ELSEVIER

Journal of Econometrics 101 (2001) 165–193

JOURNAL OF Econometrics

www.elsevier.nl/locate/econbase

A simultaneous estimation and variable selection rule

Amos Golan*

Department of Economics, American University, Roper 200, 4400 Massachusetts Ave., NW, Washington, DC 20016, USA

Received 8 January 1999; received in revised form 18 July 2000; accepted 1 August 2000

Abstract

A new data-based method of estimation and variable selection in linear statistical models is proposed. This method is based on a generalized maximum entropy formalism, and makes use of both sample and non-sample information in determining a basis for coefficient shrinkage and extraneous variable identification. In contrast to tradition, shrinkage and variable selection are achieved on a coordinate-by-coordinate basis, and the procedure works well for both ill- and well-posed statistical models. Analytical asymptotic results are presented and sampling experiments are used as a basis for determining finite sample behavior and comparing the sampling performance of the new estimation rule with traditional competitors. Solution algorithms for the non-linear inversion problem that results are simple to implement. © 2001 Elsevier Science S.A. All rights reserved.

JEL classification: C13; C14; C5

Keywords: Shrinkage estimator; Maximum entropy; Extraneous variables; Squared error loss; Data weighted prior; Subset selection

1. Introduction

Given a finite and non-experimental data set economists face two basic decisions. The first is the decision about the set of non-extraneous variables and

* Corresponding author. Tel.: 001-202-885-3783; fax: 001-202-885-3790.
E-mail address: agolan@american.edu (A. Golan).

PII: S0304-4076(00)00078-6

related functions to include in the design matrix. The second is the choice of the estimation rule to be used in recovering estimates of the unknowns in the corresponding parameterized model. Traditionally, the model selection and estimation problems have been separated (Kempthorne, 1987) where the choice of the estimation rule is done after the choice of model and variable selection. In reality, however, the two decisions are interdependent as the sampling performance of the estimation procedure is conditioned by the model selection choice.

The objective of this paper is to construct an estimation rule that *simultaneously* considers the problems of statistical variable selection and parameter estimation. This semi-parametric estimation rule has its roots in information theory and builds on the generalized maximum entropy (GME) and generalized cross-entropy (GCE) estimation rules proposed by Golan et al. (1996). Two main properties of this proposed estimation and variable selection rule are worth noting. The first is that this method makes use of both sample and non-sample (prior) information on a coordinate basis. The second is that the prior for each coordinate (or variable) is determined endogenously during the optimization. That is, the optimization is done simultaneously with respect to both the posterior and an endogenously determined weight imposed on a convex combination of informative and non-informative priors.

The statistical model and variable selection rules are specified and reviewed in Section 2. The traditional maximum entropy formalism and the GCE estimators are reviewed in Section 3. In Section 4, the proposed flexible data weighted prior (DWP) estimation rule is formulated and its corresponding sampling characteristics are discussed. In Section 5, some sampling experiments are reported. These experiments demonstrate the empirical risk and variable identification performance of the DWP estimator. Section 6 summarizes and concludes this paper.

2. The problem – background and a brief review

To make exposition easier, consider the traditional linear statistical model. Let us assume that we are unable to directly measure the unknown K-dimensional parameter vector β. Instead, we observe a T-dimensional vector of noisy sample observations $\boldsymbol{y} = (y_1, y_2, \dots, y_T)'$ that are consistent with the underlying data generation process model

$$\boldsymbol{y} = X\beta + \boldsymbol{e}, \tag{2.1}$$

where X is a fixed $(T \times K)$ full column rank design matrix known to the experimenter. Further, it is assumed that

Assumption A1. $\beta \in B$ where B is a convex set.

Example 1. $B = \{\beta \in \Re^K | \beta_k \in (\underline{z}_k, \bar{z}_k), \quad k = 1, 2, \ldots, K\}$.

Given the data generation process described by (2.1), the objective is to derive an estimator that uses *minimum assumptions* on the likelihood structure and simultaneously identifies the extraneous variables on a coordinate-wise basis. Before developing this estimator, we briefly review some of the current variable selection criteria and models.

Within the context of statistical model (2.1), the variable selection problem may be described in the following way. An investigator has a single and non-experimental sample of data that is known to have the linear functional form of (2.1). Suppose that some covariates are unrelated to the prediction of $\boldsymbol{y}$, so the true relationship may be characterized by a lower-dimensional parameter space β_0. Consequently, we visualize a K-dimensional parameter space that includes the set of K_0 relevant variables, plus an additional possible $K - K_0$ extraneous variables with coefficients of zero. Thus, the design matrix consistent with the data generation process is a proper subset of the included variables. In terms of variable selection, there are 2^K possible models that can be obtained from the general model (2.1). However, in most cases we employ our knowledge from economic theory to reduce our choice, of possible models, only to those remaining variables that include some uncertainties.

Traditionally, there are two elements in the criterion function for the various variable selection procedures. One element involves a measure of 'goodness of fit' while the other involves a penalty for complexity, which is a function of the number of variables K_0 in one of the competing models. Following Zheng and Loh (1995), let $MG(K_0)$ be the measure of goodness of fit for the competing model K_0, then the various sampling theory estimators $\tilde{K}_0$ are asymptotically equivalent to

$$\tilde{K}_0 = \underset{0 \leqslant K_0 \leqslant K}{\arg\min} \{MG(K_0) + h(K)\hat{\sigma}^2\}, \tag{2.2}$$

where $\hat{\sigma}^2$ is a consistent estimate of σ^2 and $h(K)$ is some non-linear function (e.g., Hocking, 1976; Amemiya, 1980; Laud and Ibrahim, 1995; Mallows, 1973; Miller, 1990; Mitchell and Beauchamp, 1988).[1] For a recent review and development of both parametric and non-parametric approaches to variable selection, within a general discrimination framework, see Lavergne (1998).

[1] Among the more common methods are the C_p (Mallows, 1973) and the Akaike's (1974) Information Criterion, AIC. For example, if $h(K) = 2K/T$ and $MG(K_0) = \|\boldsymbol{y} - X\beta_{K_0}\|^2$, Eq. (2.2) reduces to the AIC. For more discussion of the statistical properties of these criteria see for example Shibata (1981) and Zhang (1992). Other criteria, such as Schwarz's (1978) Criterion SC or the Zheng and Loh (1995) generalization of C_p as well as cross validation (Breiman and Spector, 1992; Stone, 1974) and penalized likelihood (Sin and White, 1996) are quite commonly used.

In most cases, however, we have some prior knowledge and/or non-sample information that come from economic theory and from understanding our data. The variable selection methods discussed above do not 'quantify' this knowledge. But some of this knowledge (priors) may be quantified and incorporated directly. This is addressed in the Bayesian approach to model selection, which also involves setting prior probabilities over the large class of models being considered together with setting the corresponding priors for the parameters of each model (see, for example, Zellner and Vandaele, 1975; George and McCulloch, 1993; Geweke, 1994; Kass and Raftery, 1995; Zellner, 1996b).[2] The model developed here also uses prior information, but introduces this knowledge in a different way. This is discussed in the next section.

3. A cross-entropy estimator – review, background and motivation

As an alternative to traditional frequentist shrinkage, pre-test and Bayes estimators for the location vector in (2.1), Golan et al. (1996) proposed, for both the symmetric and non-symmetric cases and for both well- and ill-posed problems, a new shrinkage estimation rule. This estimation rule is based on the entropy measure of Shannon (1948), a reformulation of the maximum entropy (ME) formalism of Jaynes (1957a, b; 1984), Levine (1980), Shore and Johnson (1980), Skilling (1989) and Csiszár (1991), and the cross entropy principle of Gokhale and Kullback (1978), Good (1963), and Kullback (1959). Before developing the new entropy-based variable selection model, a brief review and background, for both the ME and GME, are presented.

3.1. The classic maximum entropy model

To provide a basis for understanding the philosophy of the ME approach, consider the following example. Let $\Theta = \{\theta_1, \theta_2, \dots, \theta_M\}$ be a finite set and $\boldsymbol{p}$ be a probability mass function on Θ. The Shannon's information criterion, called entropy, is $H(\boldsymbol{p}) = -\sum_{i=1}^{M} p_i \log p_i$ with $0 \log 0 \equiv 0$. This information criterion measures the uncertainty, or informational content, in Θ which is implied by $\boldsymbol{p}$. The entropy-uncertainty measure $H(\boldsymbol{p})$ reaches a maximum when $p_1 = p_2 = \cdots = p_M = 1/M$ and a minimum with a point mass function. Given the entropy measure and structural constraints in the form of moments of the data (distribution), Jaynes (1957a, b) proposed the maximum entropy (ME) method, which is to maximize $H(\boldsymbol{p})$ subject to these structural constraints. If no constraints (data) are imposed, $H(\boldsymbol{p})$ reaches its maximum value and the distri-

[2] For a good review and recent developments of the Bayes factors within Bayesian testing and model selection see Berger and Mortera (1999).

bution of the p's is a uniform one. Thus, if we have partial information in the form of some moment conditions, Y_t ($t = 1, 2, \ldots, T$), where $T < M$, the maximum entropy principle prescribes choosing the $p(\theta_i)$ that maximizes $H(\boldsymbol{p})$ subject to the given constraints (moments) of the problem. The solution to this underdetermined problem is

$$\hat{p}(\theta_i) \propto \exp\left\{ -\sum_t \lambda_t Y_t(\theta_i) \right\}, \tag{3.1}$$

where λ are the T Lagrange multipliers.

If prior information, q_i, concerning the unknown p_i exists, then one alternative to the ME approach is to minimize the Kullback–Leibler (K–L) entropy distance between the post-data weights and the priors (Gokhale and Kullback, 1978). Under this criterion, known as cross entropy (CE), the problem of recovering $\boldsymbol{p}$, may be formulated as minimizing the CE subject to the relevant structural constraints (moments). The resulting solution is

$$\tilde{p}(\theta_i) \propto q_i \exp\left\{ \sum_t \lambda_t Y_t(\theta_i) \right\}. \tag{3.2}$$

When the prior information q_i has uniform mass, the optimal solutions of the ME and CE problems are identical.

To relate the ME formalism to the more familiar linear model (2.1), consider a special case of this model where B of Assumption A1 is $B = \{\beta \in \Re^1 | \beta_k \in [0, 1], \sum_k \beta_k = 1\}$ and $\boldsymbol{e} = 0$. Thus,

$$\boldsymbol{y} = X\boldsymbol{\beta} \equiv X\boldsymbol{p} \tag{3.3}$$

and $\boldsymbol{p}$ is a K-dimensional proper probability distribution. The ME formulation is

$$ME = \begin{cases} \hat{\boldsymbol{p}} = \arg\max\left\{ -\sum_k p_k \log p_k \right\} \\ \text{s.t.} \quad \boldsymbol{y} = X\boldsymbol{p} \text{ and} \sum_k p_\mathrm{k} = 1. \end{cases} \tag{3.4}$$

Similarly, the CE formulation is just

$$CE = \begin{cases} \tilde{\boldsymbol{p}} = \arg\min\left\{ \sum_k p_k \log\left(p_k / q_k\right) \right\} \\ \text{s.t.} \quad \boldsymbol{y} = X\boldsymbol{p} \text{ and} \sum_k p_\mathrm{k} = 1, \end{cases} \tag{3.5}$$

where $I(\boldsymbol{p}, \boldsymbol{q}) = \sum_k p_k \log(p_k / q_k)$ is the Kullback–Leibler information, or CE, measure.

The exact *CE* solution is

$$\tilde{p}_k = \frac{q_k \exp\left(\sum_{i=1}^{T} \tilde{\lambda}_i x_{ik}\right)}{\sum_k q_k \exp\left(\sum_{i=1}^{T} \tilde{\lambda}_i x_{ik}\right)} \equiv \frac{q_k \exp\left(\sum_{i=1}^{T} \tilde{\lambda}_i x_{ik}\right)}{\Omega} \equiv \frac{q_k \exp(\tilde{\eta}_k)}{\Omega(\tilde{\eta})} \tag{3.6}$$

for some natural K-dimensional vector η. The dual *CE* counterpart is

$$\operatorname*{Inf}_{\boldsymbol{p} \in \mathbf{P}} I(\boldsymbol{p}, \boldsymbol{q}) = \operatorname*{Sup}_{\lambda \in D} \{\lambda' \boldsymbol{y} - \log \Omega(X'\lambda)\} = \operatorname*{Sup}_{\lambda \in D} \{\lambda' \boldsymbol{y} - F(\eta)\} \tag{3.7}$$

for $F(\eta) = \log \Omega(\eta)$ and where $P = \{\boldsymbol{p}: X\boldsymbol{p} = \boldsymbol{y}\}$ is a set of proper (normalized) distributions satisfying the linear constraints (3.3), and D is the set $\{\lambda \in \Re^{\mathrm{T}}: \Omega(X'\lambda) \ll \infty\}$. Having solved for $\tilde{\lambda}$, one gets

$$\tilde{\boldsymbol{p}} = \frac{\partial F(\tilde{\eta})}{\partial \tilde{\eta}}. \tag{3.8}$$

3.2. The linear GCE model

Building on the above, we go back to model (2.1) and Assumption A1. Let $\boldsymbol{z}_k$ of Example 1 be an M-dimensional vector $\boldsymbol{z}_k \equiv (\underline{z}_k, \bar{z}_k) = (z_{k1}, \ldots, z_{kM})'$ for all k. Instead of searching directly for the point estimates β, we view β as the expected value over some reference set B. To do so, let $\boldsymbol{p}_k$ be an M-dimensional proper probability distribution defined on the set $\boldsymbol{z}_k$ such that

$$\beta_k = \sum_m p_{km} z_{km} \equiv E_{p_k}[\boldsymbol{z}_k] \text{ or } \beta = E_{\boldsymbol{P}}[\boldsymbol{z}]. \tag{3.9}$$

In this way the observed data $\boldsymbol{y}$ are viewed as the mean process $\boldsymbol{z}$ with a probability distribution P that is defined on B(or $\boldsymbol{z}_k$'s). Before proceeding, it helps to assume

Assumption A2. $\boldsymbol{e} \in V$ where V is a symmetric around zero convex set.

Example 2. $V = \{\boldsymbol{e} \in \Re^{\mathrm{T}} | e_t \in (\underline{v}, \bar{v}), \quad t = 1, 2, \ldots, T\}$.

Thus, similar to the β's above, each error term is redefined as

$$e_t = \sum_j w_{tj} v_j \equiv E_{w_t}[\boldsymbol{v}] \text{ or } \boldsymbol{e} = E_W[\boldsymbol{v}]. \tag{3.10}$$

Having reparameterized β and $\boldsymbol{e}$, the linear model can be specified as $y_t = \sum_{k=1}^{K}\sum_{m=1}^{M} z_{km}p_{km}x_{tk} + \sum_j v_j w_{tj}$, or $\boldsymbol{y} = XE_{\boldsymbol{P}}[\boldsymbol{z}] + E_W[\boldsymbol{v}]$, and the *GCE* rule is

$$GCE = \begin{cases} \tilde{\boldsymbol{p}} = \underset{\mathbf{p},w}{\arg\min} \sum\limits_k\sum\limits_m p_{km}\log(p_{km}/q_{km}) + \sum\limits_t\sum\limits_j w_{tj}\log(w_{tj}/u_{tj}) \\ \text{s.t.}\quad \boldsymbol{y} = XE_{\boldsymbol{P}}[\boldsymbol{z}] + E_W[\boldsymbol{v}];\quad \sum\limits_m p_{\text{km}} = 1;\quad \sum\limits_j w_{tj} = 1. \end{cases} \tag{3.11}$$

The solution is

$$\tilde{p}_{km} = \frac{q_{km}\exp\left(\sum_t \tilde{\lambda}_t z_{km}x_{tk}\right)}{\sum_m q_{km}\exp\left(\sum_t \tilde{\lambda}_t z_{km}x_{tk}\right)} \equiv \frac{q_{km}\exp\left(\sum_t \tilde{\lambda} z_{km}x_{tk}\right)}{\Omega_k(\tilde{\lambda})} \tag{3.12}$$

and

$$\tilde{w}_{tj} = \frac{u_{tj}\exp(\tilde{\lambda}_t v_j)}{\sum_j u_{tj}\exp(\tilde{\lambda}_t v_j)} \equiv \frac{u_{tj}\exp(\tilde{\lambda}_t v_j)}{\Psi_t(\tilde{\lambda})}, \tag{3.13}$$

where the prior weights for $\boldsymbol{p}_k$ are $\boldsymbol{q}_k = (q_{k1}, \ldots, q_{kM})'$ and the corresponding prior weights for $\boldsymbol{w}$, consistent with the set of discrete points $\boldsymbol{v}$, are $\boldsymbol{u}_t = (u_{t1}, \ldots, u_{tJ})'$. With these estimates we proceed to calculate the point estimates $\tilde{\beta}_k \equiv \sum_m z_{km}\tilde{p}_{km}$ and $\tilde{e}_t \equiv \sum_j v_j \tilde{w}_{tj}$.

Finally, the dual *GCE* is

$$\underset{p\in P,\ w\in W}{\operatorname{Inf}} I(\boldsymbol{p}, \boldsymbol{w}; \boldsymbol{q}, \boldsymbol{u}) = \underset{\lambda\in D}{\operatorname{Sup}}\left\{\lambda' y - \sum_k \log\Omega_k(X'\lambda) - \sum_t \log\Psi_t(\lambda)\right\}. \tag{3.14}$$

Solving the dual yields the optimal λ, which in turn yields the point estimates via (3.12) and (3.13). Finally, the Hessian matrix of the GCE problem is positive definite for $\boldsymbol{p}, \boldsymbol{w} \gg 0$, and thus satisfies the sufficient condition for a unique global minimum.[3]

3.3. Discussion

In terms of the specification of $\boldsymbol{z}_k$ and $\boldsymbol{v}$, it is traditional to assume that the elements of β and $\boldsymbol{e}$ are finite, and that β_k and e_t are drawn from a finite sample and are usually bounded (see Assumptions A1 and A2). Furthermore, most sciences have a conceptual base for identifying and defining a relevant set of

[3] Since for uniform prior distributions the GCE solution reduces to the GME solution, only the GCE method is presented here. But unlike traditional Bayes estimators, we always specify the support spaces that bound the possible parameter space.

variables, along with the characteristics of the corresponding parameters consistent with a particular problem or data generation process. For example, equality and inequality constraints on β_k commonly arise in many econometric and other scientific applications, and theory or prior empirical results permit the parameters to be signed or specified within a bounded range $[a, b]$. In economics, behavioral and technical coefficients such as marginal productivity, marginal propensity to consume or price elasticities may be classified as non-negative or positive and naturally bounded. In terms of $\boldsymbol{v}$, one possibility is to use the sample (empirical) variance of $\boldsymbol{y}$ and the three-sigma rule. This is the approach taken here.

Next, one may ask how sensitive are the estimates to the specification of z_k. The simple answer is that as long as the center of these supports remain unchanged, say zero, the estimates are practically insensitive to changes in the boundary points of the supports. For example, let $z_k = (-C, 0, C)'$ for each k. Then, a symmetric change of these supports to $z_k = (-100C, 0, 100C)'$ has practically no significant effect on the estimated β's. Since in this work the objective is to identify the extraneous and the non-extraneous variables, the choice of zero as the center point of the supports, for each k, seems to be natural. However, it is emphasized that for small data sets, the estimates may be sensitive to a change in the center of the supports z.

3.4. Comparison with other estimation rules

First, we note that, unlike the usual Stein-like estimators (e.g., Stein, 1955, 1981; James and Stein, 1960; Judge and Bock, 1978; Brown, 1966; Bock, 1988), the GME–GCE estimators shrink estimates on a coordinate-by-coordinate basis. For example, if the support vector z_k is centered at zero, the coefficients close to zero receive maximum shrinkage.[4]

Second, we note that the GCE formulation, that leads to post-data means, has many of the characteristics of the standard Bayesian conjugate analysis where the posterior mean is a matrix-weighted combination of the prior means and the OLS estimates. In the GCE approach however, the posterior means are weighted combinations of the data and the priors within the supports. Consequently, the posterior means are always within the bounds of the supports Z and $\boldsymbol{v}$.[5]

[4] For a detailed comparison and discussion of other entropy and non-entropy regularization methods, as well as the maximum entropy on the mean, see for example Donoho et al. (1992), Golan et al. (1996, Chapter 8) and Bercher et al. (1996).

[5] For other non-Bayesian methods see for example the recent work on the empirical likelihood (e.g., Owen, 1990; Qin and Lawless, 1994), weighted least squares and the GMM (e.g., Hellerstein and Imbens, 1999; Imbens et al., 1998).

Finally, it is important to note the relationship between the GCE–GME entropy-based estimation rules and Zellner's seminal BMOM approach (e.g., Zellner, 1996a, 1997; Tobias and Zellner, forthcoming). Like the GCE method, the objective behind the BMOM method is to estimate the unknown parameters with minimum assumptions on the likelihood function. As stated by Zellner (1997, p. 86), "The BMOM approach is particularly useful when there is difficulty in formulating an appropriate likelihood function. Without a likelihood function, it is not possible to pursue traditional likelihood and Bayesian approaches to estimation and testing. Using a few simple assumptions, the BMOM approach permits calculation of post-data means, variances and other moments of parameters and future observations".

In the BMOM approach one starts by maximizing the continuous entropy function (of a density function) subject to some T side conditions (moments) and normalization. This yields the average log-height of the density function, which is the least informative density given these side conditions. Thus, under the BMOM approach one works with the most uninformed, that is maxent, post-data density for the parameters. A further advantage of the BMOM approach is that many types of side conditions can be (and have been) utilized in order to obtain post-data densities for parameters and future observations. These side conditions include bounds on the parameters' values, bounds on the error terms' ranges, inequality restrictions, fractile conditions, and moment conditions. For more innovative applications of the BMOM see, for example, LaFrance (1999) and the discussion in Zellner (1999). For a full comparison of the traditional Bayes and the BMOM approaches see Zellner (1997, Table 1) and for an information theoretic derivation of Bayes' Theorem that provides a link between maximum entropy procedures and Bayes' Theorem see Zellner (1988) and Soofi (1996).

Even though the basic objectives of the BMOM and the GCE are similar, they differ in their inputs. Following Zellner, the basic inputs in the BMOM are the (i) data, (ii) prior information, (iii) mathematical form of the model, (iv) moments of the parameters and future values, and (v) the Maxent principle. The two basic differences in the GCE inputs are in the way prior information is incorporated and the assumption (input) on the moments of the parameters and future values. Specifically, under the GCE rule, the prior information is incorporated via three routes: the support spaces ($\boldsymbol{z}$ and $\boldsymbol{v}$), the priors in the objective functional ($\boldsymbol{q}$ and $\boldsymbol{u}$) and other restrictions that may enter as additional equality/inequality constraints. Further, the GCE rule is specified such that each data point enters directly in the optimization (rather than a quadratic form of the data or the moments' input). Therefore, the moment requirement can be thought of as a 'weak' moments' requirement in the sense that the sample's moments may be different (up to a certain distance which is data dependent) from the underlying population's moments. Obviously, our choice of an estimator is problem and data specific and strongly depends on the amount of information we have with respect to a specific problem, model and data.

4. A flexible, prior, data-based estimator

In this section the GCE estimator is extended so that it embraces both the model identification and estimation objectives discussed earlier. To accomplish this task the GCE estimator is reformulated such that the extraneous variables are identified and the β parameter space may be truncated, or all variables are retained but the coefficients of extraneous variables are shrunk toward zero.[6]

Given the objective of identifying possible extraneous variables, we specify the discrete support space $\boldsymbol{z}_k$ for each β_k to be symmetric around zero and employ the GCE estimator with a unit mass prior on zero. While effective for identifying extraneous variables, as an estimator of β under a squared error loss measure, this does not lead to a minimax rule. If, on the other hand, we wish to have estimators that, under a squared error loss measure, are minimax and thus superior to traditional estimators over all, or part, of the parameter space, the GCE estimator that uses uniform priors (or similarly, the GME) is a good rule. However, if our objective is to shrink but not necessarily eliminate the extraneous variables, and simultaneously produce an estimator that has a good sampling performance over the whole range of the possible parameter space, then we may combine the GME and GCE (or the GCE with uniform priors and the GCE with spike priors) estimators. This is the formulation proposed here.

We start by specifying each $\boldsymbol{z}_k$ and $\boldsymbol{v}$ to be *symmetric around zero*, with large lower and upper bounds for $\boldsymbol{z}$ and the three-empirical-sigma-rule for $\boldsymbol{v}$, so that β_k and e_t are contained in a fixed interval with arbitrarily high probability. We also specify, as a possible alternative for each β_k, a 'spike' prior, with a point mass at $z_{km} = 0$, for each $k = 1, 2, \ldots, K$. Thus, a flexible, data-based prior is specified such that for each β_k coordinate either a spike prior at the $z_{km} = 0$, a uniform prior over the discrete support space $\boldsymbol{z}_k$, or any convex combination of the two can result. The goal is to produce a natural adaptive statistical method that is data based and free of subjective choices except for the bounds on the support spaces. Because we are interested in a formulation that puts pressure on the data by including both uniform and spike prior alternatives, or some convex combination of the two, we are not able to make use of the conventional cross entropy formalism. Consequently, given the problem at hand, and within the entropy approach, we consider an extended entropy formulation. But because the $\boldsymbol{q}_k$ prior alternatives cannot be introduced as structural constraints, we must find a way to introduce them in the objective function. To accomplish this,

[6] For other approaches with similar objective see for example the Bayes–Stein estimators for the normal k-means reviewed and developed in Zellner and Vandaele (1975).

consider the following variant of the GCE formulation of Section 3:

$$\underset{p,p^{\gamma},w}{\text{Min}}\ I(\boldsymbol{p},\boldsymbol{p}^{\gamma},\boldsymbol{w}) \equiv \sum_{k}(1-\gamma_k)\left(\sum_{m} p_{km}\log p_{km}\right)/[(1-\gamma_k)(-\log M)+\gamma_k\xi]$$

$$+\sum_{k}\gamma_k\sum_{m} p_{km}\log(p_{km}/q_{km})+\sum_{k,h} p^{\gamma}_{kh}\log(p^{\gamma}_{kh}/q^{\gamma}_{kh})$$

$$+\sum_{xt,j} w_{tj}\log(w_{tj}/u_{tj}) \tag{4.1}$$

s.t.

$$y_t=\sum_{k,m} z_{km}p_{km}x_{tk}+\sum_{j} v_j w_{tj} \tag{4.2}$$

and

$$\sum_{m} p_{km}=1,\sum_{h} p^{\gamma}_{kh}=1,\sum_{j} w_{tj}=1 \tag{4.3}$$

and where the prior weight (prior mixture) γ_k is

$$\gamma_k \equiv \sum_{h} z^{\gamma}_{h}p^{\gamma}_{kh} \tag{4.4}$$

with $z^{\gamma}_{1}=0$ and $z^{\gamma}_{H}=1$ always and where $\xi \equiv -\sum_m q_{km}\log q_{km}$. Further, *except for the point mass prior* $\boldsymbol{q}_k$, *all other priors* (*i.e.*, $\boldsymbol{q}_\gamma$ *and* $\boldsymbol{u}_t$) *are specified to be uniform.*

4.1. The criterion

Having specified the new data-weighted prior (DWP) estimation rule, we now discuss the explicit structure of the objective function (4.1). There are four parts to this objective function. Just as in Section 3 (the upper part of Eq. (3.11)), the last element on the right-hand side (RHS) of (4.1) is the entropy of the noise component. Since $\boldsymbol{v}$ is specified to be symmetric, and equally spaced, support around zero, this part of the objective function shrinks the noise components toward zero. The bounds for the errors' supports are just $\pm 3\bar{\sigma}_y$ where $\bar{\sigma}_y$ is the empirical standard deviation of the sample. The first two elements on the RHS of (4.1) relate to the uniform and informative (spike) priors respectively, with corresponding weight $\gamma_k \in [0,1]$, where $\gamma_k=0$ implies use of a uniform (uninformative) prior, while $\gamma_k=1$ implies use of a spike prior. In terms of the first RHS element, as noted in Section 3, when using a uniform prior over the support points there is no need to specify this prior explicitly. The third element in the objective function relates to the total entropy of the weights γ. Finally, the first element of the objective function is scaled by the convex combination of the negative of the entropy of the priors. If, for example, $\gamma_k = 0$, then it is scaled by the negative entropy of the uniform distribution (for coordinate k), which is $-\ln(M)$. If, on the other hand, $\gamma_k = 1$, then it is scaled by $\xi \equiv -\sum_m q_{km} \log q_{km}$, which equals zero iff $\boldsymbol{q}_k$ is a spike prior. This scaling, or normalization, is needed because without it, the first two elements are of different magnitude and sign. Thus, without this normalization, the uniform prior *always* takes over, and the DWP reduces to the GME. This is because the first element of the objective function is always non-positive, while the second element is always non-negative. With this scaling, however, the two parts of the objective can 'communicate' in the sense that both are of the same sign and magnitude, and there is a data-based choice of the prior for each coordinate that is most consistent with the sample information. In this way this estimator simultaneously chooses the mixture of the two alternative priors on a coordinate-by-coordinate basis and uses this information, along with the data, to determine the shrinkage and to provide estimates of the unknown parameters. We note in conclusion that if one wishes to avoid the scaling carried in (4.1), the first component on the RHS of (4.1) can be substituted for the GCE with uniform priors. However, the formulation used here, is computationally superior and more efficient.

4.2. Solution and properties of the estimator

The solution to the optimization problem (4.1)–(4.3) yields

$$\tilde{p}_{km} = \frac{q_{km}^{\tilde{\gamma}_k/A_k} \exp\left((1/A_k)\sum_t \tilde{\lambda}_t x_{tk} z_{km}\right)}{\sum_m q_{km}^{\tilde{\gamma}_k/A_k} \exp\left((1/A_k)\sum_t \tilde{\lambda}_t x_{tk} z_{km}\right)}, \tag{4.5}$$

$$\tilde{w}_{tj} = \frac{u_{tj} \exp(\tilde{\lambda}_t v_j)}{\sum_j u_{tj} \exp(\tilde{\lambda}_t v_j)} \equiv \frac{u_{tj} \exp(\tilde{\lambda}_t v_j)}{\Psi_t(\tilde{\lambda})} \tag{4.6}$$

and

$$\tilde{\gamma}_k \equiv \sum_h \tilde{p}_{kh} z_h^{\gamma}, \tag{4.7}$$

where λ reflects the T-Lagrange multipliers for the data equations (4.2) and $A_k \equiv 1 - \tilde{\gamma}_k/((\tilde{\gamma}_k - 1)\log M + \tilde{\gamma}_k \xi) + \tilde{\gamma}_k$. As before (3.9) and (3.10) provide the basis for recovering estimates of β and $\boldsymbol{e}$. As $\gamma_k \to 0$ the designated prior becomes more uniform, with the estimates approaching those of the GME estimator. For large values of γ_k (above 0.49), the GCE estimator with an informative (in our case, point mass on zero) prior takes over.

The following conditions ensure consistency and asymptotic normality of the DWP estimation and variable selection rule:

(i) The errors' support $\boldsymbol{v}$ is symmetric around zero (see Assumption A2).
(ii) The support space $\boldsymbol{z}_k$ spans the true value of each one of the unknown parameters β. Further, the support has a finite lower and upper bounds z_{k1} and z_{kM}, respectively (see Assumption A1).
(iii) The errors are iid.
(iv) $\lim_{T\to\infty} T^{-1}X'X$ exists and is non-singular.

The proof of consistency and asymptotic normality follows directly from the empirical likelihood approach (Owen, 1990, 1991; Qin and Lawless, 1994; Golan and Judge, 1996). Similarly, these proofs can be established by following Mittelhammer and Cardell (1996).

In general, under the above four conditions, $\sqrt{T}(\hat{\beta}-\beta) \overset{\mathrm{d}}{\rightarrow} \mathrm{N}(0, Q_{\mathrm{DWP}})$ where $Q_{\mathrm{DWP}} = \sigma^2[\lim T^{-1}(X'X)^{-1}]$ is the asymptotic covariance matrix for the DWP. Since β is a continuous function of λ (the Lagrange multipliers) this statement is an immediate extension of Qin and Lawless (1994, Lemma 1 and Theorem 1). Finally, $\hat{\sigma}^2 = [1/(T-K)]\sum_t \hat{e}_t^2$ with $\hat{e}_t = \sum_j v_j \hat{w}_{tj}$ is a consistent estimator of the variance.

4.3. Diagnostics and inference

Given this testing basis, we propose a test to compare the restricted DWP model with the unrestricted one. With identical, and symmetric around zero, supports $\boldsymbol{z}$ and $\boldsymbol{v}$ for both problems, we follow Owen (1990, 1991), Qin and Lawless (1994), and Golan and Judge (1996) and define an empirical entropy (or expected log-likelihood) ratio statistic

$$\ell_E \equiv L_{E(\beta_0)} - L_{E(\tilde{\beta})}, \tag{4.8}$$

where $L_{E(\beta_0)}$ applies to the optimal value of the DWP estimator's objective function when restricting $\beta = \beta_0 = \mathbf{0}$, while $L_{E(\tilde{\beta})}$ is the optimal value of the DWP estimator's objective function when the β's are not restricted (i.e., model 4.1–4.3). Under the null, $2\ell_E$ is asymptotically distributed as a $\chi^2_{(K)}$. Similarly, one can test the hypothesis whether each one of the K covariates is extraneous. In that case, the entropy-ratio statistic $2\ell_E$ has a limiting distribution of $\chi^2_{(1)}$.

We now use the DWP rule to provide a normalized (information) entropy measure, which will be used as the criterion for identifying and eliminating the extraneous variables from the design matrix. Let the normalized entropy for each coordinate k be

$$S(\tilde{\boldsymbol{p}}_k) \equiv \frac{-\sum_{m=1}^{M} \tilde{p}_{km} \log \tilde{p}_{km}}{\log M}, \tag{4.9}$$

where $S(\tilde{p}_k) = 1$ implies maximum uncertainty and $S(\tilde{p}_k) = 0$ implies full certainty.

Next, we can relate the normalized entropy, or information measure, S, with the χ^2 statistic. We start by pointing out that each component of (4.1) is a basic cross-entropy measure for some proper set of probabilities. Ignoring the γ's for simplicity sake, let $\tilde{p}_{km}$ be any appropriate estimator of p_{km} and let $\{\tilde{\boldsymbol{p}}_k\}$ be a set of M probabilities over the M-dimensional support space $\boldsymbol{z}_k$ for each one of the K coordinates. Then, we define the statistics

$$I(\tilde{\boldsymbol{p}};\boldsymbol{q}) \equiv \sum_{k=1}^{K} \sum_{m=1}^{M} \tilde{p}_{km} \log(\tilde{p}_{km}/q_{km}) \tag{4.10}$$

and

$$\chi^2_{(M-1)} \equiv M \sum_{m=1}^{M} \frac{1}{q_{km}} (\tilde{p}_{km} - q_{km})^2 \quad \text{for each } k = 1, 2, \ldots, K. \tag{4.11}$$

A second-order approximation of $I(\tilde{\boldsymbol{p}};\boldsymbol{q})$ is just the entropy-ratio statistic for evaluating $\tilde{\boldsymbol{p}}$ vs. $\boldsymbol{q}$ discussed above. That is, for each k,

$$I(\tilde{\boldsymbol{p}}_k;\boldsymbol{q}_k) \cong \frac{1}{2} \sum_{m=1}^{M} \frac{1}{q_{km}} (\tilde{p}_{km} - q_{km})^2 \quad \text{for each } k = 1, 2, \ldots, K, \tag{4.12}$$

so $2M$ times the entropy-ratio statistic corresponds to $\chi^2_{(M-1)}$. Given this relationship, we use the normalized entropy measure, for each covariate k, in order to test the hypothesis $\mathrm{H}_0\colon \beta_k = 0$ for each k. Thus, $\chi^2_{(M-1)} = 2MI(\tilde{\boldsymbol{p}}_k;\boldsymbol{q}_k)$ for the spike priors $\boldsymbol{q}_k$.

Based on the above we can define some rules to identify the extraneous variables. First, the following variable selection rule is defined. If $[1 - S(\tilde{\boldsymbol{p}}_k)] \geqslant 0.99$, then variable k is classified as extraneous. Conversely, if $[1 - S(\tilde{\boldsymbol{p}}_k)] < 0.99$, then a real classification of the variable is suggested. An $S(\tilde{\boldsymbol{p}}_k) = 0.99$ as opposed to 1.0 is used to allow for insignificant computer rounding effects.

Alternatively, a second identification criterion is proposed. This criterion is based on the weight $\hat{\gamma}_k$ in (4.1). If for a particular variable $\tilde{\gamma}_k < 0.5$, then a non-extraneous variable is identified. As $\tilde{\gamma}_k$ decreases, the strength of this identification is increased. Conversely, a $\tilde{\gamma}_k \geqslant 0.5$ suggests the variable belongs in the extraneous category.

Finally, we relate our choice of $1 - S(\tilde{\boldsymbol{p}}_k) \geqslant 0.99$ to the χ^2 measure which, in that case, yields

$$\chi^2_{(M-1)} = 2\,MI(\tilde{\boldsymbol{p}}_k;\boldsymbol{q}^u_k) = 2M\ln(M)[1 - S(\tilde{\boldsymbol{p}}_k)], \tag{4.13}$$

where $\boldsymbol{q}^u_k = 1/M$ for all $m = 1, 2, \ldots, M$. Thus, for $M = 5$ (the number of support points used in all the experiments presented in the following sections), the 0.99 rule implies $\chi^2 = 15.9 > (\chi^2_{(4)})^c = 13.3$ for $\alpha = 0.01$.

4.4. Comparison with the ML approach

Having specified an estimator with a variable selection component and demonstrated its asymptotic properties, it is possible to compare its sampling performance with other conventional estimation rules and with other variable selection criteria. To achieve this comparison, instead of working with each data point directly, it is possible to follow common practice and transform the data representation (4.2) to a moment representation

$$X'\boldsymbol{y} = X'X\beta + X'\boldsymbol{e} = X'XZ\boldsymbol{p} + X'V\boldsymbol{w}, \tag{4.2a}$$

where all previous definitions of β and $\boldsymbol{e}$ follow and V is a matrix composed of the T vectors $\boldsymbol{v}$.

Lemma 1. If we let $\boldsymbol{v} = 0$ and substitute the K pure moment conditions $X'\boldsymbol{y} = X'XZ\boldsymbol{p}$ for the T data-consistency equations (4.2) then, under restriction (ii), the resulting estimates of problem (4.1)–(4.3) are equivalent to the least-squares (ML) estimates.

This proof is immediate since the constraints $X'\boldsymbol{y} = X'XZ\boldsymbol{p}$, that must be satisfied, are just the first-order conditions for the least-squares (ML) estimates.

Lemma 2. For any finite data set, the *approx var*$(\tilde{\beta}_{k(DWP)}) \leqslant$ *approx var*$(\tilde{\beta}_{k(GCE)})$ for all k.

The logic for the proof is as follows. Golan et al. (1996, Chapter 7) show that for all $\boldsymbol{v} \neq \boldsymbol{0}$, the *approx var*$(\tilde{\beta}_{k(GCE)}) <$ *approx var*$(\tilde{\beta}_{k(LS/ML)})$ for all finite data sets and for all k. Following their logic, for expository purposes, let X be an orthonormal matrix and the error covariance $\sum_e = \sigma^2 I_T$. Under these assumptions the approximate DWP covariance is

$$Cov\,\tilde{\beta}_{(DWP)} = \tilde{\sigma}^2 {\textstyle\sum_z} \left({\textstyle\sum_z} + {\textstyle\sum_v}\right)^{-2} {\textstyle\sum_z} \cong \tilde{\sigma}^2 (X'X)^{-1}, \tag{4.14}$$

where $\tilde{\Sigma}_z$ and $\tilde{\Sigma}_v$ are the respective covariances for $\tilde{\boldsymbol{p}}$ and $\tilde{\boldsymbol{w}}$. The kth element of (4.14) is

$$Var(\tilde{\beta}_{k(DWP)}) = \tilde{\sigma}^2 \left(\frac{\tilde{\sigma}_z^2}{\tilde{\sigma}_z^2 + \tilde{\sigma}_v^2} \right)^2, \tag{4.15}$$

where at the limit, $\tilde{\sigma}_v^2 = 0$, and (4.14) is just the LS(ML) variance $\hat{\sigma}^2$. To compare the DWP and GCE variances for the kth element, we need to evaluate $\tilde{\sigma}_z^2$, which is just

$$\tilde{\sigma}_z^2 = \sum_m \tilde{p}_{km} z_{km} - \left(\sum_m \tilde{p}_{km} z_{km} \right)^2. \tag{4.16}$$

Variance (4.16) reaches its maximum at a given z_k for $\tilde{p}_{km} = 1/M$, for all $m = 1, 2, \ldots, M$. In Section 3, given the data constraints, the GCE objective (with uniform priors) makes the estimates $\hat{p}_{km}$ as uniform as possible. Alternatively, the spike priors in the DWP estimator make the $\tilde{p}_{km}$ the least possible uniform given the data. Consequently, the less uniform the $\tilde{p}_{km}$, the smaller $\tilde{\sigma}_z^2$ in (4.16). This establishes the relationship *approx* $\mathrm{var}(\tilde{\beta}_{k(DWP)}) \leqslant$ *approx var*$(\hat{\beta}_{k(GCE)})$.

To summarize, most variable selection models (criteria) require choosing among 2^k different models while imposing some unknown smoothing function (Zheng and Loh, 1995) or smoothing parameter (e.g., C_p and AIC). Alternatively, the DWP is a data-driven penalty function estimator that is based on weak distributional assumptions. It does not require 2^k steps and/or a pre-specified smoothing parameter. It is a one-step estimation rule that requires the pre-specified support spaces. Finally, note that both the popular AIC variable selection rule and DWP are based on the Kullback–Leibler information (cross entropy) measure. If we make use of the pure moment condition (4.2a) within the DWP problem (4.1)–(4.3), then it is easy to show a proportional relationship between the DWP and an AIC criterion for each of the 2^k models.[7]

5. Sampling experiments

In this section we report some results of Monte Carlo sampling experiments to indicate the small sample performance of the DWP estimator and, under a squared error loss measure, compare it with other relevant traditional and shrinkage estimators. For the estimator comparisons, we continue to consider the multivariate estimation problem where the ML estimator is $\delta^0(\boldsymbol{b})$ and has risk $\rho(\beta, \delta^{\mathbf{b}}) = \sigma^2 \, tr\,(X'X)^{-1}$. We note here that in this work we restrict ourselves to the squared error loss measure and do not attempt to analyze predictive performance.[8] For a comprehensive comparison of the predictive powers of GME and some Bayesian methods (MELO and BMOM) see the recent work of Perloff and Shen (1999).

5.1. The symmetric case

In order to provide an experiment that involves a mean square prediction error loss and a comparison with the well-known positive rule Stein shrinkage

[7] For a nice discussion of the AIC criterion and its relationship to the CE criterion and other information criteria, within the model selection, see Lavergne (1998).

[8] In future work we will investigate the predicitive performance of these estimation rules where the parameters of each method are restricted to fall within the same bounds.

estimator, we use the orthonormal K-mean linear statistical model

$$\boldsymbol{y} = X\beta + \boldsymbol{e} = XS^{-1/2}S^{1/2}\beta + \boldsymbol{e} = A\,\theta + \boldsymbol{e}, \tag{5.1}$$

where $S^{1/2}$ is a positive-definite symmetric matrix with $S^{1/2}S^{1/2} = S = X'X$, $A'A = I_K$ and $\theta = S^{1/2}\beta$. In the experiment $\boldsymbol{e} \sim \mathrm{N}(0, \sigma^2 I_T)$, $\sigma = 1$ and $K = 4$. The risk

$$E[(\tilde{\theta} - \theta)'(\tilde{\theta} - \theta)] = E[(\tilde{\beta} - \beta)'X'X(\tilde{\beta} - \beta)] \tag{5.2}$$

yields a weighted loss function in terms of β and results in the *mean square prediction error* criterion that is often used in econometrics to evaluate performance. The parameter space investigated is $\theta = c\boldsymbol{d}_i$, where $\boldsymbol{d}_1' = (\theta_1, 0, 0, 0)$, $\boldsymbol{d}_2' = (\theta_1, \theta_2, 0, 0)$ and $\boldsymbol{d}_3' = (\theta_1, \theta_2, \theta_3, 0)$. The scalar c is chosen so that the parameter vector length $(\theta'\theta)^{1/2} = 0, 1, 2, \ldots, 66$. For selected values of the $(\theta'\theta)^{1/2}$ parameter space, 5000 samples of size $T = 10, 30$ and 100 were generated and empirical estimator risks under a $\|\delta(\boldsymbol{y}) - \theta\|^2$ measure were obtained. For the DWP estimator, $\boldsymbol{z}'_k = (-100, -50, 0, 50, 100)$ and $\boldsymbol{v}' = (-3\bar{\sigma}_y, 0, \bar{\sigma}_y)$ *where* $\bar{\sigma}_y$ *is the empirical standard deviation of each sample*. Further, $\boldsymbol{q}'_k = (0, 0, 1, 0, 0)$ for each k, which means we are putting point mass at zero for each θ in the case of the GCE estimator, and $\boldsymbol{u}' = (0.33, 0.33, 0.33)$ for each t. For comparison purposes, the risk for the ML, positive rule Stein (PRS), GME and GCE estimators are reported. To make the ML approach fully comparable to the DWP, we need to use the constrained ML where the constraints specified (for each β) are the lower and upper bounds of $\boldsymbol{z}$. But because we use very wide bounds, the constrained solution is equivalent to the unconstrained solution and therefore we refer to it as ML.

The risk for the PRS over the $(\theta'\theta)^{1/2}$ parameter space was numerically evaluated using an algorithm by Bohrer and Yancey (1984). The $\boldsymbol{z}_k$ and $\boldsymbol{v}$ support spaces noted above are also used for the GME and GCE estimators. It is worth noting that the support spaces for $\boldsymbol{z}_k$ where chosen to reflect very wide bounds. Increasing these bounds did not change the estimates.

5.1.1. Experiment 1 – $\boldsymbol{d}_1' = (\theta_1, 0, 0, 0)$

5.1.1.1. Variable identification. First, we focus on the variable selection objective with a design matrix involving one non-extraneous and three extraneous variables (Table 1). The last three columns of Table 1 provide information that forms a basis for identifying the correct design matrix for different points in the $(\theta'\theta)^{1/2}$ parameter space. The column labeled $\tilde{\gamma}$ identifies the weight between the uniform and the point mass prior for each coordinate at each point in the $(\theta'\theta)^{1/2}$ parameter space. Note at the origin the weight is 0.5, which signifies that the informative prior with point mass at zero is the active prior. This choice remains active over the whole $(\theta'\theta)^{1/2}$ parameter space for the extraneous variables. For the non-extraneous variable, as $(\theta'\theta)^{1/2}$ increases, the weight on the point mass prior decreases and finally all weight is allocated to the uniform prior.

Using the normalized entropy measure, column 5 reports the probability of correctly identifying each of the variables. Note at the origin, where all coordinates are extraneous, and the point mass prior is consistent with the location parameters, the probability of correct identification is about 0.98. Over the parameter space this identification probability is maintained for the extraneous variables. For the *non-extraneous* variable x_1, over the range $(\theta'\theta)^{1/2} \in (1,2)$, the DWP rule underfits, and the probability of correct variable identification is on average less than 0.5. However, for $(\theta'\theta)^{1/2} \geqslant 3$, the probability of correct identification, on average, approaches 1.0. In terms of the normalized entropy measure, these results mirror those reported under the $\tilde{\gamma}$ column. For the non-extraneous variable, x_1, the normalized entropy measure, which reflects the probability of being an extraneous variable, decreases as $(\theta'\theta)^{1/2}$ increases, and thus reinforces the $\tilde{\gamma}_k$ measure in identifying this variable from its extraneous counterparts. One reason for these nice results is that when all but a single variable are extraneous, a model that identifies these extraneous variables correctly, must also identify the single non-extraneous variable correctly. The next set of experiments investigate the more complex case of more than a single non-extraneous variable.

5.1.1.2. Empirical risks. In this section, we assume that an investigator wishes to use the DWP formulation as an estimator that shrinks but does not eliminate. To show the statistical consequences of this rule, we present in Fig. 1 empirical risks over the $(\theta'\theta)^{1/2}$ parameter space. As a basis of comparison, we present the corresponding risks for the ML and PRS estimators. The comparison with the PRS points out the risk implications of Stein-like estimators where all coordinates are shrunk by the same amount versus estimators such as GCE, GME and DWP, where shrinkage may vary from coordinate to coordinate. Thus, in this experiment we contrast the performance of the estimators where extraneous variables exist and are shrunk but not identified and removed. Variability and bias results for this experiment (for the DWP) are presented in column 2 of Table 1.

Under this sample design the GCE estimator is risk superior to the PRS estimator over the whole range of the $(\theta'\theta)^{1/2}$ parameter space. At the origin the empirical risks of the PRS estimator and the GCE estimator, with spike prior at zero, are 1.472 and 0.53, respectively. From this point in the $(\theta'\theta)^{1/2}$ parameter space the risk of the positive part Stein increases sharply, reaches 2.8 at 10 and finally becomes equal to the ML risk of 4.0 around $(\theta'\theta)^{1/2} = 60$. Alternatively, the GCE empirical risk increases more slowly and finally reaches 3.37 as $(\theta'\theta)^{1/2}$ approaches 60. Consequently, in these sampling experiments the GCE estimator is risk superior, over the range of the θ parameter space, to both the ML and PRS estimators. If, instead of using a point mass prior at zero, we had used a uniform prior over the elements in $\boldsymbol{q}_k$, the empirical risk of the GME estimator is about 3.90 over the whole $(\theta'\theta)^{1/2}$ parameter space.

Table 1
Performance of the DWP estimator for selected points where $\theta = (\theta_1, 0, 0, 0)'$ and x_2, x_3, x_4 are extraneous and the ML $MSE(\delta^b) = 4$

$(\theta'\theta)^{1/2}$	MSE	$tr\{Cov(\tilde{\theta})\}$	$\tilde{\gamma}$				Frequency of correctly identifying				$1 - S(\tilde{p}_k)$			
			x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
0	0.246	0.246	0.50	0.5	0.5	0.5	0.98	0.97	0.98	0.98	0.999	0.999	0.999	0.999
1	1.159	0.50	0.49	0.5	0.5	0.5	0.12	0.97	0.98	0.97	0.99	0.999	0.999	0.999
2	2.576	1.30	0.47	0.5	0.5	0.5	0.47	0.97	0.97	0.97	0.97	0.999	0.999	0.999
3	2.617	1.62	0.43	0.5	0.5	0.5	0.84	0.97	0.97	0.97	0.94	0.999	0.999	0.999
4	2.182	1.46	0.40	0.5	0.5	0.5	0.98	0.96	0.96	0.97	0.91	0.999	0.999	0.999
5	1.891	1.29	0.38	0.5	0.5	0.5	1.00	0.97	0.96	0.97	0.89	0.999	0.998	0.999
6	1.802	1.30	0.35	0.5	0.5	0.5	1.00	0.96	0.96	0.97	0.87	0.999	0.999	0.999
10	1.548	1.28	0.26	0.5	0.5	0.5	1.00	0.96	0.96	0.97	0.81	0.998	0.998	0.999
20	1.276	1.22	0.14	0.5	0.5	0.5	1.00	0.96	0.95	0.96	0.69	0.998	0.998	0.998
60	1.202	1.20	0.01	0.5	0.5	0.5	1.00	0.96	0.96	0.97	0.58	0.998	0.998	0.999

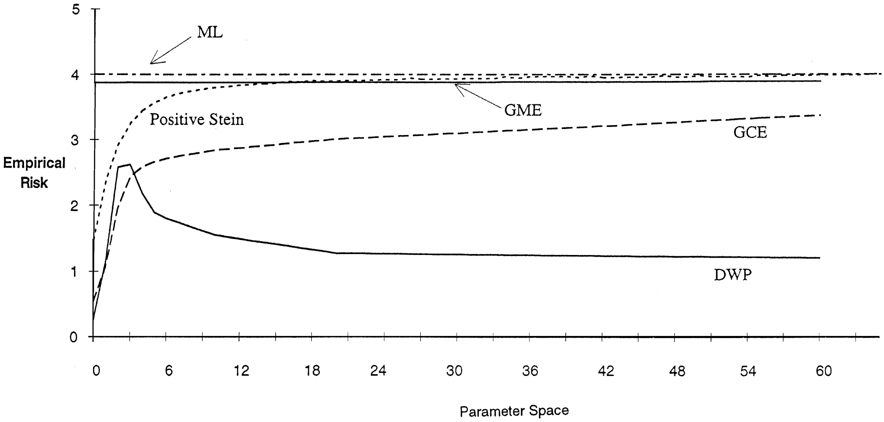

Fig. 1. Empirical risk functions for the ML, GME, GCE, PRS and DWP estimators.

Alternatively, the empirical risk of the DWP estimator is 0.246 at the origin where all variable coordinates are correctly shrunk toward zero, and then reaches a maximum of 2.61 at $(\theta'\theta)^{1/2} = 3$, where θ_1 continues to be shrunk toward zero and the bias is a major part of the risk component. From there on the risk decreases sharply and stabilizes at about 1.20 for the remainder of the parameter space. In this range of the parameter space, where θ_1 is shrunk relatively little, the coefficients of the extraneous variables θ_2, θ_3 and θ_4 continue to have a maximum shrinkage toward zero. The DWP estimator identifies and shrinks the extraneous variables correctly and is superior to all the other estimators over the whole parameter space.[9]

Finally, it is important to remember that under the GME, GCE and DWP rules, specific bounds are imposed on all of the parameters' values. Naturally, these bounds may have a significant impact on the sampling results. Such bounds can of course be introduced in traditional Bayes and other sampling theory approaches. To provide a fair comparison, the bounds used in this paper covered a very large range of the parameter spaces such that the restricted and unrestricted ML yielded the same results. Nevertheless, one should keep in mind that the use of bounds can cause moments to exist and eliminate outlying estimates. For example, Perloff and Shen (1999) show that when estimating $\gamma = 1/\beta$ with β restricted to the range $0 < a < \beta < b$, this restriction precludes

[9] These experiments were repeated for different sample sizes but to save space are not reported here. For example, for $T = 30$ the MSE for the DWP is about 3.2 time smaller than that of the ML over the whole parameter space. For $T = 100$, the DWP maintains its superiority as well. In this case it is between 2.5 and 3.2 times smaller than the ML over the parameter space.

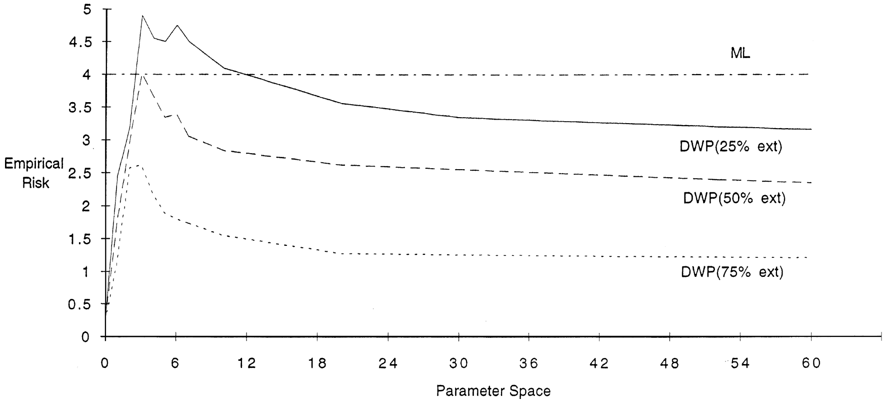

Fig. 2. Empirical risk functions for the DWP estimator for different levels of extraneous variables.

estimates of β to take on values that are close to zero, implying that the effect on the estimated values could be quite large. Consequently, the results presented here should be interpreted with these qualifications in mind. To summarize, there is a need for much more research in order to identify and establish the exact impact of imposing bounds on the properties of any estimator. As the method developed here, as well as the GME/GCE, uses pre-specified bounds on the parameters, the effects of these bounds on the estimator's properties is a subject for future studies.

5.1.2. Experiment 2 - $\boldsymbol{d}'_3 = (\theta_1, \theta_2, \theta_3, 0)$

The previous experiment pretty much tells the estimation and variable selection story because the DWP results evolve on a coordinate-by-coordinate basis. Thus, risk and variable selection results for any mix of non-extraneous and extraneous in the full design matrix can be approximated from the results of Section 5.1.1 .

To give a sense of the risk results for different ratios of non-extraneous to extraneous variables, we present the empirical risk functions for the DWP and ML rules in Fig. 2. Note that, over a small portion of the parameter space (for small values of θ_k), the DWP rule with 25% extraneous variables underfits and is inferior to the ML and Stein rules. However, over most of the $(\theta'\theta)^{1/2}$ parameter space, the DWP rule is risk superior. Finally, in terms of variable selection, the $\tilde{\gamma}_k$ weight parameter and the normalized entropy measure work as in Section 5.1.1.1. Both the non-extraneous and extraneous variables are identified, over the range of the θ parameter space, with probabilities $\geqslant 0.95$.

5.1.3. Experiment 3 – the χ^2 error distribution

To demonstrate the robustness of the DWP estimator in a non-Gaussian setting, we repeated the experimental design of Section 5.1 with $\chi^2_{(4)}$ random errors, normalized to have a unit variance. The results for this case (not reported here)[10] are similar in structure and magnitude to those presented in Fig. 1 and Table 1. In terms of identifying the extraneous and non-extraneous variables, the DWP mirrors the results reported in Table 1. In particular, at least 93% of the extraneous variables are identified over all points in the $(\theta'\theta)^{1/2}$ parameter space.

5.2. The non-symmetric case

5.2.1. Experiment 4 – high condition number

Consider for this experiment the general linear statistical model $\boldsymbol{y} = X\beta + \boldsymbol{e}$, where the ML estimator δ^b is distributed as $\mathrm{N}_K(\beta, \sigma^2(X'X)^{-1})$, and $X'X$ is a positive-definite matrix. For constructing the design matrix we use the condition number's definition $\kappa(X'X) = \pi_1/\pi_2$ which is the ratio of the largest and smallest singular values of the design matrix X, with columns scaled to unit length. As $\kappa(X'X)$ increases, traditional estimators, such as the ML estimator, become unstable and have low precision. For a review of different regularization methods that exhibit relatively good risk characteristics see, for example, Hoerl and Kennard (1970), O'Sullivan (1986), Breiman (1995), Tibshirani (1996), and Titterington (1985).

The sampling experiment for the non-symmetric case is similar to Experiment 1, reported in Section 5.1.1. However, to reflect a design matrix more in line with data that are by and large non-experimentally generated, instead of a condition number of $\kappa(X'X) = 1$ for the design matrix, we now specify a very moderate condition number of $\kappa(X'X) = 90$.[11] Under the SEL, the ML risk is approximately 47.5 and the iterative ridge estimator risk is approximately 14.1. The risk performance of the DWP and GCE rules mirrors that for the well-posed case, where $\kappa(X'X) = 1$. The DWP risk starts at about 0.21, increases to a maximum of 2.18 for $(\beta'\beta)^{1/2} = 2$, and decreases rapidly to 1.0 from about $(\beta'\beta)^{1/2} = 12$ in the parameter space. Unlike the DWP, the GCE risk increases monotonically with $(\beta'\beta)^{1/2}$ and is unbounded (or bounded by the support space Z). Selected points on the parameter space for the DWP estimator are reported in Table 2. Note in comparing the MSE for the well-posed design matrix in

[10] All results discussed in this Section are available, upon request, from the author.

[11] For example, the commonly used 'Longley' aggregated employment data (Longley, 1967) have a condition number of approximately 11,100; a highly collinear case. Our choice of 90 is completely arbitrary with the objective of maintaining a very moderate, but yet realistic, level of collinearity.

Table 2
Performance of the DWP estimator for $\beta = (\beta_1, 0, 0, 0)'$, $T = 10$ and $\kappa(X'X) = 90$

$(\beta'\beta)^{1/2}$	MSE	$tr\{Cov(\theta\}$	$\tilde{\gamma}$				Frequency of correctly identifying				$1 - S(\tilde{p}_k)$			
			x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
0	0.209	0.209	0.50	0.5	0.5	0.5	0.97	0.999	0.98	0.96	0.998	1.0	0.999	0.999
1	1.096	0.52	0.49	0.5	0.5	0.5	0.16	0.998	0.97	0.96	0.99	1.0	0.999	0.999
2	2.180	1.26	0.46	0.5	0.5	0.5	0.58	0.995	0.96	0.95	0.97	1.0	0.999	0.998
3	2.051	1.38	0.43	0.5	0.5	0.5	0.91	0.995	0.96	0.95	0.94	1.0	0.998	0.998
4	1.689	1.21	0.40	0.5	0.5	0.5	0.99	0.999	0.96	0.95	0.91	1.0	0.998	0.998
6	1.521	1.19	0.34	0.5	0.5	0.5	0.999	0.999	0.96	0.95	0.87	1.0	0.998	0.998
10	1.238	1.06	0.26	0.5	0.5	0.5	1.0	1.0	0.95	0.95	0.80	1.0	0.998	0.998
20	1.074	1.02	0.13	0.5	0.5	0.5	1.0	1.0	0.95	0.95	0.69	1.0	0.998	0.998
60	0.996	0.996	0.02	0.5	0.5	0.5	1.0	1.0	0.96	0.95	0.58	1.0	0.998	0.998

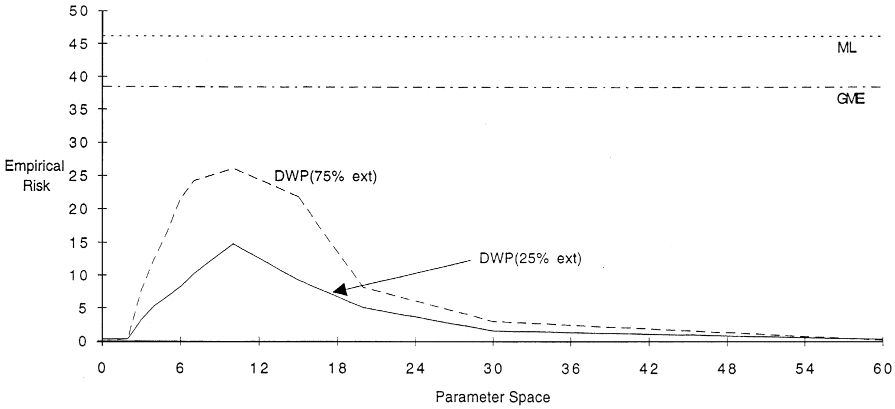

Fig. 3. Empirical risk functions for the DWP, ML and GME with $t(3)$ errors and different levels of extraneous variables.

Table 1 to the moderately ill-posed design of Table 2, that the performance of the DWP rule actually improves as the condition number $\kappa(X'X)$ increases.

The results of this case are basically similar in nature to the empirical risk results of Fig. 2 and are in each case, greatly superior to the risk results for the ML and ridge estimators. The risk results from experiments with higher and lower condition numbers were virtually the same as those reported in the table and are not reported here.

In terms of identifying extraneous variables, the sampling performance for the case of a $\kappa(X'X) = 90$ design is reported in Table 2. Again, the results for the ill-posed case, mirror those of the well-posed case reported in Table 1. Note the probability (or frequency) of identifying the extraneous variables exceeds 0.95 in all cases.

5.2.2. Experiment 5 – ill-conditioned design and t_3 error distribution

Within the linear model and the above framework, in this experiment we investigate the small sample performance of a moderately ill-conditioned design matrix and a non-normal error process. The design matrix consists of $K = 4$ covariates, generated such that $\mathbf{x}_4 = \mathbf{x}_1 + 0.15\mathbf{e}^*$, where $\mathbf{e}^*$ is a normal (0, 1) random vector. This experiment follows George and McCulloch (1993). The errors are generated from a student-t distribution with three degrees of freedom, normalized to a unit variance. The results plotted in Fig. 3 exhibit, in general, the risk behavior of the previous experiments. The ML empirical risk is 46.167 and the GME risk is 38.43. Thus, the DWP estimator exhibits superior risk performance over the range of $(\beta'\beta)^{1/2}$ parameter space. If $\tilde{\gamma}$ or S are used as variable selection measures, the performance of the experiments in the previous sections is duplicated.

Table 3
Empirical risks of a range of competing estimators for the symmetric case

	Empirical risk of					
	ML	θ^+	*AIC*	*SC*	C_p	*DWP*
Model 1	8.77	8.18	8.50	8.73	8.64	7.25
Model 2	8.89	8.63	8.03	8.40	8.88	7.00

Table 4
Variable selection results of the DWP estimator for the symmetric case

Model	Estimator	Frequency of identifying correctly									Risk
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	
Model 1	DWP	1.00	0.99	0.94	0.89	0.39	0.83	0.82	0.80	0.77	7.25
Model 2	DWP	1.00	1.00	0.93	0.73	0.39	0.81	0.82	0.82	0.82	7.00

5.3. Experiment 6

To include experiments where the design matrix has more variables, we report two experiments that duplicate those carried out by Judge et al. (1987). The experiments involve the linear statistical model $\boldsymbol{y} = A\theta + \boldsymbol{e}$ with nine location parameters, four extraneous variables, a (20×9) design matrix A, where $A'A = I_9$, an error process, $\boldsymbol{e}$, distributed as $\mathrm{N}(0, \sigma^2 I_{20})$ with $\sigma^2 = 1$ and 500 replications.[12] The supports $\boldsymbol{z}_k$ and $\boldsymbol{v}$ are specified as before. The first experiment, Model 1, involves the location vector $\theta = (5, 4, 3, 2, 1, 0, 0, 0, 0,)'$ and the second, Model 2, involves $\theta = (10, 8, 3, 2, 1, 0, 0, 0, 0,)'$.

Table 3 compares the empirical risk of six estimators: ML; DWP; the positive part Stein (1981), θ^+, that shrinks all coefficients toward zero; AIC (Akaike, 1974), SC (Schwarz, 1978); and C_p (Mallows, 1973). The ML estimator is close to the theoretical risk of 9. The traditional and the Stein θ^+ variable selection estimators have empirical risks that are superior to the ML but are *significantly inferior* to the DWP that shrinks but does not eliminate variables. In terms of variables selection, the DWP results are presented in Table 4.

[12] Due to the larger size of each sample (relative to most of the previous experiments), the sampling experiment is performed for 500 samples rather than for 5000 samples for each point on the parameter space.

The results presented here are consistent with the earlier experiments where the higher the parameter space $(\theta'\theta)^{1/2}$, the better the DWP performs relative to the other estimators. Further, increasing the $(\theta'\theta)^{1/2}$ to 37 where $\beta = (25, 20, 15, 10, 5, 0, 0, 0, 0)'$ yields risk of 6.32 while the risk for θ^+ practically equals the ML risk, which is 8.8. In terms of variable identification, the non-extraneous variables are *always* identified while the extraneous variables are identified at least 80% of the time.

6. Summary remarks

In this work a new simultaneous variable selection and estimation rule is developed and investigated. This new rule provides a basis for identifying the non-extraneous and extraneous variables in the design matrix of a linear statistical model and simultaneously yields good estimates. The result is a simple, consistent, one-stage estimator, based on *one* sample of data and a K variable design specification that leads to a basis for semi-parametric entropy-based inference. This data-based procedure is based on weak distributional assumptions and it performs well for both ill and well-posed problems, non-Gaussian error distributions and small samples of data. In this approach coefficient shrinkage and variable elimination are data determined and done on an individual coordinate basis. Further, the choice of prior is data based and endogenously determined. Consequently, the method provides a simple way of introducing and evaluating prior information in the estimation and variable selection process. In contrast to other shrinkage-variable selection procedures, that require a tuning parameter for variable identification and/or determining the degree of shrinkage, upper and lower bounds on the estimated parameters are specified here.

This method is applicable to a wide range of econometric-statistical models (linear and non-linear) and flexible enough to simultaneously cope with a variety of model specification uncertainties. Solution algorithms for these types of non-linear inversion problems are available and easy to implement.

Acknowledgements

This work is based on an earlier joint work with George Judge presented at the 1996 joint Statistical Meetings. In addition to thanking George for his infinite contribution and time put into an earlier version of this work, I also thank Leo Breiman, Douglas Miller, Dale Poirier, Arnold Zellner, an Associate Editor and the referee for their helpful suggestions and comments.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

Amemiya, T., 1980. Selection of regressors. International Economic Review 21, 331–354.

Bercher, J.F., Le Besnerais, G., Demoment, G., 1996. The maximum entropy on the mean method, noise and sensitivity. In Skiling, J., Sibisi, S. (Ed.), Maximum Entropy and Bayesian Studies. Kluwer, Dordrecht.

Berger, J.O., Mortera, J., 1999. Default Bayes factors for nonnested hypothesis testing. Journal of the American Statistical Association 94, 542–554.

Bock, M.E., 1988. Shrinkage estimators: pseudo-Bayes rules for normal mean vectors. In: Gupta, S.S., Berger, J.O. (Eds.), Statistical Decision Theory and Related Topics IV. Springer, New York, pp. 281–298.

Bohrer, R., Yancey, T.A., 1984. Algorithms for numerical evaluation of Stein-like and limited translation estimators. Journal of Econometrics 25, 235–239.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373–384.

Breiman, L., Spector, P., 1992. Submodel selection and evaluation in regression. The X-random case. International Statistical Institute Review 60, 291–319.

Brown, L.D., 1966. On the inadmissibility of invariant estimators of one or more location parameters. Annals of Mathematics and Statistics 37, 1087–1136.

Csiszár, I., 1991. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. The Annals of Statistics 19, 2032–2066.

Donoho, D.L., Johnstone, I.M., Hoch, J.C., Stern, A.S., 1992. Maximum entropy and the nearly black object. Journal of the Royal Statistical Society, Series B 54, 41–81.

George, E., McCulloch, R., 1993. Variable selection in Gibbs sampling. Journal of the American Statistical Association 88 (423), 881–889.

Geweke, J., 1994. Variable selection and model comparison in regression, Working Paper No. 539, Federal Reserve Bank of Minneapolis.

Gokhale, D.V., Kullback, S., 1978. The Information in Contingency Tables. Marcel Dekker, New York.

Golan, A., Judge, G., 1996. A maximum entropy approach to empirical likelihood estimation and inference. University of California, Berkeley, unpublished paper. Presented at the 1997 Summer Econometric Society Meetings.

Golan, A., Judge, G.G., Miller, D., 1996. Maximum Entropy Econometrics: Robust Estimation with Limited Data. Wiley, New York.

Good, I.J., 1963. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. Annals of Mathematical Statistics 34, 911–934.

Hellerstein, J.K., Imbens, G.W., 1999. Imposing moment restrictions from auxiliary data by weighting. The Review of Economics and Statistics 81, 1–14.

Hocking, R.R., 1976. The analysis and selection of variables in linear regression. Biometrics 32, 1–51.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for non-orthogonal problems. Technometrics 1, 55–67.

Imbens, G.W., Johnson, P., Spady, R.H., 1998. Information-theoretic approaches to inference in moment condition models. Econometrica 66, 333–357.

James, W., Stein, C., 1960. Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, CA, pp. 361–379.

Jaynes, E.T., 1957a. Information theory and statistical mechanics. Physics Review 106, 620–630.

Jaynes, E.T., 1957b. Information theory and statistical mechanics II. Physics Review 108, 171–190.

Jaynes, E.T., 1984. Prior information and ambiguity in inverse problems. In: McLaughlin, D.W. (Ed.), Inverse Problems, SIAM Proceedings, American Mathematical Society, Providence, RI, pp. 151–166.

Judge, G., Yi, G., Yancey, T., Teräsvirta, T., 1987. The extended Stein procedure for simultaneous model selection and parameter estimation. Journal of Econometrics 35, 375–391.

Judge, G.G., Bock, M.E., 1978. The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics. North-Holland, Amsterdam.

Kass, R., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Kempthorne, P.J., 1987. Estimating the mean of a normal distribution with loss equal to squared error plus complexity cost. Annals of Statistics 15, 1389–1400.

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

LaFrance, J.T., 1999. Inferring the nutrient content of food with prior information. American Journal of Agricultural Economics 81, 728–734.

Laud, P.W., Ibrahim, J.G., 1995. Predictive model selection. Journal of the Royal Statistical Society, Series B 57, 247–262.

Lavergne, P., 1998. Selection of regressors in econometrics: parametric and nonparametric methods. Econometric Reviews 17, 227–273.

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

Longley, J., 1967. An appraisal of least squares programs from the point of the user. Journal of the American Statistical association 62, 819–841.

Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–665.

Miller, A.J., 1990. Subset Selection in Regression. Chapman & Hall, London.

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression (with discussion). Journal of the American Statistical Association 83, 1023–1036.

Mittelhammer, R., Cardell, S., 1996. On the consistency and asymptotic normality of the data constrained GME estimator of the GML. Mimeo. Washington State University, Pullman, WA.

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

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

Owen, A., 1991. Empirical likelihood for linear models. The Annals of Statistics 19, 1725–1747.

Perloff, J., Shen, Z., 1999. Maximum entropy and Bayesian approaches to the ratio problem. University of California, Berkeley, unpublished paper.

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

Schwarz, G., 1978. Estimating the dimensions of a model. The Annals of Statistics 6, 461–464.

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

Shibata, R., 1981. An optimal selection of regression variables. Biometrika 68, 45–54.

Shore, J.E., Johnson, R.W., 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Transactions on Information Theory IT-26 (1), 26–37.

Sin, C.-Y., White, H., 1996. Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71, 207–225.

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

Soofi, E., 1996. Information theory and bayesian statistics. In: Berry, D.A., Chaloner, K.M., Geweke, J.K. (Eds.), Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner. Wiley, New York, pp. 179–189.

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

Stein, C., 1981. Estimation of the mean of a multivariate normal distribution. The Annals of Statistics 9, 1135–1151.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B 36, 111–147.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Titterington, D.M., 1985. Common structures of smoothing techniques in statistics. International Statistical Review 53, 141–170.

Tobias, J., Zellner, A. Further results on the Bayesian method of moments analysis of multiple regression model. International Economic Review, forthcoming.

Zellner, A., 1988. Optimal information processing and Bayes's theorem. American Statistician 42, 278–284.

Zellner, A., 1996a. Bayesian method of moments/instrumental variables (BMOM/IV) analysis of mean and regression models. In: Lee, J.C., Zellner, A., Johnson, W.O. (Eds.), Prediction and Modelling Honoring Seymour Geisser. Springer, Berlin.

Zellner, A., 1996b. Models, prior information, and Bayesian analysis. Journal of Econometrics 75, 51–68.

Zellner, A., 1997. The Bayesian method of moments (BMOM): theory and applications. In: Fomby, T., Hill, R., (Eds.), Advances in Econometrics, vol. 12, pp. 85–105.

Zellner, A., 1999. New information-based econometric methods in agricultural economics: discussion. American Journal of Agricultural Economics 81, 742–746.

Zellner, A., Vandaele, W., 1975. Bayes–Stein estimators for *k*-means, regression and simultaneous equation model. In: Fienberg, S.E., Zellner, A. (Eds.), Studies in Bayesian Econometrics and Statistics, in Honor of Leonard J. Savage. North-Holland, Amsterdam.

Zhang, P., 1992. On the distributional properties of model selection criteria. Journal of the American Statistical Association 87, 732–737.

Zheng, X., Loh, W.-Y., 1995. Consistent variable selection in linear models. Journal of the American Statistical Association 90, 151–156.