x} \mathbf{D}_{21} \mathbf{1}' + \mathbf{x} \mathbf{D}_{22} \mathbf{x}'$$

and $\mathbf{1} \mathbf{1}' = \mathbf{1}_N$.

Next, we solve for the optimal $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$. Recalling the optimal solution is $\boldsymbol{\lambda}^* = -\mathbf{M}^{-1} \mathbf{B}$ and $\Omega(\boldsymbol{\lambda}^*) = \omega(\mathbf{A}^T \boldsymbol{\lambda}^*)$, then following the derivations of Section 3 we get:

$$\Omega(\boldsymbol{\lambda}^*) = \exp \left\{ \frac{1}{2} \langle \mathbf{A}' \mathbf{M}^{-1} \mathbf{B}, \mathbf{D} \mathbf{A}' \mathbf{M}^{-1} \mathbf{B} \rangle + \langle \mathbf{A}' \mathbf{M}^{-1} \mathbf{B}, \mathbf{c}_0 \rangle \right\}$$

and:

$$\rho^*(\xi) = \frac{e^{-\langle \boldsymbol{\lambda}^*, \mathbf{A} \xi \rangle}}{\Omega(\boldsymbol{\lambda}^*)} = \frac{e^{\langle \mathbf{B}, \mathbf{M}^{-1} \mathbf{A} \xi \rangle}}{\Omega(\boldsymbol{\lambda}^*)},$$

so:

$$\begin{aligned}
dP^*(\xi) &= \frac{e^{\langle \mathbf{B}, \mathbf{M}^{-1} \mathbf{A} \xi \rangle}}{\Omega(\boldsymbol{\lambda}^*)} \frac{e^{\frac{1}{2} \langle (\xi - \mathbf{c}_0), \mathbf{D}^{-1} (\xi - \mathbf{c}_0) \rangle}}{(2\pi)^{d/2} (\det \mathbf{D})^{1/2}} d\xi \\
&= \frac{e^{\langle \mathbf{B}, \mathbf{M}^{-1} \mathbf{A} \mathbf{c}_0 \rangle} e^{\langle \mathbf{B}, \mathbf{M}^{-1} \mathbf{A} (\xi - \mathbf{c}_0) \rangle} e^{-\frac{1}{2} \langle (\xi - \mathbf{c}_0), \mathbf{D}^{-1} (\xi - \mathbf{c}_0) \rangle}}{\Omega(\boldsymbol{\lambda}^*) (2\pi)^{d/2} (\det \mathbf{D})^{1/2}} d\xi.
\end{aligned}$$

Rewriting the exponent in the numerator as:

$$\begin{aligned}
&\langle \mathbf{A}' \mathbf{M}^{-1} \mathbf{B}, (\xi - \mathbf{c}_0) \rangle - \frac{1}{2} \langle (\xi - \mathbf{c}_0), \mathbf{D}^{-1} (\xi - \mathbf{c}_0) \rangle \\
&= -\frac{1}{2} \langle (\xi - \mathbf{c}_0 - \mathbf{D} \mathbf{A}' \mathbf{M}^{-1} \mathbf{B}), \mathbf{D}^{-1} (\xi - \mathbf{c}_0 - \mathbf{D} \mathbf{A}' \mathbf{M}^{-1} \mathbf{B}) \rangle + \frac{1}{2} \langle \mathbf{A}' \mathbf{M}^{-1} \mathbf{B}, \mathbf{D} \mathbf{A}' \mathbf{M}^{-1} \mathbf{B} \rangle
\end{aligned}$$

and incorporating it in dP^* yields:

$$dP^* = \frac{e^{-\frac{1}{2} \langle (\xi - \mathbf{c}_0 - \mathbf{D} \mathbf{A}' \mathbf{M}^{-1} \mathbf{B}), \mathbf{D}^{-1} (\xi - \mathbf{c}_0 - \mathbf{D} \mathbf{A}' \mathbf{M}^{-1} \mathbf{B}) \rangle}}{(2\pi)^{d/2} (\det \mathbf{D})^{1/2}} \times \frac{e^{\langle \mathbf{B}, \mathbf{M}^{-1} \mathbf{A} \mathbf{c}_0 \rangle + \frac{1}{2} \langle (\mathbf{A}' \mathbf{M}^{-1} \mathbf{B}), \mathbf{D} \mathbf{A}' \mathbf{M}^{-1} \mathbf{B} \rangle}}{\Omega(\boldsymbol{\lambda}^*)} d\xi,$$

where the second right-hand side term equals 1. Finally:

$$E_{P^*} \begin{pmatrix} \mathbf{z} \\ \mathbf{v} \end{pmatrix} = E_{P^*}(\xi) = \begin{pmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\varepsilon}^* \end{pmatrix} = \mathbf{c}_0 + \mathbf{D} \mathbf{A}' \mathbf{M}^{-1} (\mathbf{y} - \mathbf{A} \mathbf{c}_0) = \boldsymbol{\xi}^*.$$

To check our solution, note that $\mathbf{A}\xi^* = \mathbf{A}\mathbf{c}_0 + \mathbf{A}\mathbf{D}\mathbf{A}^t\mathbf{M}^{-1}(\mathbf{y} - \mathbf{A}\mathbf{c}_0) = \mathbf{y}$, so:

$$\begin{pmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\varepsilon}^* \end{pmatrix} = \begin{pmatrix} \mathbf{z}_0 \\ \mathbf{v}_0 \end{pmatrix} + \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}^t \\ \mathbf{I} \end{pmatrix} \{ \mathbf{M}^{-1}\mathbf{y} - \mathbf{M}^{-1}\mathbf{A}\mathbf{c}_0 \}$$

and finally:

$$\begin{aligned} \boldsymbol{\beta}^* &= \mathbf{z}^0 + \mathbf{D}_1\mathbf{X}^t\mathbf{M}^{-1}\mathbf{B} \\ \boldsymbol{\varepsilon}^* &= \mathbf{v}^0 + \mathbf{D}_2\mathbf{M}^{-1}\mathbf{B}, \end{aligned}$$

which is (14), where $\mathbf{B} = \mathbf{y} - \mathbf{A}\mathbf{c}_0$. Within the basic model, it is clear that $[\mathbf{X} \ \mathbf{I}] \begin{pmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\varepsilon}^* \end{pmatrix} = \mathbf{y}$, or $\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^* = \mathbf{y}$. Under the natural case where the errors' priors are centered on zero ($\mathbf{v}^0 = \mathbf{0}$), $\mathbf{M}^{-1}\mathbf{B} = \mathbf{D}_2^{-1}\boldsymbol{\varepsilon}^*$ and $\boldsymbol{\beta}^* = \mathbf{z}^0 + \mathbf{D}_1\mathbf{X}^t\mathbf{D}_2^{-1}\boldsymbol{\varepsilon}^*$. If in addition $\mathbf{z}^0 = \mathbf{0}$, then $\boldsymbol{\beta}^* = \mathbf{D}_1\mathbf{X}^t\mathbf{M}^{-1}\mathbf{y}$.

Appendix 3: Model Comparisons — Analytic Examples

We provide here the detailed analytical formulations of constructing the dual (concentrated) GE model for different priors. It can be used to derive the entropy ratio statistics based on Equation (39).

Example 1. The priors for both the signal and the noise are normal. In that case, the final post-data entropy, computed in Section 3, is:

$$\Sigma(\boldsymbol{\lambda}^*) = -\frac{1}{2} \langle \mathbf{y}, \mathbf{M}^{-1}\mathbf{y} \rangle.$$

This seems to be the only case amenable to full analytical computation.

Example 2. Laplace prior for state space variables plus an uniform prior (in $[-e, e]$) for the noise term. The full post-data entropy is:

$$\Sigma(\boldsymbol{\lambda}^*) = \sum_{j=1}^K \ln \left(\frac{\sigma_j^2}{\sigma_j^2 - (\mathbf{X}^t\boldsymbol{\lambda}^*)^2} \right) + \sum_{i=1}^N \ln \left(\frac{\sinh(\boldsymbol{\lambda}^* e)}{\boldsymbol{\lambda}^* e} \right) + \langle \boldsymbol{\lambda}^*, \mathbf{y} - \mathbf{X}\mathbf{z} \rangle.$$

Example 3. Normal prior for state space variables and an uniform prior (in $[-e, e]$) for the noise. The post-data entropy is:

$$\Sigma(\boldsymbol{\lambda}^*) = \frac{1}{2} \langle \boldsymbol{\lambda}^*, \mathbf{X}\mathbf{D}_1\mathbf{X}^t\boldsymbol{\lambda}^* \rangle + \sum_{i=1}^N \ln \left(\frac{\sinh(\boldsymbol{\lambda}^* e)}{\boldsymbol{\lambda}^* e} \right) + \langle \boldsymbol{\lambda}^*, \mathbf{y} - \mathbf{X}\mathbf{z}_0 \rangle$$

where $\mathbf{z}_0$ is the center of the normal priors and $\mathbf{D}_1$ is the covariance matrix of the state space variables.

Example 4. A Gamma prior for the state space variables, and an uniform prior (in $[-e, e]$) for the noise term. In this case the post-data entropy is:

$$\Sigma(\boldsymbol{\lambda}^*) = \sum_{j=1}^K (b_j + 1) \ln \left(\frac{a_j}{a_j + (\mathbf{X}^t\boldsymbol{\lambda}^*)^2} \right) + \sum_{i=1}^N \ln \left(\frac{\sinh(\boldsymbol{\lambda}^* e)}{\boldsymbol{\lambda}^* e} \right) + \langle \boldsymbol{\lambda}^*, \mathbf{y} - \mathbf{X}\mathbf{I} \rangle.$$

Example 5. The priors for the state space variables are Laplace and the priors for the noise are normal. Here, the post-data entropy is:

$$\Sigma(\lambda^*) = \sum_{j=1}^K \ln \left(\frac{\sigma_j^2}{\sigma_j^2 - (\mathbf{X}^t \lambda^*)^2} \right) + \frac{1}{2} \langle \lambda^*, \mathbf{D}_2 \lambda^* \rangle + \langle \lambda^*, \mathbf{y} - \mathbf{X} \mathbf{z}_0 \rangle.$$

Example 6. Both signal and noise have bounded supports, and we assume uniform priors for both. The post-data is:

$$\Sigma(\lambda^*) = \sum_{j=1}^K \ln \left(\frac{e^{-(\mathbf{X}^t \lambda^*)_j a_j} - e^{-(\mathbf{X}^t \lambda^*)_j b_j}}{(b_j - a_j)(\mathbf{X}^t \lambda^*)_j} \right) + \sum_{i=1}^N \ln \left(\frac{\sinh(\lambda_i^* e)}{\lambda_i^* e} \right) + \langle \lambda^*, \mathbf{y} \rangle.$$

Finally, we complete the set of examples with, probably, the most common case.

Example 7. Uniform priors on bounded intervals for the signal components and normal priors for the noise. The post-data entropy is:

$$\Sigma(\lambda^*) = \sum_{j=1}^K \ln \left(\frac{e^{-(\mathbf{X}^t \lambda^*)_j a_j} - e^{-(\mathbf{X}^t \lambda^*)_j b_j}}{(b_j - a_j)(\mathbf{X}^t \lambda^*)_j} \right) + \frac{1}{2} \langle \lambda^*, \mathbf{D}_2 \lambda^* \rangle + \langle \lambda^*, \mathbf{y} \rangle.$$

We reemphasize that this model comparison can only be used to compare models after each model has been completely worked out and for a given data set. Finally, we presented here the case of comparing the total entropies of post-data to priors, but as Equation (39) shows, one can just compare the post and pre data entropies of only the signal, $S_Q(P_s^*)$, or only the noise, $S_Q(P_n^*)$.

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).