Institute for Food and Resource Economics
University of Bonn

Discussion Paper 2008: 2

A BAYESIAN ALTERNATIVE TO GENERALIZED CROSS ENTROPY SOLUTIONS FOR UNDERDETERMINED ECONOMETRIC MODELS

Thomas Heckelei
University of Bonn, Institute for Food and Resource Economics, Bonn, Germany

Ron Mittelhammer
Washington State University, School of Economic Sciences, Pullman, USA

Torbjoern Jansson
Wageningen University, LEI, The Hague, Netherlands

Editor: Thomas Heckelei
Institute for Food and Resource Economics
University of Bonn
Nußallee 21
53115 Bonn, Germany

Phone: +49-228-732332
Fax: +49-228-734693
E-mail: thomas.heckelei@ilr.uni-bonn.de

A BAYESIAN ALTERNATIVE TO GENERALIZED CROSS ENTROPY SOLUTIONS FOR UNDERDETERMINED ECONOMETRIC MODELS

Thomas Heckelei, Ron Mittelhammer, Torbjoern Jansson

Abstract

This paper presents a Bayesian alternative to Generalized Maximum Entropy (GME) and Generalized Cross Entropy (GCE) methods for deriving solutions to econometric models represented by underdetermined systems of equations. For certain types of econometric model specifications, the Bayesian approach provides fully equivalent results to GME-GCE techniques. However, in its general form, the proposed Bayesian methodology allows a more direct and straightforwardly interpretable formulation of available prior information and can reduce significantly the computational effort involved in finding solutions. The technique can be adapted to provide solutions in situations characterized by either informative or uninformative prior information.

Keywords: Underdetermined Equation Systems, Maximum Entropy, Bayesian Priors, Structural Estimation, Calibration.
JEL-classification: C11, C13, C51

1 Introduction

In 1996, GOLAN, JUDGE AND MILLER published a book on "Maximum Entropy Econometrics" introducing Generalized Maximum Entropy (GME) and Generalized Cross Entropy techniques (GCE) to a wider range of applied econometricians. These estimation approaches were attractive to empirical modelers mainly for two reasons: First, they allow empirical specification and estimation of underdetermined models, i.e. models where the number of unknowns is larger than the number of equations, a capability not provided by classical solution techniques. Second, prior information on model unknowns can be included in a technically straightforward way, making estimates potentially more efficient in a mean square error sense, or at least more "plausible" for model simulation, interpretation, and analysis subsequent to estimation.

Since their introduction, a notable number of applications of GME and GCE have appeared in the empirical economics literature. A significant area of application relates to balancing large raw data sets using accounting identities and prior information

to fill gaps and reconcile conflicting data sources. The techniques allow setting ranges for missing data values and provide a means of differentiating the reliability of various sources in the balancing process (e.g. ROBINSON, CATTANBO AND EL-SAID 2000; BRITZ AND WIECK 2002, ROBILLIARD AND ROBINSON 2003). A related line of work deals with allocating input quantities to outputs from data on total input use and prior information on the input-output relationships (e.g. LENCE AND MILLER 1998a and b, LÉON ET AL. 1999). Calibration of simulation models to base year quantities and theory-consistent parameter sets is often done using entropy methods (e.g. PARIS AND HOWITT 1998; WITZKE AND BRITZ 1998; PARIS 2001) and a fairly new but increasingly important area is the spatial disaggregation of technological and economic data (HOWITT AND REYNAUD 2003). However, GME and GCE applications are not reserved for data recovery and calibration issues, and have been employed in attempts to better solve traditional estimation problems or analyze new ones (e.g. GOLAN, JUDGE AND PERLOFF 1996; OUDE LANSINK 1999; ZHANG AND FAN 2001; ARNDT, ROBINSON AND TARP 2002; HECKELEI AND WOLFF 2003). In essence, any economic model characterized by a vector of M equations in $K > M$ unknowns, say $\mathbf{g}(\mathbf{z}) = \mathbf{0}$, is an underdetermined model that can be solved through the use of GME or GCE techniques.

Despite the growing number of applications, GME and GCE techniques are arguably subject to at least three difficulties, the first being the specification and interpretation of prior information imposed via the use of discrete support points and a corresponding reference prior probability distribution on that support. In fact, the actual prior information ultimately imposed is a rather complicated composite of the choice of support points, the choice of reference prior probabilities on support points, and their interaction with the criterion of maximum entropy or minimum cross entropy in determining the final estimated subject probabilities on the support points. A second issue – connected to the first – relates to challenges in characterizing the nature of the estimation objective that is actually being used to combine prior and data information, with attendant difficulties in evaluation of the estimation results by the scientific community. Thirdly, the entropy approach introduces additional variables (the probabilities linked to the supports) and equations (adding up constraints for the probabilities) to the estimation process, which leads to a potential computational challenge especially for large data balancing applications. We elucidate as well as address these issues in the sections ahead.

The overall objective of this paper is to introduce a Bayesian alternative to GME and GCE techniques that allows for a direct and straightforwardly interpretable formulation of prior information and a clearly defined estimation objective while also reducing computational demands considerably when estimating an underdetermined

economic model. The specific objectives are reflected in the organization of the remaining sections of the paper, which is as follows. Section 2 reviews the GME-GCE approach in the context of estimating an underdetermined linear model without noise. We clarify the interpretation of the effective prior information imposed as being a combined effect of supports, reference probabilities on supports, and the solution for the subject probabilities via the maximum entropy criterion. Section 3 introduces a formulation of the underdetermined linear model estimation problem using a Bayesian approach that is fully equivalent to GME-GCE, where the underdetermined model equations and the data together represent the "Likelihood" information and all prior information is represented in terms of a prior density on model unknowns. This approach is then extended to solving general systems of underdetermined equations. In section 4, the approach is extended to accommodate the situation where the prior information is uninformative over the relevant parameter space. Section 5 provides illustrative applications, followed by concluding remarks.

2 Prior information in GME-GCE approaches

The principles of GME (later extended to GCE) estimation as introduced by GOLAN, JUDGE AND MILLER (1996) and discussed further in MITTELHAMMER, JUDGE AND MILLER (2000) are briefly reviewed here in the linear model context without noise to provide a conceptual foundation and identify notation for use in later sections. Within this basic model context, we elucidate the actual nature of the prior information that is implicitly used in the GME and GCE approaches.

Consider the underdetermined linear regression model, without noise, given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

where $\mathbf{y}$ is a T-dimensional column vector of observations on the dependent variable, $\mathbf{X}$ is a T×K matrix of observations on independent regressors with T<K, and $\boldsymbol{\beta}$ **is** a K-dimensional column vector of unknown parameters. The values of $\boldsymbol{\beta}$ cannot be uniquely identified with classical estimation techniques, such as ordinary least squares, because the number of observations is smaller than the number of parameters. The basic GME approach is to "reparameterize" the vector of parameters $\boldsymbol{\beta}$ such that each element is expressed as an expectation of a discrete probability distribution. Let $\mathbf{S}$ be a block-diagonal K×KL matrix of *support points*, where L is the number of support points associated with each parameter, and let $\mathbf{p}$ be a corresponding KL×1 vector of weights that have the properties of probabilities. The vector $\boldsymbol{\beta}$ can then be represented as

$$\boldsymbol{\beta} = \mathbf{Sp} = \begin{bmatrix} \mathbf{s}_1' & 0 & \cdots & 0 \\ 0 & \mathbf{s}_2' & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \mathbf{s}_K' \end{bmatrix} \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_K \end{bmatrix} \tag{2}$$

with $\mathbf{s}_k' = \begin{bmatrix} \mathbf{s}_{k1} & \mathbf{s}_{k2} & \cdots & \mathbf{s}_{kL} \end{bmatrix}$ such that $s_{k1} < s_{k2} < \ldots < s_{kL}$. A reparameterized version of (2) is then given by

$$\mathbf{y} = \mathbf{XSp} \tag{3}$$

which corresponds to the 'data constraints' of the GME approach. Realizing that the elements of each $\mathbf{p}_k$, k = 1,...,K sum to 1 consistent with their interpretation as having the properties of probabilities, equation (2) defines the *admissible values* for the elements of $\boldsymbol{\beta}$ as convex combinations of the corresponding support points $\mathbf{s}_k$, k = 1,...,K. This implies that the range of possible values for $\boldsymbol{\beta}_k$ is given by the interval $[s_{k1}, s_{kL}]$. The GME approach chooses among the infinite number of vectors $\mathbf{p}$ satisfying (3) so as to maximize the entropy criterion[1]

$$H[\mathbf{p}] = -\mathbf{p}' \ln \mathbf{p} \tag{4}$$

The objective function (4) attains an unconstrained maximum when all elements of $\mathbf{p}$ have the value 1/L, i.e. when the probabilities are uniform. Since the uniform distribution treats each outcome as equally likely one can view this distribution as the maximally uninformative distribution with respect to anticipating outcomes of a random variable. Thus, the *maximum value of entropy* is uniquely associated with the *maximally uninformative weight- probability distribution*" (MITTELHAMMER, JUDGE AND MILLER 2000, E3: 8). However, the notion of "uninformative" probabilities has caused some confusion in some applications of GME in that it has been incorrectly interpreted as characterizing the prior probabilities associated with various possible values of the parameters in the GME problem formulation. We will address this issue in more detail shortly.

The complete estimation problem can now be stated as

---

[1] The value of $p_{ij} \ln\left(p_{ij}\right)$ is defined to equal its limiting value of 0 when $p_{ij} = 0$

$$\begin{aligned} &\max_{\mathbf{p}} \mathrm{H}(\mathbf{p}) = -\mathbf{p}' \ln \mathbf{p} \\ &\text{subject to} \\ &\mathbf{y} = \mathbf{XSp} \\ &\boldsymbol{\iota}' \mathbf{p}_k = 1 \quad \forall \, \mathrm{k} \end{aligned} \tag{5}$$

where the last constraint ensures that the probabilities appropriately sum to one, with $\boldsymbol{\iota}$ being a L×1 'summation vector', i.e. a conformable vector of ones. The values of $\boldsymbol{\beta}$ can be recovered after optimization by the definition given in (2).

A crucial question for interpreting the results of the GME estimation approach is how one can interpret the notion of "uninformative" claimed above for the entropy criterion in the GME context. Of principal interest is the interpretation of the *expectation* of the probability distribution over the support points, since it is this expectation that represents the final estimate of the parameter vector $\boldsymbol{\beta}$, as defined in (4). The probability distributions inherent in the solved value of $\mathbf{p}$ merely serves as a vehicle for the entropy criterion to choose particular values of the expectation that maximize entropy. Or as PRECKEL (2001, p. 375) states: "*Thus, the role of the distribution is simply to serve as intermediary in expressing the desirability of the value of a parameter…*".

Preckel reinterprets GME as minimizing a penalty function on these expectations subject to the data constraints, and compares the approach to the case of the penalty function implied by a least squares criterion. We instead conceptualize the GME-implied weighting on expectations as the prior probability distribution in a Bayesian context. This prior density turns out to be a reflection of Preckel's penalty function (see his equation (5), p. 368).

For an explicit illustration of the implied prior, consider just one parameter β_k from the linear model in (1) and suppose that only two support points s_{k1} and s_{k2} are used, i.e. L=2. Recalling that $p_{k1} + p_{k2} = 1$ we write the expectation of β_k as

$$\mathrm{E}\beta_k = p_{k1} s_{k1} + (1 - p_{k1}) s_{k2} \tag{6}$$

Solving for the probability as a function of $\mathrm{E}\beta_k$ obtains

$$p_{k1}(\mathrm{E}\beta_k) = (\mathrm{E}\beta_k - s_{k2}) / (s_{k1} - s_{k2}) \tag{7}$$

The component of the entropy criterion in (5) relating to the expectation of β_k can then be expressed as

$$
\begin{aligned}
H(E\beta_k) &= -p_{k1}\ln(p_{k1}) - (1-p_{k1})\ln(1-p_{k1}) \\
&= -(E\beta_k - s_{k2})/(s_{k1}-s_{k2})\ln\left((E\beta_k - s_{k2})/(s_{k1}-s_{k2})\right) \\
&\quad -(s_{k1} - E\beta_k)/(s_{k1}-s_{k2})\ln\left((s_{k1} - E\beta_k)/(s_{k1}-s_{k2})\right)
\end{aligned} \tag{8}
$$

which defines the prior weight that the entropy criterion assigns to each possible value of the expectation of β_k. The criterion is maximized if the distance of $E\beta_k$ from the lower support point s_{k1} is equal to the distance of $E\beta_k$ from the upper support point s_{k2}, which coincides with $p_{k1} = p_{k2} = 0.5$, i.e. a uniform distribution over the supports, and a value for $\beta_k = (s_{k1} + s_{k2})/2$. All other values of $E\beta_k$ are assigned lower prior weights via $H(E\beta_k)$. A graphical illustration of the weight distribution is given in Figure 1, where we chose $s_{k1} = 0$ and $s_{k2} = 10$.

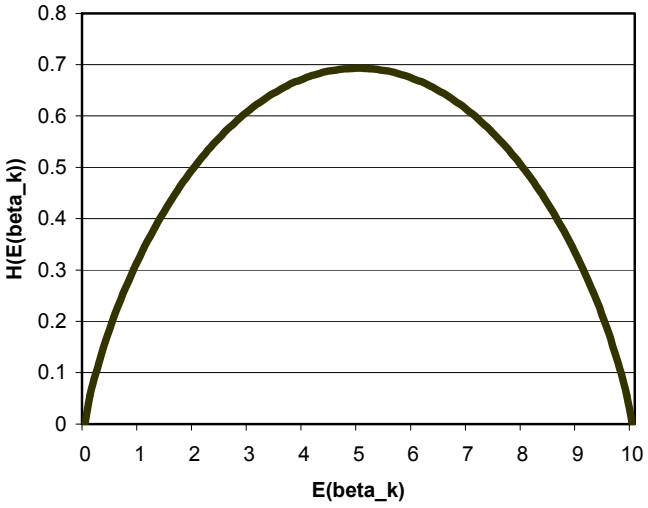

Figure 1. Prior weighting of parameter expectations based on the entropy criterion. Source: Maximum entropy calculations with $s_{k1} = 0$, $s_{k2} = 10$.

The mathematical and graphical illustration above demonstrates that the use of the maximum entropy criterion implies different prior weights on the different possible outcomes, $E\beta_k$, of the GME estimator. These are prior weights in that they are independent of any data information. The highest weight is given to the parameter expectation that would be generated by a uniform probability distribution over the supports.

The GME approach is a special case of GCE, where the latter method allows defining a reference probability distribution over the support points. Denoting the vector of reference distribution probabilities as **q**, the cross entropy criterion can be written as

$$I(\mathbf{p},\mathbf{q}) = \mathbf{p}'\ln(\mathbf{p}/\mathbf{q}) \tag{9}$$

where $\mathbf{p}/\mathbf{q}$ is to be interpreted as a vector with elements p_{sk}/q_{sk}. The value of $I(\mathbf{p},\mathbf{q})$ is smallest if all elements of the vector **p** are equal to the corresponding elements of the vector **q**. Consequently, an unconstrained *minimization* of the cross entropy measure over **p** will result in a probability distribution equal to **q**, and provides estimates of parameters according to expectations implied by the probabilities in **q**. The GME approach considered above is equivalent to an application of the GCE approach with a uniform *reference* distribution.

The use of a non-uniform reference distribution leads to modifications in the implicit prior weighting on parameter expectations under the GCE approach. Without repeating what amounts to a similar mathematical derivation to that in (6)-(8), we illustrate in Figure 2 the impact on the prior weights for the two support points example above. The reference probabilities were chosen such that $q_{k1} = 0.3$ and $q_{k2} = 0.7$. Note that we reflected the cross entropy value – which is minimized rather than maximized as in the GME case – around 0.6 to make the graph more easily comparable to Figure 1. In this case the highest cross-entropy weight is given to $E\beta_k = 7$, which would be the parameter estimate chosen by the GCE approach if data constraints render the value $\beta_k = 7$ feasible. A general principle of GCE is illustrated by the two examples — the prior that is actually implied by the method places the highest prior weight on the expectation that is implied by the *reference* probability distribution.

In summary, the GME/GCE approaches imply the use of *informative* prior information on parameters to be estimated. This is true, even if the reference distribution employed is uniform over the set of support points because the actual GME/GCE estimates are defined as expectations with respect to the discrete probability distribution used to reparameterize the parameters of interest. To solve underdetermined systems of equations, the use of prior information is unavoidable and by itself is not a caveat regarding the use of GME techniques. It is in fact this specific feature, i.e. the flexibility in formulating prior information, that makes the GCE/GME framework of analysis so interesting to applied modelers who seek plausible simulation models and consistent data sets. The prior information actually employed is, however, a result of interactions between chosen support points and the reference distributions on the cho-

sen supports as well as the final weighting on support points implied by the maximum entropy criterion. The total effect of this interaction — especially for applications with many parameters and more than two support points — is not transparent. Furthermore, the introduction of a set of probabilities for each parameter to be estimated increases the computational demand on solving complex problems, which renders some very complex data reconciliation and estimation exercises intractable with currently available hardware and optimization solvers. In the next section we develop a Bayesian alternative to the GME approach which allows a direct and transparent formulation of prior information and potentially reduces the computational demand significantly.

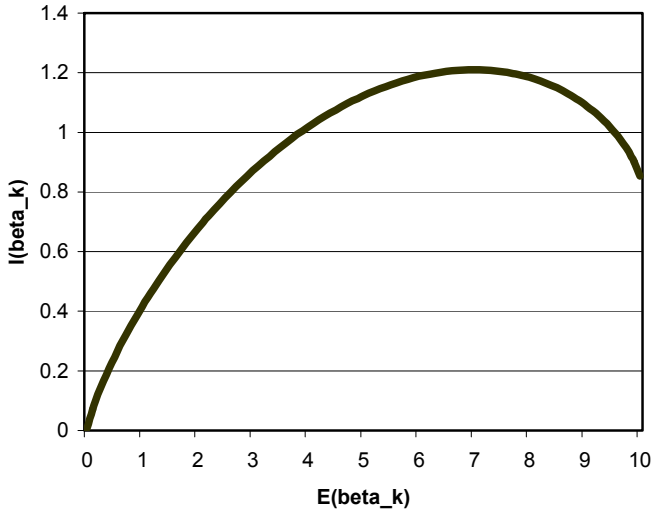

Figure 2.Prior weighting of parameter expectations with the cross-entropy criterion. Source: Minimum cross entropy calculations with sk1 = 0, sk2 = 10, and a reference distribution where qk1 = 0.3, qk2 = 0.7.

3 A Bayesian approach to the solution of underdetermined systems

To motivate the general concepts underlying the Bayesian alternative to GCE/GME we first reconsider the linear model without noise used in the previous section. We then extend the approach to a general system of underdetermined structural equations.

3.1 The Linear Model Revisited

The Bayesian approach to parameter estimation treats model parameters as stochastic variables. In this context the method distinguishes between the prior density, $p(\boldsymbol{\beta})$, summarizing prior information on parameters, the Likelihood function, $L(\boldsymbol{\beta}|\mathbf{y})$, representing information obtained from the data in conjunction with the assumed model, and the posterior density, $h(\boldsymbol{\beta}|\mathbf{y})$, which is the result of combining prior and data information based on Bayes' theorem. The relationship between these three elements can be expressed as (e.g. ZELLNER 1971, p.14)

$$h(\boldsymbol{\beta}|\mathbf{y}) \propto p(\boldsymbol{\beta}) L(\boldsymbol{\beta}|\mathbf{y}), \tag{10}$$

where the posterior density is proportional to the prior density multiplied by the Likelihood function. The posterior density allows drawing statistical inference about $\boldsymbol{\beta}$ using probability statements or by deriving point estimates that are optimal with respect to some loss criterion. For example, the mean of the posterior (density) is the value which minimizes quadratic loss.

Through appropriate interpretation of its components, the GME approach to estimating the parameters of the underdetermined linear model given in the previous section can be subsumed within the Bayesian formalism. For the case of two support points, using (8) and suppressing the GCE/GME expectation operator henceforth by simply representing the resultant estimator by $\boldsymbol{\beta}$, the GME optimization problem can be represented as

$$\max_{\boldsymbol{\beta}} \left\{ h(\boldsymbol{\beta}|\mathbf{y}) \propto p(\boldsymbol{\beta}) L(\boldsymbol{\beta}|\mathbf{y}) \propto \left[\sum_{k=1}^{K} H(\beta_k) \right] I_{\{\boldsymbol{\beta}:\, \mathbf{y}=\mathbf{X}\boldsymbol{\beta}\}}(\boldsymbol{\beta}) \right\} \tag{11}$$

where $I_A(\boldsymbol{\beta})$ is the standard indicator function that takes the value 1 when $\boldsymbol{\beta} \in \mathbf{A}$ and equals 0 otherwise. If $H(\beta_k)$ is chosen according to (8), the optimal value for $\boldsymbol{\beta}$ will be equal to the optimal $E\boldsymbol{\beta} = \mathbf{Sp}$ obtained in the GME solution, with an analogous result holding for GCE with $H(\beta_k)$ defined appropriately. In the Bayesian context, the objective function can be interpreted as the joint posterior density of the model parameters,

$h(\boldsymbol{\beta}|\mathbf{y})$, defined via a prior density defined by $p(\boldsymbol{\beta}) \propto \sum_{k=1}^{K} H(\beta_k)$ [2] that is multiplied by a likelihood function that assigns zero weights to values of $\boldsymbol{\beta}$ that do not satisfy the linear model constraints $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ and a positive constant weight to the values of $\boldsymbol{\beta}$ that are compatible with the data and the linear model relationship.[3] This implies zero *posterior* density weights for the values of $\boldsymbol{\beta}$ not satisfying the constraints and differential *posterior* weighting according to the prior (density) for all other values of $\boldsymbol{\beta}$. The value of $\boldsymbol{\beta}$ that maximizes $h(\boldsymbol{\beta}|\mathbf{y})$ is the mode of the posterior distribution of $\boldsymbol{\beta}$, which leads to the *Highest Posterior Density* (HPD)-estimate of $\boldsymbol{\beta}$.[4]

The preceding interpretation of GCE/GME within the Bayesian formalism suggests a general Bayesian alternative to the entropy approach that has three useful characteristics: (1) it can be formulated such that it is fully equivalent to the GCE/GME approach if support point choice and implicit weighting by the entropy criterion are appropriately represented, (2) the prior information on unknowns can be transparently formulated by assigning any appropriate prior density $p(\boldsymbol{\beta})$ *directly* to the unknowns, and (3) the optimization model has a smaller number of variables and, for an appropriate choice of the prior density functions, can be less computationally demanding.

---

[2] The function $\sum_{k=1}^{K} H(\beta_k)$ would need to be scaled appropriately to integrate to unity in order to be interpreted as a proper density, but this scaling is irrelevant for the outcome of the maximization.

[3] In the classical case of a linear model with noise, the Likelihood function would also have the error variance as an argument and would imply some continuous differential weighting according to the assumed error distribution. All that can be learned from the current model (the underdetermined data constraints, without noise) is which parameter vectors satisfy the data constraints and which do not, which motivates the dichotomous nature of the likelihood weighting in this case.

[4] Using the mode of the posterior for estimation was suggested before in the context of well-posed estimation problems, for example by DeGroot (1970), who called the estimator "generalized maximum likelihood". More frequently used terms are "maximum a-posteriori estimator" and "posterior mode estimator". In accordance with the Bayesian confidence intervals we prefer the HPD-estimator.

Having motivated the Bayesian alternative with a basic underdetermined linear model example, we now turn to a more general treatment of the Bayesian solution to underdetermined systems and the connection to entropy-based approaches.

3.2 General Structural Equation System

The general mathematical problem now being addressed is one where there are M equations, represented in vector function form as $\mathbf{g}(\mathbf{z}) = \mathbf{0}$, involving an unknown (K×1) vector argument $\mathbf{z}$, with M < K, so that the system of equations *underdetermines* the unknown vector $\mathbf{z}$.[5] Thus, in the absence of any additional information, and assuming the original equation system $\mathbf{g}(\mathbf{z}) = \mathbf{0}$ is consistent so that at least some solution actually exists, then indeterminacy implies that there is generally an infinite number of solution vectors that solve the system of equations.

One method of obtaining a unique solution to the system of equations is to choose $\mathbf{z}$ so as to optimize an extremum metric $v(\mathbf{z})$, subject to the constraints that $\mathbf{g}(\mathbf{z}) = \mathbf{0}$. So long as there exists a unique optimum of $v(\mathbf{z})$ within the feasible space of $\mathbf{z}$ values determined by $\mathbf{z} \in \Psi = \{\mathbf{z}: g(\mathbf{z}) = \mathbf{0}\}$, a unique solution to the original equation system can be identified. In general terms, such a solution could be represented as

$$\mathbf{z}^* = \arg\max_{\mathbf{z}} \left\{ v(\mathbf{z}) \text{ s.t. } g(\mathbf{z}) = \mathbf{0} \right\} \tag{12}$$

where it is assumed without loss of generality that *maximization* is the type of optimization pursued.

In fact *any* extremum metric $v(\mathbf{z})$ that exhibits an optimum within the feasible space $\mathbf{z} \in \Psi$ defines a possible solution to the equation system. There is thus a problem of deciding *which* metric to optimize, which in turn determines *which* solution from among a generally infinite number will be chosen as *the* solution to the original equation system. In general, any of the solutions in Ψ can be obtained given an appropriate corresponding choice of extremum metric $v(\mathbf{z})$. Thus, the solution obtained to a system of equations in this way is only defensible to the extent that the extremum metric used to obtain that solution is defensible. Before returning to this issue we discuss some necessary conditions for the solution.

Assume that the equation system of $\mathbf{g}(\mathbf{z}) = \mathbf{0}$ is a collection of functionally independent equations, so that the equations effectively determine M of the z_i's as a func-

---

[5] The elements of $\mathbf{z}$ are not restricted to model parameters. They could also represent unknown variable values in a data reconciliation exercise where data are measured with errors or not observed at all.

tion of the remaining K-M z_i-values. It is not necessary, conceptually, that *explicit* solutions exist for M of the variables in terms of the other K-M variables, but only that solutions exists. The solution might only be implicitly defined (which would then require numerical solution techniques). It is apparent that a general *necessary* condition for an extremum solution to exist is that $v(\mathbf{z})$ for $\mathbf{z} \in \Psi$ be informative, i.e. non-constant, in at least K-M of the variables in the vector $\mathbf{z}$. Among other things, this means that $v(\mathbf{z})$ cannot be uniform (or "uninformative" in prior distribution parlance) in more than M of the z_i arguments.[6] We note that there are other conditions that might be necessary in any given application, because depending on the nature of the equations in the system, it may be that informative information would have to exist on a specific as opposed to an arbitrary subset of $\mathbf{z}$ arguments given the solution space to $\mathbf{g}(\mathbf{z}) = \mathbf{0}$. It should also be noted that if $v(\mathbf{z})$ is informative on precisely K-M variables in the $\mathbf{z}$ vector, then the solution can be trivial in the sense that unconstrained optimization of the $v(\mathbf{z})$ metric in these K-M dimensions could be pursued independent of the equation system $\mathbf{g}(\mathbf{z}) = \mathbf{0}$ to determine K-M of the unknowns. The remaining arguments in the $\mathbf{z}$ vector could then be solved based on the relationships among the z_i's determined by the equation system.

Given that the data information serves only to narrow the feasible space of solutions for the unknowns and is otherwise uninformative, a useful and defensible choice for the extremum metric, $v(\mathbf{z})$, is the additional prior information held by the analyst, which summarizes the available non-data information on $\mathbf{z}$. If $p_i(z_i)$ represents general prior distribution weights on the possible solution values for the i^{th} component of the $\mathbf{z}$ vector, and if the prior weightings of the different components are considered to be independent, then the optimization metric used to obtain a solution to the equation system could be specified as

$$v(\mathbf{z}) = p(\mathbf{z}) = \prod_{i=1}^{K} p_i(z_i) \tag{13}$$

as example of which was given in (11). In the absence of independence, $p(\mathbf{z})$ can represent any joint prior distribution on potential solution values $\mathbf{z}$.

---

[6] Given this observation, it is clear that the GME approach to solving underdetermined systems works because it "automatically" implies a non-uniform prior weighting with respect to the variable of interest.

Now consider Bayes' rule applied to the problem of solving the equation system for **z**. In the absence of any information that would link **z** values to data and allow a likelihood function to be specified, the likelihood function would be considered undetermined or undefined. In this case, the Bayesian posterior and prior on the **z** vector would be identical and the maximization of the prior v(**z**) = p(**z**) would yield the maximum of the posterior. But in the current problem context the system of equations **g**(**z**) = **0** in effect constrains the support of the posterior h(**z**) to $\mathbf{z} \in \Psi = \{\mathbf{z} : \mathbf{g}(\mathbf{z}) = \mathbf{0}\}$. The Likelihood function in this case can be interpreted as an indicator function $I_{\psi}(\mathbf{z})$ that assigns weights of 1 to admissible values of **z** and 0 otherwise (It is straightforward to introduce non-dichotomous likelihood weightings if the model specification supports such information. Such a case will be illustrated in the applications section). The posterior is then in the form

$$h(\mathbf{z}) \propto p(\mathbf{z}) I_{\psi}(\mathbf{z}) . \tag{14}$$

Consequently, the argument that maximizes the prior probability p(**z**) subject to the constraint $\mathbf{z} \in \Psi$ (or **g**(**z**) = **0**) will provide a *Bayesian highest posterior density (HPD) solution* to the equation system.

3.3 Solutions for Uninformative Priors

For reasons discussed earlier, the HPD approach to solving the system of equations cannot be applied in cases where the prior weighting on solution values is not sufficiently informative, i.e. p(**z**) cannot be uniform in more than M of the z_i arguments as the optimum will not be unique. However, in this case, solving for the posterior *mean*, which is the posterior risk-minimizing Bayesian estimate under quadratic loss, will generally be possible as long as the uniform distribution is proper in the sense of integrating to 1. This will follow naturally if the prior support space is a priori compact, so that there is indifference among values of **z** within a hyperrectangle of values having finite boundaries. In the extreme case of no informative prior information at all, the values in the support space defined by the equation system. $\Psi = \{\mathbf{z} : \mathbf{g}(\mathbf{z}) = \mathbf{0}\}$, are all equally likely, so that the Bayes' posterior mean solution would be the mean of **z** from among all equally likely values in this support space. A computational method of finding such a solution would be to draw uniform random outcomes of **z** from Ψ, forming their sample mean, and for large enough simulated sample sizes, the sample mean would converge in probability (or almost surely) to the true mean by the weak (strong) law of large numbers.

In some cases, the posterior mean solution might be identifiable analytically. For example, consider again the underdetermined linear model without noise,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{X}$ is a T × K matrix, with T < K and rank ($\mathbf{X}$) = T. Since rank ($\mathbf{X}$) is smaller than the number of columns, an infinite number of solutions exist for $\boldsymbol{\beta}$. These solutions will form a hyperplane in $\Re^K$, which can be described by a linear function of the form

$$\boldsymbol{\beta} = \boldsymbol{\beta}^0 + \mathbf{B}\boldsymbol{\xi} \qquad (15)$$

where $\mathbf{B}$ is a K × (K – T) matrix that is a basis for the subspace of solutions to the *homogeneous model* $\mathbf{0} = \mathbf{X}\boldsymbol{\beta}$, and $\boldsymbol{\xi}$ is an arbitrary (K – T) vector. This follows from the following results:

Lemma 1: *Any* solution $\boldsymbol{\beta}^*$ to the inhomogeneous linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ can be written as the sum of a *particular* solution $\boldsymbol{\beta}^0$ to the inhomogeneous model plus *some* solution $\boldsymbol{\beta}^1$ to the homogeneous linear model $\mathbf{0} = \mathbf{X}\boldsymbol{\beta}$ (e.g. DE LA FUENTE 2000, p. 197).

Lemma 2: $\boldsymbol{\beta}^1$ in lemma 1 can be written as $\mathbf{B}\boldsymbol{\xi}$, for *some* matrix $\mathbf{B}$ and *any* vector $\boldsymbol{\xi}$ of dimension (K – T).

If there are uniform priors for at least K – T of the elements of $\boldsymbol{\beta}$, then those priors constitute a hyperrectangle $\mathbf{U}$ in $\Re^K$, and the posterior mean is the geometrical centre of the intersection between the solution hyperplane and the hyperrectangle $\mathbf{U}$. We can then compute the posterior mean through a sequence of four steps that include first finding $\boldsymbol{\beta}^0$, then computing the matrix $\mathbf{B}$**,** next finding the intersection between the solution hyperplane and the prior hyper rectangle, and finally finding the center of the intersection. A specific algorithm for accomplishing these steps is as follows:

Step 1. A particular solution $\boldsymbol{\beta}^0$ to the inhomogeneous system can be found by solving $\boldsymbol{\beta}^0 = \mathbf{X}^+\mathbf{y}$, where $\mathbf{X}^+$ is the generalized inverse of $\mathbf{X}$.

Step 2. Since K > T and rank ($\mathbf{X}$) = T, K – T columns of $\mathbf{X}$, together forming the matrix $\mathbf{X}_{(i)}$, can be written as linear combinations of the other T columns, which are kept in the T × T matrix $\mathbf{X}_{(-i)}$. The coefficients of each of the K – T columns in $\mathbf{X}_{(i)}$ can be chosen arbitrarily. If this is repeatedly done for each column in $\mathbf{X}_{(i)}$, the following expression is obtained, where the columns of $\mathbf{B}_{(i)}$ are the arbitrary coefficient vectors for $\mathbf{X}_{(i)}$:

$$-\mathbf{X}_{(i)}\mathbf{B}_{(i)} = \mathbf{X}_{(-i)}\mathbf{B}_{(-i)}$$

Choosing the (T − K) × (T − K) identity matrix for $\mathbf{B}_{(i)}$, the above expression can be solved for $\mathbf{B}_{(-i)} = -(\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}$, and $\mathbf{B}$ can be obtained by vertical concatenation of $\mathbf{B}_{(i)}$ and $\mathbf{B}_{(-i)}$, keeping the rows in proper order.

Step 3. Find the values of $\boldsymbol{\xi}$ for which the resulting $\boldsymbol{\beta}$ is inside the prior hyperrectangle. This can be done by trial and error if the dimension of $\boldsymbol{\xi}$ is low, and numeri-

cally by repeated linear programming (solving $\{\min \mathbf{p}'\boldsymbol{\xi}: \boldsymbol{\beta}_0 + \mathbf{B}\boldsymbol{\xi} = \boldsymbol{\beta} \in \mathbf{U}\}$, with $\mathbf{p}$ being some permutation of (-1) and 1 of length K – T, for all such permutations) if the dimension is higher. The set of solutions will be the bounds of a hyperrectangle in $\Re^{K-T}$.

Step 4. Since $\boldsymbol{\beta}$ is linearly dependent on $\boldsymbol{\xi}$, and $\boldsymbol{\beta}$ is uniformly distributed, the expected value of $\boldsymbol{\beta}$ is found by computing the geometrical mid point of the hyperrectangle found in step 3.

4 Illustrative Applications

This section presents two illustrative applications of the HPD-estimator based on underdetermined problem specifications that are typical of applications for entropy estimators: Balancing of a Social Accounting Matrix (SAM) and a linear regression problem.

4.4 Balancing a Social Accounting Matrix

In 1994, GOLAN, JUDGE AND ROBINSON (GJR) used entropy based estimators to create a consistent SAM. Variants of their approaches can be found in the empirical Computable General Equilibrium literature to prepare complete databases out of incomplete and uncertain data information.

The basic problem of balancing a SAM can be formulated as follows: find a square matrix of coefficients $\mathbf{A}$ and vectors $\mathbf{x}$ and $\mathbf{y}$ satisfying the equations

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{16}$$

$$\mathbf{A}'\boldsymbol{\iota} = \boldsymbol{\iota}. \tag{17}$$

with $\boldsymbol{\iota}$ the vector of ones of appropriate dimension. In general, information about the $\mathbf{x}$ and $\mathbf{y}$ are available from observable data, whereas the coefficient matrix $\mathbf{A}$ is difficult to obtain. A common situation is thus that $\mathbf{x}$ and $\mathbf{y}$ are given, and $\mathbf{A}$ needs to be determined subject to the restrictions (16) and (17), possibly given some prior information about $\mathbf{A}$, perhaps in the form of the same matrix for another region or for the same region for a different period. We take the example studied by GJR and provide a Bayesian alternative.

Table 1 in their paper provides the “true parameters”,

$$\mathbf{A} = \begin{bmatrix} 0.726 & 0.000 & 0.165 & 0.301 \\ 0.161 & 0.268 & 0.000 & 0.451 \\ 0.113 & 0.678 & 0.714 & 0.000 \\ 0.000 & 0.054 & 0.121 & 0.248 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} 62 \\ 56 \\ 91 \\ 266 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 140 \\ 145 \\ 110 \\ 80 \end{bmatrix}.$$

The authors proceed to construct a (synthetic) prior for **A** by multiplying each entry in **A** by a random number drawn from a normal distribution, N(1, .05). They present the outcome

$$\mathbf{A}^o = \begin{bmatrix} 0.730 & 0.000 & 0.172 & 0.278 \\ 0.159 & 0.259 & 0.000 & 0.480 \\ 0.111 & 0.688 & 0.694 & 0.000 \\ 0.000 & 0.053 & 0.135 & 0.243 \end{bmatrix},$$

and estimate **A** with GCE using $\mathbf{A}^o$ as a prior.

The GCE problem is

$$\begin{aligned} \min \quad & \mathbf{p}'\ln(\mathbf{p}/\mathbf{q}) \\ \text{such that} \quad & \mathbf{p} \geq \mathbf{0},\ \mathbf{Sp} = vec(\mathbf{A}),\ \mathbf{y} = \mathbf{Ax},\ \mathbf{A}\iota = \mathbf{1} \end{aligned} \quad (18)$$

where the prior probabilities $\boldsymbol{q}$ of the support point matrix **S** are selected so that $\mathbf{Sq} = \mathbf{A}^o$, with vec(**A**) being the operator that reshapes the matrix **A** to a column vector by vertically concatenating respective columns, and **p**/**q** as in section 2 the vector whose i^{th} element is p_i/q_i, Note that this approach requires the researcher to define a set of at least two (GJR use five) support points for each parameter, and also to define a corresponding set of prior probabilities such that the prior SAM is recovered. GJR use the same support points for all elements of **A**, and choose **q** using an initial GME estimation of $\mathbf{Sq} = \mathbf{A}^o$, which effectively doubles the computational effort needed to produce the final estimates of the **A** matrix.

Now construct an alternative Bayesian estimator for the same problem. The HPD framework allows the use of any prior distribution. Assume, for example, that the researcher had a-priori knowledge that the observed matrix $\mathbf{A}^o$ was generated as in GJR. Taking $\mathbf{A}^o$ as prior mean, and continuing to follow GJR, the corresponding prior density function would be vec(**A**) ~ N(vec($\mathbf{A}^o$),**Σ**), The covariance matrix **Σ** is set equal to a diagonal matrix with elements $\left(\text{vec}(\mathbf{A}^o)0.05\right)^2$, the square taken element-wise.

Formulating the HPD estimator as discussed previously, taking natural logs, and restricting the objective function to the terms that are relevant for optimization leads to the following extremum estimation problem:

$$\max_{\mathbf{A}} \left[\mathrm{vec}(\mathbf{A}) - \mathrm{vec}(\mathbf{A}^{o})\right]' \boldsymbol{\Omega}^{-1} \left[\mathrm{vec}(\mathbf{A}) - \mathrm{vec}(\mathbf{A}^{o})\right]$$

$$\text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y} \qquad (19)$$

$$\mathbf{A}\boldsymbol{\iota} = \mathbf{1}$$

For the synthetic data provided in GJR, GCE and HPD give very similar results, shown below (results for GCE as printed in GJR). Note that the HPD estimation tacitly assumed degenerate priors for **x** and **y**. The estimation is easily extended to encompass the fact that **x** and **y** are not known with certainty.

$$\mathbf{A}^{GCE} = \begin{bmatrix} 0.732 & 0.000 & 0.168 & 0.298 \\ 0.155 & 0.251 & 0.000 & 0.456 \\ 0.114 & 0.697 & 0.702 & 0.000 \\ 0.000 & 0.052 & 0.129 & 0.246 \end{bmatrix}, \quad \mathbf{A}^{HPD} = \begin{bmatrix} 0.731 & 0.000 & 0.167 & 0.299 \\ 0.157 & 0.248 & 0.000 & 0.456 \\ 0.112 & 0.699 & 0.702 & 0.000 \\ 0.000 & 0.053 & 0.131 & 0.245 \end{bmatrix}$$

As can be seen from (19), the choice of a normal prior distribution results in a weighted least squares approach implying numerically desirable properties for large scale problems. Compared to GME or GCE approaches, explicit accounting for support points and adding up constraints for probabilities are unnecessary and infeasibilities are less likely to lead to numerical problems. Other prior distributions can be flexibly accommodated and will be considered in the next example.

4.5 Regression models

In this section we consider an ill-posed linear regression model with and without noise, and characterized by three equations and four parameters. In total five cases are studied which are distinguished by the available prior information on parameters, and the type of estimation objective applied. All cases have in common that there is prior information available *only for two of the four parameters*:

1. Uniform priors given bounds [**u**,**v**], with parameters estimated by posterior means;

2. Symmetric triangular distributed priors over the support [**u**,**v**] and application of the Bayesian HPD-estimator;

3. GME estimation with [**u**,**v**] as range of support points and a uniform reference distribution, represented and solved equivalently as a Bayesian HPD-estimator;

4. Same as 2), but with priors distributed as Beta(2,2) between the bounds [**u**,**v**];

5. As in 4), but also including additive white noise;

True parameters for the model without noise, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, were chosen arbitrarily and the columns 2-4 of **X** were drawn from a normal distributions with means 20, 8 and 12 and variances equal to ¼ of the means. By multiplication with the selected true parameters, the true **y** was obtained. The procedure resulted in the following numbers.

$$\boldsymbol{\beta} = \begin{bmatrix} 10.0 \\ 0.5 \\ 1.5 \\ 1.0 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & 20.733 & 8.656 & 8.830 \\ 1 & 17.827 & 7.443 & 13.619 \\ 1 & 20.001 & 6.715 & 12.596 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 42.180 \\ 43.697 \\ 42.668 \end{bmatrix}$$

In all five cases, the support of the prior densities for β_2 and β_3 is defined by the interval

$$(\mathbf{u}, \mathbf{v}) = \begin{bmatrix} 0 & 0.868 \\ 0 & 2.903 \end{bmatrix}.$$

Note that the mid points between the bounds (**u**,**v**) are *not* equal to the true parameter values.

Case 1: Since we are dealing with a linear system with (-1) degrees of freedom, the vector $\boldsymbol{\xi}$ in equation (15) is a scalar, and all feasible $\boldsymbol{\beta}$ lie on a line segment limited by (**u**,**v**). Following the steps indicated in section 4, choosing the second column of **X** for $\mathbf{X}_{(i)}$, we obtain

$$\boldsymbol{\beta} = \begin{bmatrix} 0.1132 \\ 0.7284 \\ 1.8604 \\ 1.2300 \end{bmatrix} + \xi \begin{bmatrix} -43.2306 \\ 1.0000 \\ 1.5735 \\ 1.0054 \end{bmatrix} \tag{20}$$

for arbitrary ξ. In order for β_2 and β_3 to be within (**u**,**v**), it is required that $\xi \in$ (-0.7284, 0.1396). Since the uniform density indicates the same posterior density weight for all values for ξ in that interval and zero elsewhere, we can compute the posterior mean as the mid point of the interval, or $\hat{\xi} = -0.2944$. Inserting that value into the expression (20) gives us the point estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 12.842 \\ 0.434 \\ 1.397 \\ 0.934 \end{bmatrix}.$$

Case 2: Let the prior density for β_2 and β_3 have the same bounds as before, but now follow a symmetric triangular distribution, i.e. the mid point of the interval is favored. Now a unique posterior mode exists, and we may apply the HPD estimator. Since we strive to maximize the posterior, the piecewise linear formulation of the triangular density can be relaxed to three linear inequalities, each representing a side of the triangle. For ease of notation, we first introduce the subvector $\boldsymbol{\beta}_p$ consisting of the elements (β_2,β_3) for which there are priors, the corresponding subset of probability densities $\mathbf{p}_p$, and the linear mapping $\mathbf{g}$:[**u**,**v**] $\rightarrow$ [0,1]×[0,1]. Thus $g_k(\beta_k)$ expresses β_k in terms of the *share* of the distance from u_k to v_k, for k = 2,3.

The HPD estimator is then

$$\max_{\mathbf{p}_p, \boldsymbol{\beta}_p} \; p_2 p_3$$

subject to $\boldsymbol{\beta}_p$ being triangular distributed, i.e.

$$\mathbf{p}_p \le 4\mathbf{g}(\boldsymbol{\beta}_p), \;\; \mathbf{p}_p \le 4 - 4\mathbf{g}(\boldsymbol{\beta}_p), \;\; \mathbf{p}_p \ge 0,$$

and the data constraints,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

The outcome of the estimator is identical to the outcome of the posterior mean estimator and is not repeated here.

Case 3: Let (**u**,**v**) be support points for a GME estimation. Using the normalization $\mathbf{g}(\boldsymbol{\beta}_p)$ as before, the support points become (0,1), the probabilities of the supports become $\mathbf{p} = vec\left(\begin{bmatrix} \mathbf{g}(\boldsymbol{\beta}_p) \\ \mathbf{1}-\mathbf{g}(\boldsymbol{\beta}_p) \end{bmatrix}'\right)$ and we may write the GME estimator as

$$\max_{\boldsymbol{\beta}_p} \quad H$$

$$= -\left[\mathbf{g}(\boldsymbol{\beta}_p)'\ln\left(\mathbf{g}(\boldsymbol{\beta}_p)\right)+\left(\mathbf{1}-\mathbf{g}(\boldsymbol{\beta}_p)\right)'\ln\left(\mathbf{1}-\mathbf{g}(\boldsymbol{\beta}_p)\right)\right]$$

$$\text{subject to} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{u} \le \boldsymbol{\beta}_p \le \mathbf{v}$$

For the sake of illustration, we re-write this as a fully equivalent HPD estimation problem. Note that the GME problem is equivalent to maximizing e^H (the maximum is maintained under monotonic transformation). Substitution and some algebra lead to the equivalent HPD problem

$$\max \quad f_2(\beta_2)f_3(\beta_3)$$

$$\text{subject to} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{u} \le \boldsymbol{\beta}_p \le \mathbf{v}$$

where $f_k(\beta_k) = c g_k(\beta_k)^{-g_k(\beta_k)}(1-g_k(\beta_k))^{(g_k(\beta_k)-1)}$ (for $k = 2,3$) is a probability density function if the constant c is chosen properly ($c \approx 0.6$ makes f integrate to unity, non-negative values are prevented by the mapping $\boldsymbol{g}$ and the bounds $\mathbf{u} \le \boldsymbol{\beta}_p \le \mathbf{v}$). We see that, interpreted in this way, the GME estimator is an instance of a HPD estimator. The GME estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 12.586 \\ 0.440 \\ 1.406 \\ 0.940 \end{bmatrix}$$

Case 4: The upper and lower bounds on two of the parameters make it natural to describe the estimates in terms of a fraction of the distance between the bounds (as expressed in the mapping **g**). In such cases the beta distribution is sometimes used. Let the distribution of β_k for $k = 2,3$ be such that $g_k(\beta_k) \sim$ beta(2,2). The HPD estimate of $\boldsymbol{\beta}$ with beta-distributed priors is identical to the GME estimate at least up to three decimal places in this case, and not repeated here.

Case 5: In the previous cases we assumed that **X** and **y** were observable without noise. We now introduce white noise for **y** by adding iid errors drawn from N(0,1) whereas **X** is still assumed to be known with certainty. The resulting stochastic vector of left hand side variables is denoted by $\mathbf{y}_s$, and the outcome of a draw was

$$\mathbf{y}_s = \mathbf{y} + \mathbf{e} = \begin{bmatrix} 44.064 & 42.976 & 41.369 \end{bmatrix}'$$

where **e** is an outcome of the error $\boldsymbol{\varepsilon}$, and the system to estimate is $\mathbf{y}_s = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

If we consider $\boldsymbol{\varepsilon}$ yet another parameter to determine, and introduce the prior information that errors were drawn from N(0,1) and still assume that β_2 and β_3 belong to the same beta distributions as in the previous example, the HPD estimator for $\boldsymbol{\beta}$ is found by solving the problem

$$\max_{\boldsymbol{\beta}_p, \boldsymbol{\varepsilon}} \quad h = \prod_{k \in \{2,3\}} p_b\left(g_k(\beta_k)\right) \prod_{i=1}^{4} p_e\left(\varepsilon_i\right)$$

$$\text{subject to} \quad \mathbf{u} \le \boldsymbol{\beta}_p \le \mathbf{v}$$

$$\mathbf{y}_s = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $p_b(\cdot)$ being the beta density function as in the previous example and $p_e(\cdot)$ being the standard normal univariate density. A form more easily computed is obtained by recognizing that $p_b(x) = 6(x\text{-}x^2)I_{(0,1)}(x)$ and taking the logarithm of the objective function, which then becomes

$$\max \ \ln(h) = \sum_{k\in\{2,3\}} \ln\left(g_k(\beta_k) - g_k(\beta_k)^2\right) - \tfrac{1}{2}\sum_{i=1}^{4}\varepsilon_i^2$$

The resulting estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 16.668 & 0.379 & 1.820 & 0.419 \end{bmatrix}'$.

4.6 Summary and Conclusions

This paper presents a Bayesian alternative to the solution of underdetermined systems of equations. First, we reviewed the GME-GCE approach in the context of estimating an underdetermined linear model without noise and identified the effective prior information as a combined effect between supports, reference probabilities, and the entropy criterion. It was indicated that a "uniform distribution over supports" does not imply a "non-informative" prior on the parameter of interest, but rather a clear prior preferential weighting on estimation outcomes. In the suggested Bayesian alternative the underdetermined model equations and the data represent the "Likelihood" information. Deviating from standard Likelihood functions of conventional models with a predefined family of distributions, the Likelihood implies a constant positive weight for all possible solutions of the model equations and a zero weight for infeasible values. All prior information is represented in a standard Bayesian way via prior probability densities on model parameters. Highest Posterior Density (HPD) estimates are obtained using an optimization algorithm.

The Bayesian approach can be formulated to mimic the behavior of GME-GCE models perfectly. However, more interesting is its general structure allowing full flexibility in formulating directly and transparently the prior information held by the analyst. For a unique solution to exist, a certain amount of informative prior information is necessary. However, if this is not the case, a solution based on the posterior mean can — at least conceptually — still be provided.

The suggested approach lends itself easily to the type of problems currently solved with GME or GCE techniques. It has been successfully applied to large scale estimation and calibration exercises (BRITZ, WITZKE AND HECKELEI 2004; JANSSON 2007). It facilitates the peer review of methodology and underlying assumptions by making the employed prior information directly visible. Further research could examine computational approaches for generating posterior mean estimates under insufficient identifying prior information.

References

ARNDT, C., S. ROBINSON AND F. TARP (2002). Parameter Estimation for a Computable General Equilibrium Model: a Maximum Entropy Approach. *Economic Modelling 19*, 375-398.

BRITZ, W. AND C. WIECK (2002). *Completeness and Consistency in a Multidimensional Data Base using Constrained Simultaneous Estimation Techniques.* Poster presented at the 42. Jahrestagung der GEWISOLA (German Agricultural Economics Society).

BRITZ, W., H.P. WITZKE AND T. HECKELEI (2004). Estimating Trade Matrices and Supply Utilization Accounts using a Bayesian Estimator." Final Report of EuroCARE to FAO, Bonn.

DEGROOT, M.H. 1970. *Optimal statistical decisions*. New York: McGraw-Hill.

DE LA FUENTE, A. (2000). *Mathematical methods and models for economists*. New York: Cambridge University Press.

GOLAN, A., G. JUDGE AND S. ROBINSON (1994). "Recovering Information from Incomplete or Partial Multisectoral Economic Data." *The Review of Economics and Statistics.* 76(3), 541-549.

GOLAN, A., G. JUDGE AND D. MILLER (1996). *Maximum Entropy Econometrics*. Chichester UK: Wiley.

GOLAN, A., G. JUDGE AND J.M. PERLOFF (1996). A Maximum Entropy Approach to Recovering Information From Multinomial Response Data." *Journal of the American Statistical Association 91(434)*, 841-853.

HECKELEI T. AND H. WOLFF (2003). Estimation of Constrained Optimisation Models for Agricultural Supply Analysis Based on Generalised Maximum Entropy. *European Review of AgriculturalEconomics 30(1)*, 27-50.

HOWITT, R.E. AND A. REYNAUD (2003). Spatial Disaggregation of Agricultural Production Data using Maximum Entropy. *European Review of Agricultural Economics 30*, 359-387.

JANSSON, T. (2007). Econometric specification of constrained optimization problems. dissertation, University of Bonn, Bonn, forthcoming.

LENCE, H.L. AND D. MILLER (1998a). Estimation of Multioutput Production Functions with Incomplete Data: A Generalized Maximum Entropy Approach." *European Review of Agricultural Economics 25*, 188-209.

— (1998b). Recovering Output Specific Inputs from Aggregate Input Data: A Generalized Cross-Entropy Approach. *American Journal of Agricultural Economics 80(4)*, 852-867.

LÉON Y., L. PEETERS, M. QUINQU AND Y. SURRY (1999). The Use of Maximum Entropy to Estimate Input-Output Coefficients from Regional Farm Accounting Data. *Journal of Agricultural Economics 50*, 425-439.

MITTELHAMMER, R.C., G.G. JUDGE AND D.J. MILLER (2000). *Econometric Foundations*. Cambridge: Cambridge University Press.

OUDE LANSINK, A.G.J.M (1999). Generalized Maximum Entropy and Heterogeneous Technologies. *European Review of Agricultural Economics 26,* 101-115.

PARIS, Q. (2001). Symmetric Positive Equilibrium Problem: A Framework for Rationalizing Economic Behavior with Limited Information. *American Journal of Agricultural Economics 83(4),* 1049-1061.

PARIS, Q. AND R.E. HOWITT (1998). An Analysis of Ill-Posed Production Problems Using Maximum Entropy. *American Journal of Agricultural Economics 80(1)*, 124-138.

PRECKEL, P.V. (2001). Least Squares and Entropy: A Penalty Function Perspective. *American Journal of Agricultural Economics 83(2)*, 366-377.

ROBILLIARD, A.-S. AND S. ROBINSON (2003). Reconciling Household Surveys and National Accounts Data using a Cross Entropy Estimation Method. *Review of Income and Wealth 49,* 395-406.

ROBINSON, S., A. CATTANBO AND M. EL-SAID (2000). Updating and Estimating a Social Accounting Matrix using Cross Entropy Methods. *Economic Systems Research 13*, 47-67.

WITZKE, H.P. AND W. BRITZ (1998). A Maximum Entropy Approach to the Calibration of Highly Differentiated Demand Systems. Working Paper, CAPRI 98-06, University of Bonn, Bonn.

ZELLNER, A. (1971). *Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

ZHANG, X. AND S. FAN (2001). Estimating Crop-Specific Production Technologies in Chinese Agriculture: A Generalized Maximum Entropy Approach. *American Journal of Agricultural Economics 83(2)*, 378-388.