ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY
Lönnrotinkatu 4 B, 00120 Helsinki 12, Finland, tel. 601322

Keskusteluaiheita
Discussion papers

Timo Teräsvirta, Gang Yi* and
George Judge*

MODEL SELECTION, SMOOTHING AND PARAMETER ESTIMATION IN LINEAR MODELS UNDER SQUARED ERROR LOSS†

No. 213 17 July 1986

* Department of Economics, University of Illinois, Champaign, IL 61820, USA

† To be presented at the Econometric Society European Meeting, Budapest, 1-5 September 1986. The authors wish to thank Ilkka Mellin for help in preparing the paper. A major part of the work was done when the first author was visiting the Department of Economics, University of Illinois at Urbana-Champaign. His research was supported by grants from Academy of Finland, the Finnish Cultural Foundation and the Yrjö Jahnsson Foundation.

ISSN 0781-6847

MODEL SELECTION, SMOOTHING AND PARAMETER ESTIMATION IN LINEAR MODELS UNDER SQUARED ERROR LOSS

by

Timo Teräsvirta, Gang Yi, and George Judge

Abstract. This paper considers the joint problem of variable selection and estimation of the parameters in a linear model under certain smoothness conditions. An example of a situation where this problem occurs in practice is polynomial distributed lag estimation with smoothness priors. The model builder usually has to choose the lag length, the order of polynomial and the degree of smoothing in estimating the parameters of the model. Another example could be non-parametric regression when the response is non-zero only if the input value exceeds an unknown threshold. It is suggested that the joint model selection and parameter estimation problem be solved by first generalizing existing model selection criteria in such a way that they can at the same time be applied to choosing the lag length and the order of the lag polynomial, and to determining the value of the smoothness parameter. The performance of this model selection/parameter estimation procedure is investigated through simulation. The measure of performance is the mean squared error of prediction. The procedure seems no worse than certain well-known model selection criteria used for selecting the lag length in the situation where the shape of the lag function is far from that of a smooth low-order polynomial. If the lag function is smooth, the present technique outperforms by a clear margin the model selection criteria which do not make use of any polynomial or other smoothness assumptions.

Key words. Convex smoother, mean squared error of prediction, polynomial distributed lag, shrinkage, smoothing criterion

Timo Teräsvirta: Research Institute of the Finnish Economy (ETLA), Lönnrotinkatu 4 B, SF-00120 Helsinki, Finland

Gang Yi and George Judge: Department of Economics, University of Illinois, Champaign, IL 61820, USA

MODEL SELECTION, SMOOTHING AND PARAMETER ESTIMATION IN LINEAR MODELS UNDER SQUARED ERROR LOSS

1. Introduction

In this paper we continue to pursue the problem of the applied econometrician who seeks a simple criterion for choosing one of a set of competing models to describe the sampling process for a given set of data. In this context we visualize an investigator who has a single sample of data and wants to estimate the unknown parameters of a linear statistical model that are known to lie in a high dimensional space B. However, the investigator may suspect that the data should be modeled by a lower dimensional parameter space $B_1 \subset B$, where B_1 is a subset of B.

In an earlier paper (Judge, et al., 1986), we considered this problem within an orthonormal (K mean) linear statistical model context and proposed an extended Stein criterion called ESP for truncating and partitioning the parameter space. This criterion made use of Stein's (1981) estimator that is a partitioned analogue of the Efron and Morris (1973a) limited translation estimator. In this paper we extend the problem to a general linear statistical model and consider the problem of model selection when estimating a multivariate normal mean under quadratic loss. Our focus will be on simultaneously selecting the model and estimating the unknown parameters in a small sample situation.

The estimator to be derived for this double purpose which also considers how the results will be used, is essentially a Stein-type estimator where shrinkage occurs toward a restricted least squares estimator. The restrictions are essentially exclusion and smoothness restrictions, and a choice between different restrictions is made by a

generalized model selection criterion called the shrinkage criterion. Minimizing the shrinkage criterion with respect to the combination of regressors, the linear restrictions and the amount of shrinkage yields the model and its parameter estimates. This procedure can be seen as a generalization of ordinary model selection criteria which choose a model among a set of alternatives and then use maximum likelihood procedures to estimate the unknown parameters. It may also be viewed as a generalized operational version of a convex smoother (Titterington, 1985).

We shall evaluate the sampling performance of the new procedure by the mean squared error of prediction (MSEP) measure. As an example a polynomial distributed lag estimation problem is analyzed. Within this context a shrinkage criterion estimator behaves quite well even if the polynomial restrictions on the parameters are not correct. When there exists a true set of polynomial restrictions, the procedure is capable of taking advantage of this fact and shows, relative to more traditional procedures, a considerable reduction in the MSEP.

2. Statistical Model and Estimators

Let the p dimensional random vector $\underset{\sim}{\beta}$ have a multivariate normal distribution with mean $\underset{\sim}{\beta}$ and known positive definite covariance matrix Σ. The location vector $\underset{\sim}{\beta}$ is unknown and the objective is to estimate it using an estimator $\delta(\underset{\sim}{b})$ under a squared error loss measure

$$L(\underset{\sim}{\beta},\delta(\underset{\sim}{b})) = (\delta(\underset{\sim}{b}) - \underset{\sim}{\beta})'\Sigma^{-1}(\delta(\underset{\sim}{b}) - \underset{\sim}{\beta}) \qquad (2.1)$$

where the sampling performance of the estimator $\delta(\underset{\sim}{b})$ will be evaluated by its risk function

$$\rho(\underset{\sim}{\beta},\delta(\underset{\sim}{b})) = E[L(\underset{\sim}{\beta},\delta(\underset{\sim}{b}))]. \tag{2.2}$$

In econometrics one common problem that gives rise to the above involves estimating the location vector for the normal linear statistical model

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{e} \qquad \underset{\sim}{e} \sim N(\underset{\sim}{0},\sigma^2 I). \tag{2.3}$$

where $\underset{\sim}{y}$ is an (n×1) vector, X is an (n×p) matrix of regressors of rank p, $\underset{\sim}{\beta}$ is a (p×1) parameter vector, and $\underset{\sim}{e}$ is an (n×1) vector of normal random variables. The maximum likelihood (ML) estimator of the unknown location vector $\underset{\sim}{\beta}$ is $\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{y}$ and $\Sigma = \sigma^2(X'X)^{-1}$ is the precision matrix.

For the orthonormal case when $X'X = I_p$ and $\Sigma = \sigma^2 I_p$, Stein (1956) showed that the usual ML estimator $\delta^o(\hat{\underset{\sim}{\theta}}) = \hat{\underset{\sim}{\theta}} = X'y$ is inadmissible, under squared error loss, for $p \geq 3$. James and Stein (1961) demonstrated that the estimator

$$\delta^{(JS)}(\hat{\underset{\sim}{\theta}}) = (1 - a\sigma^2/\hat{\underset{\sim}{\theta}}'\hat{\underset{\sim}{\theta}})\hat{\underset{\sim}{\theta}} \tag{2.4}$$

has uniformly smaller risk than $\delta^o(\hat{\underset{\sim}{\theta}})$ under conditions normally fulfilled in practice. Baranchik (1964) demonstrated the inadmissibility of $\delta^{(JS)}(\hat{\underset{\sim}{\theta}})$ by the positive part estimator

$$\delta^{(JS)+}(\hat{\underset{\sim}{\theta}}) = I_{[a\sigma^2,\infty)}(\hat{\underset{\sim}{\theta}}'\hat{\underset{\sim}{\theta}})(1 - a\sigma^2/\hat{\underset{\sim}{\theta}}'\hat{\underset{\sim}{\theta}})\hat{\underset{\sim}{\theta}}. \tag{2.5}$$

If $X'X \neq I_p$ the above results hold under a mean squared error of prediction (MSEP) criterion $E[(\delta(\underset{\sim}{b}) - \underset{\sim}{\beta})'X'X(\delta(\underset{\sim}{b}) - \underset{\sim}{\beta})]$. Using, for example, the spectral decomposition $X'X = C\Lambda C'$ and defining $S^{-1/2} = C\Lambda^{-1/2}$, the problem may be transformed into an estimation problem of an orthonormal model $\underset{\sim}{y} = Z\underset{\sim}{\theta} + \underset{\sim}{e}$, with $Z = XS^{-1/2}$ and $\underset{\sim}{\theta} = S^{1/2}\underset{\sim}{\beta}$. The above results are

valid for $\delta^{(JS)}(\hat{\underset{\sim}{\theta}})$ and since the MSEP is invariant to this transformation. Consequently, for estimation purposes the general case can be handled by transforming the model into the orthogonal space and estimating the parameters there.

Unfortunately, risk gains with practical importance may be expected from $\delta^{(JS)}(\hat{\underset{\sim}{\theta}})$ compared to $\delta^{o}(\hat{\underset{\sim}{\theta}})$ only when the prior with mean $\underset{\sim}{\theta}_0 = 0$ is relevant, that is, $\|\underset{\sim}{\theta}\|$ is not far from zero. If the original model contains redundant variables and/or exact or approximate linear relationships between parameters, it is often advisable to focus attention to them instead of transforming the problem first. Omitting redundant variables and applying exact or approximate linear restrictions may, in that case, lead to more substantial risk gains than does shrinkage in the orthogonal space. Within the context of a squared error loss measure the problems are i) which variables to delete, ii) how to find the possible linear restrictions and the degree of approximation, and iii) how to estimate the corresponding parameters. A solution to these problems is outlined in this paper. The estimator that results may not be minimax but if achieving a risk gain relative to MLE over a large range of the parameter space is our main concern, the present approach will be of interest.

3. Shrinkage Criteria for Model Selection and Parameter Estimation

The model selection-parameter estimation procedure to be discussed is related to the parameter truncation and partitioning procedures developed in Judge, et al. (1986) and to the partitioned Stein-type estimator proposed recently by Teräsvirta and Yi (1986). The latter is based on the idea of Efron and Morris (1973b) who suggested partitioning

the problem of estimating $\underset{\sim}{\theta}$ into a set of subproblems, each of which is solved by estimating the corresponding subvector $\underset{\sim}{\theta}_j$, j=1, ...,m; $\underset{\sim}{\theta} = (\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_m)'$ by a Stein estimator (2.3) or (2.4). If the partition is known, the resulting estimator of $\underset{\sim}{\theta}$ is minimax provided the Stein estimators for $\underset{\sim}{\theta}_j$ are minimax. To ensure a large risk gain over the ML estimator, an appropriate partition may in practice be determined from the data. Let $\hat{\underset{\sim}{\theta}} = (\hat{\underset{\sim}{\theta}}_1,\ldots,\hat{\underset{\sim}{\theta}}_m)'$ where the $(p_j x 1)$ vector $\underset{\sim}{\theta}_j$ is the ML estimator of $\underset{\sim}{\theta}$, and $\sum_{j=1}^{m} p_j = p$. The j-th subvector of the partitioned Stein-type estimator is $\delta_j(\hat{\theta}) = (1-\hat{c}_j)\hat{\theta}_j$, where $\hat{c}_j = 1 - h\tilde{\sigma}^2 p_j/\hat{\theta}'\hat{\underset{\sim}{\theta}}_j$, so that $\delta(\hat{\theta}) = (1-\hat{C})\hat{\underset{\sim}{\theta}}$ and $\hat{C} = \text{diag}(\hat{c}_1 I_{p_1},\ldots,\hat{c}_m I_{p_m})$. Teräsvirta and Yi (1986) suggest that a good partition α may be obtained by minimizing the shrinkage criterion

$$SC(\alpha,\underset{\sim}{c}^{\alpha}) = \hat{\sigma}^2(\underset{\sim}{c}^{\alpha}) + \tilde{\sigma}^2 p(\underset{\sim}{c}^{\alpha}) f(n,p)/n \qquad (3.1)$$

where $\underset{\sim}{c}_{\alpha} = (c_1^{\alpha},\ldots,c_{m(\alpha)}^{\alpha})$, $\hat{\sigma}^2(\underset{\sim}{c}^{\alpha}) = n^{-1}(y - X\delta^{\alpha}(\hat{\theta}))'(y - X\delta^{\alpha}(\hat{\theta}))$, $\tilde{\sigma}^2 = (n-p)^{-1}(y-Xb)'(y-Xb)$, $p(\underset{\sim}{c}^{\alpha}) = \sum_{j=1}^{m(\alpha)} (1-c_j^{\alpha})p_j^{\alpha}$ and $\sum_{j=1}^{m(\alpha)} p_j^{\alpha} = p$, over α and $\underset{\sim}{c}^{\alpha}$. Furthermore, f(n,p) is a positive function of n and p, and $\lim_{n\to\infty} f(n,p)/n = 0$. This yields $\delta_j^{\alpha}(\hat{\theta}) = (1-\hat{c}_j^{\alpha})\hat{\underset{\sim}{\theta}}_j^{\alpha}$, j=1,...,m(α), where

$$\hat{c}_j^{\alpha} = h\tilde{\sigma}^2 p_j^{\alpha}/\hat{\theta}_j^{\alpha\prime}\hat{\theta}_j^{\alpha}, \quad j=1,\ldots,m(\alpha)$$

with $h = f(n,p)/2$.

We shall apply the idea of using a shrinkage criterion to the problem of reducing the dimension of the model and selecting a set of possibly approximate linear restrictions when the design is non-orthogonal. For simplicity, only the case where $C^{\alpha} \equiv cI$ is considered.

The shrinkage criteria of this paper are a generalization of ordinary model selection criteria. Consider the model (2.3) with a set of linear smoothness restrictions

$$R\underset{\sim}{\beta} = \underset{\sim}{0} \tag{3.2}$$

where R is an (m(R)×p) matrix of fixed finite elements of rank m(R). Assume furthermore that we have a set $\bar{R}$ of feasible restriction matrices under consideration. Our estimator is

$$\underset{\sim}{b}_R(c) = c\underset{\sim}{b}_R + (1-c)\underset{\sim}{b} = \underset{\sim}{b} - cUR'(RUR')^{-1}R\underset{\sim}{b} \tag{3.3}$$

where $\underset{\sim}{b}_R = \underset{\sim}{b} - UR'(RUR')^{-1}R\underset{\sim}{b}$ with $U = (X'X)^{-1}$ is the restricted least squares estimator of $\underset{\sim}{\beta}$ based on the restriction (3.2). The estimator (3.3) is a Stein-like rule that shrinks $\underset{\sim}{b}$ toward the restricted least squares estimator $\underset{\sim}{b}_R$. To make (3.3) operational, an optimal combination of c, p and $R \in \bar{R}$ may be chosen by minimizing a member of the following class of shrinkage criteria over c, p and $\bar{R}$:

$$SC(c,p,R) = \hat{\sigma}^2(c,p,R) + \tilde{\sigma}^2 p(c,R)f(n,p)/n \tag{3.4}$$

where $\hat{\sigma}^2(c,p,R) = n^{-1}(\underset{\sim}{y}-X\underset{\sim}{b}_R(c))'(\underset{\sim}{y}-X\underset{\sim}{b}_R(c))$ and

$$\begin{aligned} p(c,R) &= tr(XUX' - cXUR'(RUR')^{-1}RUX') \\ &= p - cm(R). \end{aligned} \tag{3.5}$$

Expression (3.5) is called the equivalent number of regressors in the model and it is a decreasing function of c; p(1,R) = p - m(R) and p(0,R) = p. For further discussion see Engle et al. (1986), Teräsvirta (1986) and Teräsvirta and Yi (1986).

Function f is as in (3.1) and specifies a member of the class of criteria of type (3.4). For example, if $f(n,p) \equiv 2$, we are dealing with a generalization of the C_p model selection criterion of Mallows (1973). Thus, if we minimize (3.4) with $f(n,p) \equiv 2$, this amounts to minimizing an unbiased estimator of the conditional MSEP

$$\mathrm{MSEP}(b_R(\hat{c}) \mid c = \hat{c}) = E\{(\underset{\sim}{b}_R(\hat{c}) - \underset{\sim}{\beta})'X'X(\underset{\sim}{b}_R(\hat{c}) - \underset{\sim}{\beta}) \mid c = \hat{c}\}$$

over c, p and R.

Let us for a moment assume that p and R are fixed and find $\hat{c}$ by minimizing (3.4) over c. A differentiation yields

$$\frac{\partial SC(c,R)}{\partial c} = \hat{\sigma}^2(c,p,R)' + n^{-1}\tilde{\sigma}^2 p'(c,R)f(n,p)$$

$$= 2c\hat{a}_p - n^{-1}\tilde{\sigma}^2 m(R)f(n,p) \qquad (3.6)$$

where $\hat{a}_p = n^{-1}b'R'(RUR')^{-1}Rb$. Setting (3.6) equal to zero and solving for c yields

$$\hat{c} = \frac{\tilde{\sigma}^2 f(n,p)m(R)}{2n\hat{a}_p} . \qquad (3.7)$$

In practice, a positive part estimator

$$(1-\hat{c})^+ = [1 - \frac{\tilde{\sigma}^2 f(n,p)m(R)}{2n\hat{a}_p}]^+ \qquad (3.8)$$

where $[a]^+ = \max(a,0)$ is always recommended.

Note that $\underset{\sim}{b}_R(\hat{c})$ with (3.7) or (3.8) is a Stein-type estimator of the form considered by Judge and Bock (1978, pp. 240-241). Its risk properties and superiority over the ML estimator have been discussed by Mittelhammer (1984).

It might be pointed out that defining another family of model selection criteria

$$SC_2(c,p,R) = \ln\hat{\sigma}^2(c,R) + p(c,p,R)f(n,p)/n \qquad (3.9)$$

and using it for finding a good value of c leads to a second-order equation in c; see Teräsvirta and Yi (1986). Selecting a proper root of this equation yields a solution that differs from (3.7) by a factor of order $O(n^{-1})$. For example, choosing $f(n,p) \equiv 2$ in (3.9) is equivalent to generalizing AIC to the problem of determining c.[1]

As pointed out above, the shrinkage criteria may be used as model selection criteria by varying the combination of regressors and thus p, and selecting the regressors (and p) and the degree of shrinkage c which minimize the shrinkage criterion. Extending this procedure to cover the selection of smoothness conditions requires that for each combination of regressors considered, there is a set of well-defined smoothness restrictions. The model selection-parameter estimation problem is then solved by minimizing the shrinkage criterion over the permitted combinations of regressors, the degree of shrinkage c and the feasible set of smoothness restrictions $\overline{R}$. The criteria (3.4) and (3.9) permits one to i) select the model, ii) choose the linear restrictions for the structural parameters in the model and, iii) determine the degree of shrinkage towards the restricted least squares estimator specified by the chosen restrictions.

A typical case of this model selection-parameter estimation problem is such that there is an ordering among the regressors. Selecting a combination of regressors is then equivalent to specifying p. An example of such a situation is polynomial distributed lag estimation where the

parameters of a finite distributed lag model are related by assuming that they lie, either exactly or approximately, on a polynomial of low degree. This topic has recently been reviewed and discussed, for example, in Hendry, et al. (1984), Judge, et al. (1985, Chapter 9), and Trivedi (1984). If lags are removed from the model one after the other starting with the longest one, one linear restriction will be removed for each lag; but the remaining restrictions will not be affected by an omission of the longer lags. In practical applications, the lag length is often unknown, and usually there is uncertainty as to the degree of polynomial that is most appropriate. The analytical properties of the shrinkage criterion estimators are unknown. We shall investigate them through simulation when the shrinkage criteria are applied to the problem of selecting the lag length and the degree of polynomial in finite distributed lag models. This topic is discussed in sections 5 and 6.

4. Model Selection in an Orthogonal Space

The shrinkage criterion (3.1) that operates in the orthogonal space can easily be generalized for model selection purposes. The idea is similar to that of the ESP criterion discussed in Judge, et al. (1986). In this context, consider the omission of variables whose coefficient estimates are small in absolute value, where the decision of omitting variables is based on minimizing (3.1). An advantage of these criteria is that they offer a possibility of differentiating the degree of shrinkage.

An estimation problem with approximate linear restrictions may be transformed into an orthogonal space such that the restrictions in the

transformed space are simply exclusion restrictions; for the transformation see Judge and Bock (1978, p. 84). It is, therefore, appropriate to shrink ML estimates in the orthogonal space toward zero. However, this can be expected to work only if (3.2) holds. If, in reality, $R\underset{\sim}{\beta} \neq \underset{\sim}{0}$, all the parameters of the transformed model may be large in absolute value. Because of this, applying ESP or shrinkage criteria, which omits variables and shrinks the remaining coefficient towards zero in the orthogonal space, is unlikely to lead to substantial risk gains compared to the risk of the ML estimator. Consequently, procedures that select models and perform the shrinkage in the original space may in general be expected to do a better job. One exception is the case of ill-conditioned data where one or more of the characteristic roots of the X'X matrix may be near zero. For these reasons, within the context of the principal components parameter space, ESP may work well in reducing the MSEP (Hill and Judge, 1986).

5. Sampling Experiment

To obtain information on the sampling performance of the alternative model selection-estimation procedures Monte Carlo sampling experiments were performed. The experiments involved the general linear statistical model $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{e}$, where $\underset{\sim}{\beta}$ contains 8 non-zero location parameters and 7 zeroes, and $\underset{\sim}{e} \sim N(0,\sigma^2 I)$ with $\sigma^2 = .25, 1, 4$. The relevant part of the design matrix X was (30x8) where the jth column was the jth lag of the first column x_t. Variable x_t was generated by an AR(1) process with $x_t = \rho x_{t-1} + v_t$ with $\rho = 0.9$ and $v_t \sim N(0,1)$. Models M1, M2 and M3 with the following location vectors that reflect alternative lag structures for the distributed lag were used:

M1: $\underset{\sim}{\beta}_1 = (1.87, -2.48, -.25, -.25, -.25, -.25, -.25, -.25, \underset{\sim}{0})'$

M2: $\underset{\sim}{\beta}_2 = (1.02, 1.54, 2.05, 1.23, 0.74, 0.45, 0.26, 0.16, \underset{\sim}{0})'$

M3: $\underset{\sim}{\beta}_3 = (2.07, 1.63, 1.25, 0.92, 0.64, 0.41, 0.22, 0.10, \underset{\sim}{0})'$.

The first location vector $\underset{\sim}{\beta}_1$ represents a situation where no low-order polynomial is a reasonable approximation of the lag function. The lag function depicted by $\underset{\sim}{\beta}_2$ has a sharp peak at lag 2 while the tail is smooth. Finally, the non-zero regression coefficients of $\underset{\sim}{\beta}_3$ lie exactly on a second degree polynomial. In this case it can be expected that the shrinkage of $\underset{\sim}{b}$ towards $\underset{\sim}{b}_R$, where R contains the second (or higher) order polynomial conditions is strong, i.e., $\hat{c}$ is close to one in the operational version of (3.3). For comparability, the elements of all three vectors have been chosen in such a way that the norms of the vectors are equal.

Interest is focused on the performance of (3.3) where c, p and R are determined by minimizing (3.4) with $f(n,p) \equiv 2$. Within this context, the shrinkage criterion is thus a generalization of C_p. The degree of polynomial q-1 is varied from zero to six and an alternative with no polynomial restrictions at all is included in the minimization. If the degree of polynomial equals q-1, say and $p > q$, then the jth row of the ((p-q) x p) constraint matrix R_q is $[R_q]_j = (0,\ldots,0,1, -\binom{q}{1},\ldots, (-1)^j\binom{q}{j},\ldots,(-1)^q,0,\ldots,0)$ cf. for example, Teräsvirta (1976). The number of zeros in the beginning of the row is j-1 and at the end p-q-j. The lag length is varied from 6 to 15; however, for $p \leq 7$ we set $q_{max} = p-1$.

To evaluate the performance of the "generalized C_p" (GC_p) we included some ordinary model selection criteria in the MSEP comparisons. Since all the criteria to be used are based on the squared sum of residuals, the MSEP is a natural choice for assessing their performance. The model selection criteria were allowed to select one of the models with no polynomial restrictions and lag lengths from 6 to 15. The criteria used were C_p, AIC and SC; for exact definitions cf. e.g., Judge, et al. (1985, Chapter 21). Of these, the SC is dimension consistent in the sense that asymptotically it selects the correct lag length with probability one. For the other two criteria there is a positive probability of overestimating the lag length even asymptotically. In fact, they have the same asymptotic behavior. However, in this experiment n = 30, so that we are far from the asymptotic situation. This will be obvious from the results.

Table 1 contains the theoretical values of R^2 for each of location vectors M1, M2 and M3 and thus conveys an idea of how large the values σ^2 actually were. When $\sigma^2 \leq 1$, the models are quite good with high R^2. When $\sigma^2 = 4$, R^2 is below 0.9 and as n = 30, at least this case may call for smoothness restrictions and estimators of type (3.3).

6. Results

The computations were performed on a CYBER 175 computer. The main subroutine library was ISML and the random number generator GGNML of ISML was used to generate x_t (once) and 100 samples of y_t. The parameters of M1, M2 and M3 were estimated for each sample, and an MSEP estimate was computed from the 100 samples. Another interesting statistic, the sample mean of the equivalent number of regressors, is also reported.

Before studying the performance of the shrinkage criterion we shall note some general features of the model selection criteria in these experiments. Teräsvirta and Mellin (1986) have derived finite sample approximations to significance levels of model selection criteria in the case of nested model alternatives. When n = 30 and the models considered have 8 to 15 regressors (7 to 14 lags), the significance levels of AIC, C_p, and SC are .537, .340, and .229, respectively. Table 2 contains the corresponding observed frequencies for these criteria in M1 to M3 when $\sigma^2 = .25$. Although in the present experiment, underestimation of the true lag length (seven) is allowed, the frequencies correspond quite closely to the significance levels. The tendency for underestimation increases with σ^2. From the significance levels we may conclude that the sample mean of the equivalent number of regressors will be largest for AIC, clearly smaller for C_p and smallest for SC. AIC and C_p are asymptotically equivalent criteria, but their significance levels imply considerable differences in their behavior in this experiment.

6.1 Model M1

Table 3 contains the results for the M1 parameter vector, where large gains cannot be expected from imposing polynomial restrictions. The shrinkage criterion GC_p has almost as small an MSEP as SC when $\sigma^2 = 0.25$. As the error variance increases ($\sigma^2 \geq 1$), the GC_p is still better than the AIC and C_p but not as good as SC. It is worth mentioning that, depending on σ^2, the polynomial restrictions selected by the GC_p are those of a polynomial of order zero in 75 to 79 cases out of 100. Thus using the criterion very often leads to deleting a few lags and shrinking the remaining ML estimates towards their grand mean.

Theoretically, one could then expect the GC_p to improve upon the ordinary C_p, which does not shrink the ML estimates of the selected lag coefficients. That also happens in this experiment when $\sigma^2 \leq 1$. However, considering a whole set of polynomial restrictions together with the lag length and choosing a combination of them adds to the uncertainty compared to the case where the only polynomial considered is of degree zero. The extra added uncertainty translates into increased risk in estimation. It is thus conceivable that GC_p could have higher risk than C_p in some cases where all low-order polynomials are very bad approximations of the true lag function. This is the case here when $\sigma^2 = 4$.

The AIC criterion is inferior to the C_p and SC criteria in M1 for reasons explained above. Note that the strong tendency of AIC of overestimating the dimension of the model does not depend on $\underset{\sim}{\beta}$ or σ^2 but is dependent on the sample size which was not varied in this experiment.

The sample mean of the equivalent number of regressors (ENR) decreases as σ^2 increases. This is natural as an increase in the error variance is equivalent to a decrease in the relative amount of sample information and thus in the number of parameters it is profitable to estimate from the data.

6.2 Model M2

From Table 4 it is seen that M2 is already a favorable case for shrinkage criteria, in spite of the fact that the lag function still has a rather jagged shape. For $\sigma^2 \geq 1$, GC_p has a lower MSEP than the

model selection criteria. The shrinkage in GC_p is in that case substantial as indicated by the values of ENR. A wide variety of combinations of lag length and degree of polynomial are chosen in the 100 trials. On average, however, the longer the lag length, the higher the degree of polynomial.

6.3 Model M3

The parameters of M3 lie exactly on a second degree polynomial, so that the actual number of coefficients to be estimated is only three. As can be expected, the model selection criteria are overwhelmed by GC_p because the former are only applied to selecting the lag length and not the degree of polynomial; see Table 5. This limitation is probably an advantage in M1 and to an extent in M2 but has an adverse effect on the performance of the criteria in M3. Under this specification the GC_p criterion results in the selection of a wide spectrum of lag lengths with the second order polynomial as the most popular alternative. This is understandable from (3.5) where now $m(R) = p-q$. When $c \to 1$, the weight of p in (3.5) approaches zero and only q matters. Of course, the choice of p still affects the residual variance of the model.

Overall, the shrinkage criterion behaves well in the experiments conducted here. It is not appreciably inferior to model selection criteria when the potential gain from polynomial restrictions is small. On the other hand, the GC_p is able to benefit from polynomial restrictions if they do hold. Both the model selection criteria and the shrinkage criterion are clearly superior to the ML estimator because M1-M3 contain several redundant lags.

6.4 Other experiments

We also experimented with the use of an orthogonal space in solving the model selection and estimation problems. The lag length was varied from 6 to 15 while it was assumed that the non-zero coefficients of the lag function were points on a second-order polynomial. The parameter space was transformed for each lag length using the Judge and Bock (1978, p. 84) transformation. In the orthogonal space, the ESP criterion was employed to omit variables and shrink the remaining coefficients towards zero. The final estimator was chosen by the ESP (the minimization was extended over the lag length). It was not possible to vary the degree of polynomial due to lack of comparability of the values of the statistic for different polynomials.

In theory this approach should have worked in connection with M3 where the coefficients indeed were points of a second order polynomial. In spite of this the results were disappointing and showed at best only marginal risk gains over the ML estimator. Therefore, we were not able to suggest a suitable technique for doing the model selection and parameter estimation partially in an orthogonal space when the original model is non-orthogonal.

7. Directions for Further Research

Another way of taking account of smoothness restrictions (3.2) is to apply the estimator

$$b_R^*(\lambda) = (X'X+\lambda R'R)^{-1} X'y \qquad (7.1)$$

that has been discussed, among others, by Golub et al. (1979), Engle et al. (1983) and Titterington (1985). In (7.1), λ is a smoothing

parameter. This estimator coincides with (3.3) when $\lambda \to 0$ and $c \to 1$ so that both (7.1) and (3.3) in that case approach b_R. Their paths leading to b_R are different, however, which makes it interesting to know whether these differences are also reflected in the MSEP of these estimators. It is likely that (7.1) will emphasize local smoothness of the lag function more than (3.3); see e.g., Titterington (1985).

Using (7.1) requires determining the value of the smoothing parameter λ. Golub, et al. (1979) have suggested a solution which involves minimizing the generalized cross-validation criterion (GCV). Within this context Teräsvirta (1986) has considered the possibility of generalizing other existing model selection criteria to find a good smoother and evaluated the asymptotic properties of the generalized criteria. These generalizations can also be extended to choosing p and R. However, λ can only be determined numerically, which may cause a relatively heavy computational burden if p and R are varied as well. For this reason, (3.3) combined with a member of (3.4) to determine c is computationally an attractive alternative because there is an analytical expression for $\hat{c}$. The research to compare the small sample MSEP properties of (3.3) with those of (7.1) is currently under way.

8. Summary comments

For the general linear statistical model we have considered under squared error loss and within a simultaneous context the problems of i) which explanatory variables to include and ii) how to estimate the corresponding parameters. Consequently, we have sought a solution as to which zero order constraints to impose when estimating $\underset{\sim}{\beta}$ and how to

best estimate $\underset{\sim}{\beta}$ subject to these constraints. In this situation, under a MSEP criterion, it is tempting to reparameterize the model in an orthonormal ($\underset{\sim}{\theta}$ space) and use the ESP or some other model selection criteria to select a proper subspace. Unfortunately, our prior information is usually on the original $\underset{\sim}{\beta}$ space and not on the transformed θ parameters. Because of this, some model selection procedures such as ESP and (3.1) do not perform any better than unconstrained maximum likelihood procedures.

To mitigate this outcome a generalized C_p (GC_p) criterion was developed and applied to alternative parameter structures for a distributed lag problem. Using Monte Carlo procedures the proposed GC_p procedure under a MSEP measure, in each case compared well with the traditional information criteria and outperformed the unconstrained ML estimator. The GC_p criterion is easy to apply and it has the novel feature of simultaneously considering, within a decision theory context, the joint problems of model specification and estimation in a non-orthogonal setting.

Footnote

[1]Teräsvirta and Mellin (1986) define three categories of model selection criteria, two of which appear generalized in (3.4) and (3.9). A corresponding generalization of their third category would be

$$SC_3(c,p,R) = \hat{\sigma}^2(c,p,R) + \hat{\sigma}^2(c,p,R)p(c,R)f(n,p(c,R))/n. \quad (3.10)$$

This category contains several well-known criteria: one example is the generalized cross-validation criterion GCV (Golub et al., 1979). However, deriving the "shrinker" c by minimizing this generalization would have to be done numerically. The reasons for that are that f may now be a function of c and $\hat{\sigma}^2(c,p,R)$ also appears in the second term of (3.10). Minimizing the value of the criterion then amounts to solving a higher degree equation in c.

References

Baranchik, A. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. Department of Statistics, Stanford University, Technical Report No. 51.

Efron, B. and C. Morris (1973a). The Stein estimation rule and its competitors--An empirical Bayes approach. Journal of the American Statistical Association 68, 117-130.

Efron, B. and C. Morris (1973b). Combining possibly related estimation problems. Journal of the Royal Statistical Society B 35, 379-421 (with Discussion).

Engle, R. F., C. W. J. Granger, J. Rice and A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. Journal of the American Statistical Association 81, 310-320.

Golub, G. H., M. Heath and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215-223.

Hendry, D. F., A. R. Pagan and J. D. Sargan (1984). Dynamic specification. Chapter 18 in: Z. Griliches and M. A. Intriligator (eds.): Handbook of econometrics, Vol. 2. Amsterdam: North-Holland.

Hill, R. C. and G. G. Judge (1986). Improved prediction in the presence of multicollinearity. Department of Economics, University of Illinois at Urbana-Champaign, unpublished paper.

James, W. and C. Stein (1961). Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium, Vol. 1. Berkeley: University of California Press, 361-379.

Judge, G. G. and M. Bock (1978). The statistical implications of pretest and Stein-rule estimators in econometrics. Amsterdam: North-Holland.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl and T. C. Lee (1985). The theory and practice of econometrics, second edition. New York: Wiley.

Judge, G. G., G. Yi, T. Yancey and T. Teräsvirta (1986). The extended Stein procedure (ESP) for simultaneous model selection and parameter estimation. Department of Economics, University of Illinois at Urbana-Champaign, unpublished paper.

Mallows, C. L. (1973). Some comments on C_p. Technometrics 15, 661-676.

Mittelhammer, R. C. (1984). Restricted least squares, pre-test, OLS and Stein rule estimators: Risk comparisons under model misspecification. Journal of Econometrics 25, 151-164.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium, Vol. 1. Berkeley: University of California Press, 197-206.

Stein, C. (1981). Estimation of the parameters of a multivariate normal distribution. Annals of Statistics 9, 1131-1151.

Teräsvirta, T. (1976). A note on bias in the Almon distributed lag estimator. Econometrica 44, 1317-1321.

Teräsvirta, T. (1986). Smoothness in regression: Asymptotic considerations, in: I. B. MacNeill and G. J. Umphrey (eds.): Proceedings of the Symposia in Statistics and Festschrift in Honour of V. M. Joshi. Dordrecht: Reidel.

Teräsvirta, T. and I. Mellin (1986). Model selection criteria and model selection tests in regression models. Scandinavian Journal of Statistics (in press).

Teräsvirta, T. and G. Yi (1986). Partitioned Stein-type estimators for linear models. Department of Economics, University of Illinois at Urbana-Champaign, unpublished paper.

Titterington, D. M. (1985). Common structure of smoothing techniques in statistics. International Statistical Review 53, 141-170.

Trivedi, P. K. (1984). Uncertain prior information and distributed lag estimation. Chapter 7 in: D. F. Hendry and K. F. Wallis (eds.): Econometrics and quantitative economics. Oxford: Blackwells.

Table 1

Theoretical Values of the Coefficient Determination R^2 in Models M1, M2 and M3

Model	σ^2		
	0.25	1.0	4.0
M1 - R^2	.974	.904	.701
M2 - R^2	.989	.958	.851
M3 - R^2	.988	.955	.842

Table 2

Observed Frequencies in 100 Trials for Choosing a Longer Lag than Eight for M1-M3 and $\sigma^2 = .25$ Using Model Selection Criteria, and Significance Levels for the Same Criteria when the True Model has Lag Length 7, the Longest Lag Considered is 14 and n = 30

Criterion	Model M1	M2	M3	Significance level
C_p	36	33	33	.340
AIC	55	53	48	.537
SC	27	26	18	.229

Table 3

The Mean Squared Error of Prediction and Equivalent Number of Regressors (ENR) of Shrinkage and Model Selection Procedures for M1

Shrinkage or model selection criterion	$\sigma^2 = 0.25$		$\sigma^2 = 1$		$\sigma^2 = 4$	
	MSEP	ENR	MSEP	ENR	MSEP	ENR
GC_p (3.4)	2.50	8.38	10.19	7.75	41.14	6.42
C_p	2.64	8.93	10.39	8.37	39.34	7.58
AIC	2.97	10.14	11.94	9.80	45.10	9.03
SC	2.44	8.53	10.10	7.66	36.67	7.00
MLE*	3.66	15.00	14.64	15.00	58.54	15.00

*The ML estimator estimates all 15 coefficients; no model selection.

Table 4

The Mean Squared Error of Prediction and Equivalent Number of Regressors (ENR) of Shrinkage and Model Selection Procedures for M2

Shrinkage or model selection criterion	$\sigma^2 = 0.25$		$\sigma^2 = 1$		$\sigma^2 = 4$	
	MSEP	ENR	MSEP	ENR	MSEP	ENR
GC_p (3.4)	2.72	7.25	9.34	5.72	33.76	4.65
C_p	2.63	8.58	10.03	7.84	36.06	7.25
AIC	3.02	9.90	11.78	9.50	43.51	8.76
SC	2.54	8.19	9.88	7.48	34.66	6.85
MLE*	3.66	15.00	14.64	15.00	58.54	15.00

*The ML estimator estimates all 15 coefficients; no model selection.

Table 5

The Mean Squared Error of Prediction and Equivalent Number of Regressors (ENR) of Shrinkage and Model Selection Procedures for M3

Shrinkage or model selection criterion	$\sigma^2 = 0.25$		$\sigma^2 = 1$		$\sigma^2 = 4$	
	MSEP	ENR	MSEP	ENR	MSEP	ENR
GC_p (3.4)	1.55	3.89	6.74	3.59	23.95	3.30
C_p	2.47	8.18	9.68	7.54	34.91	7.17
AIC	2.90	9.63	11.42	9.21	42.70	8.72
SC	2.41	7.58	9.36	7.13	32.60	6.69
MLE*	3.66	15.00	14.64	15.00	58.54	15.00

*The ML estimator estimates all 15 coefficients; no model selection.

ELINKEINOELÄMÄN TUTKIMUSLAITOS (ETLA)
The Research Institute of the Finnish Economy
Lönnrotinkatu 4 B, SF-00120 HELSINKI Puh./Tel. (90) 601 322

KESKUSTELUAIHEITA - DISCUSSION PAPERS ISSN 0781-6847

No 193 KARI ALHO, An Analysis of Financial Markets and Central Bank Policy: A Flow-of-Funds Approach to the Case of Finland. 15.01.1986. 44 p.

No 194 PAAVO OKKO, Julkisen rahoitustuen tehokkuus ja sen kohdentuminen eteläsuomalaisiin teollisuusyrityksiin. 15.01.1986. 46 s.

No 195 JUSSI KARKO, The Measurement of Productivity and Technical Change at Industry Level: An Application of Micro-Data to Industry Analysis. 16.01.1986. 40 p.

No 196 MARKKU RAHIALA, Teollisuusyritysten tuotantosuunnitelmien toteutumiseen vaikuttavat tekijät suhdannebarometriaineiston valossa tarkasteltuina. 20.01.1986. 53 s.

No 197 ERKKI KOSKELA, Taxation and Timber Supply under Uncertainty and Liquidity Constraints. 31.01.1986. 24 p.

No 198 PEKKA YLÄ-ANTTILA, Metsäteollisuuden kannattavuusvaihteluiden kokonaistaloudellisista vaikutuksista. 13.02.1986. 14 s.

No 199 JUHA KETTUNEN, Kansaneläke- ja sairausvakuutuksen rahoituksesta. 10.03.1986. 28 s.

No 200 JUKKA LESKELÄ, Välitysvaluutat ja ulkomaankaupan laskutus. 10.03.1986. 22 s.

No 201 VESA KANNIAINEN - HANNU HERNESNIEMI, Asset Structure, Indebtedness, and the Rate of Return on Capital in a Sample of Finnish Manufacturing Firms in 1961 - 1983. 11.03.1986. 31 s.

No 202 ANTTI RIPATTI, Teollisuus- ja ulkomaankauppatilaston yhdisteen hyödyntäminen. 20.03.1986. 31 s.

No 203 SYNNÖVE VUORI, Returns to R & D in Finnish and Swedish Manufacturing Industries. 20.03.1986. 23 p.

No 204 VESA KANNIAINEN, On the Effects of Inflation: The Debtor-Creditor Hypothesis Reconsidered. 20.03.1986. 15 p.

No 205 PEKKA ILMAKUNNAS, Aggregation of Micro Forecasts. 01.04.1986. 17 p.

No 206 JUSSI RAUMOLIN, Recent Trends in the Development of the Forest Sector in Finland and Eastern Canada. 04.04.1986. 40 p.

No 207 VESA KANNIAINEN - JUHA VEHVILÄINEN, On Instability of a Keynesian Macro Model: Some Notes. 08.04.1986. 14 p.

No 208 PEKKA YLÄ-ANTTILA, Investment Structure, Productivity and Technical Change - Implications for Business Organizations and Management. 17.04.1986. 19 p.

No 209 JUHA AHTOLA, Consequences from Improper Use of Ordinary Least Squares Estimation with Time Series Data. 12.05.1986. 11 p.

No 210 TIMO AIRAKSINEN, Vertaileva analyysi pääomatulojen verotuksesta Suomessa ja Ruotsissa vuonna 1986. 29.05.1986. 36 s.

No 211 JUSSI RAUMOLIN, Kaivos- ja metallituotteiden maailmantalous. 18.06.1986. 40 s.

No 212 TARMO VALKONEN, Vakuutusyhtiöiden sijoitustoiminnan puitteet ja sijoitusten jakautuminen Suomessa vuosina 1962-1984. 19.06.1986. 68 s.

No 213 TIMO TERÄSVIRTA, GANG YI and GEORGE JUDGE, Model Selection, Smoothing and Parameter Estimation in Linear Models under Squared Error Loss. 17.07.1986. 21 p.

Elinkeinoelämän Tutkimuslaitoksen julkaisemat "Keskusteluaiheet" ovat raportteja alustavista tutkimustuloksista ja väliraportteja tekeillä olevista tutkimuksista. Tässä sarjassa julkaistuja monisteita on rajoitetusti saatavissa ETLAn kirjastosta tai ao. tutkijalta.

Papers in this series are reports on preliminary research results and on studies in progress; they can be obtained, on request, by the author's permission.