Econometric Theory, **24**, 2008, 338–376. Printed in the United States of America.
DOI: 10.1017/S0266466608080158

CAN ONE ESTIMATE THE UNCONDITIONAL DISTRIBUTION OF POST-MODEL-SELECTION ESTIMATORS?

HANNES LEEB
Yale University

BENEDIKT M. PÖTSCHER
University of Vienna

We consider the problem of estimating the unconditional distribution of a post-model-selection estimator. The notion of a post-model-selection estimator here refers to the combined procedure resulting from first selecting a model (e.g., by a model-selection criterion such as the Akaike information criterion [AIC] or by a hypothesis testing procedure) and then estimating the parameters in the selected model (e.g., by least squares or maximum likelihood), all based on the same data set. We show that it is impossible to estimate the unconditional distribution with reasonable accuracy even asymptotically. In particular, we show that no estimator for this distribution can be uniformly consistent (not even locally). This follows as a corollary to (local) minimax lower bounds on the performance of estimators for the distribution; performance is here measured by the probability that the estimation error exceeds a given threshold. These lower bounds are shown to approach $\frac{1}{2}$ or even 1 in large samples, depending on the situation considered. Similar impossibility results are also obtained for the distribution of linear functions (e.g., predictors) of the post-model-selection estimator.

1. INTRODUCTION AND OVERVIEW

In many statistical applications a data-based model-selection step precedes the final parameter estimation and inference stage. For example, the specification of the model (choice of functional form, choice of regressors, number of lags, etc.) is often based on the data. In contrast, the traditional theory of statistical inference is concerned with the properties of estimators and inference proce-

The research of the first author was supported by the Max Kade Foundation and by the Austrian National Science Foundation (FWF), Grant no. P13868-MAT. A preliminary draft of the material in this paper was written in 1999. Address correspondence to Benedikt M. Pötscher, Department of Statistics, University of Vienna, Universitätsstrasse 5, A-1010 Vienna, Austria; e-mail: Benedikt.Poetscher@univie.ac.at.

 © 2008 Cambridge University Press 0266-4666/08 $15.00

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

dures under the central assumption of an a priori given model. That is, it is assumed that the model is known to the researcher prior to the statistical analysis, except for the value of the true parameter vector. As a consequence, the actual statistical properties of estimators or inference procedures following a data-driven model-selection step are not described by the traditional theory, which assumes an a priori given model; in fact, they may differ substantially from the properties predicted by this theory; cf., e.g., Danilov and Magnus (2004), Dijkstra and Veldkamp (1988), Pötscher (1991, Sect. 3.3), and Rao and Wu (2001, Sect. 12). Ignoring the additional uncertainty originating from the data-driven model-selection step and (inappropriately) applying traditional theory can hence result in very misleading conclusions.

Investigations into the distributional properties of post-model-selection estimators, i.e., of estimators constructed after a data-driven model-selection step, are relatively few and of recent vintage. Sen (1979) obtained the unconditional large-sample limit distribution of a post-model-selection estimator in an independent and identically distributed (i.i.d.) maximum likelihood framework when selection is between two competing nested models. In Pötscher (1991) the asymptotic properties of a class of post-model-selection estimators (based on a sequence of hypothesis tests) were studied in a rather general setting covering nonlinear models, dependent processes, and more than two competing models. In that paper, the large-sample limit distribution of the post-model-selection estimator was derived, both unconditional and conditional on having chosen a correct model, not necessarily the minimal one. See also Pötscher and Novak (1998) for further discussion and a simulation study and Nickl (2003) for extensions. The finite-sample distribution of a post-model-selection estimator, both unconditional and conditional on having chosen a particular (possibly incorrect) model, was derived in Leeb and Pötscher (2003) in a normal linear regression framework; this paper also studied asymptotic approximations that are in a certain sense superior to the asymptotic distribution derived in Pötscher (1991). The distributions of corresponding linear predictors constructed after model selection were studied in Leeb (2005, 2006). Related work can also be found in Sen and Saleh (1987), Kabaila (1995), Pötscher (1995), Ahmed and Basu (2000), Kapetanios (2001), Hjort and Claeskens (2003), Dukić and Peña (2005), and Leeb and Pötscher (2005a). The latter paper provides a simple exposition of the problems of inference post model selection and may serve as an entry point to the present paper.

It transpires from the papers mentioned previously that the finite-sample distributions (and also the large-sample limit distributions) of post-model-selection estimators typically depend on the unknown model parameters, often in a complicated fashion. For inference purposes, e.g., for the construction of confidence sets, estimators of these distributions would be desirable. Consistent estimators of these distributions can typically be constructed quite easily, e.g., by suitably replacing unknown parameters in the large-sample limit distributions by estimators; cf. Section 2.2.1. However, the merits of such "plug-in"

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

estimators in small samples are questionable: It is known that the convergence of the finite-sample distributions to their large-sample limits is typically not uniform with respect to the underlying parameters (see Appendix B in the present paper and Leeb and Pötscher, 2003, Cor. 5.5), and there is no reason to believe that this nonuniformity will disappear when unknown parameters in the large-sample limit are replaced by estimators. This observation is the main motivation for the present paper to investigate in general the performance of estimators of the distribution of a post-model-selection estimator, where the estimators of the distribution are not necessarily "plug-in" estimators based on the limiting distribution. In particular, we ask whether estimators of the distribution function of post-model-selection estimators exist that do not suffer from the nonuniformity phenomenon mentioned earlier. As we show in this paper the answer in general is "No." We also show that these negative results extend to the problem of estimating the distribution of linear functions (e.g., linear predictors) of post-model-selection estimators. Similar negative results apply also to the estimation of the mean squared error or bias of post-model-selection estimators; cf. Remark 4.7.

To fix ideas consider for the moment the linear regression model

$$Y = V\chi + W\psi + u, \tag{1}$$

where V and W, respectively, represent $n \times k$ and $n \times l$ nonstochastic regressor matrices ($k \geq 1, l \geq 1$) and the $n \times 1$ disturbance vector u is normally distributed with mean zero and variance-covariance matrix $\sigma^2 I_n$. We also assume for the moment that $(V : W)'(V : W)/n$ converges to a nonsingular matrix as the sample size n goes to infinity and that $\lim_{n\to\infty} V'W/n \neq 0$ (for a discussion of the case where this limit is zero see Example 1 in Sect. 2.2.2). Now suppose that the vector χ represents the parameters of interest and the parameter vector ψ and the associated regressors in W have been entered into the model only to avoid possible misspecification. Suppose further that the necessity to include ψ or some of its components is then checked on the basis of the data; i.e., a model-selection procedure is used to determine which components of ψ are to be retained in the model, the inclusion of χ not being disputed. The selected model is then used to obtain the final (post-model-selection) estimator $\tilde{\chi}$ for χ. We are now interested in the unconditional finite-sample distribution of $\tilde{\chi}$ (appropriately scaled and centered). Denote this k-dimensional cumulative distribution function (cdf) by $G_{n,\theta,\sigma}(t)$. As indicated in the notation, this distribution function depends on the true parameters $\theta = (\chi', \psi')'$ and σ. For the sake of definiteness of discussion assume for the moment that the model-selection procedure used here is the particular "general-to-specific" procedure described at the beginning of Section 2; we comment on other model selection procedures, including the Akaike information criterion (AIC) and thresholding procedures, later.

As mentioned previously, it is not difficult to construct a consistent estimator of $G_{n,\theta,\sigma}(t)$ for any t, i.e., an estimator $\hat{G}_n(t)$ satisfying

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

$$P_{n,\theta,\sigma}(|\hat{G}_n(t) - G_{n,\theta,\sigma}(t)| > \delta) \xrightarrow{n\to\infty} 0 \tag{2}$$

for each $\delta > 0$ and each θ, σ; see Section 2.2.1. However, it follows from the results in Section 2.2.2 that *any* estimator satisfying (2), i.e., *any consistent* estimator of $G_{n,\theta,\sigma}(t)$, necessarily also satisfies

$$\sup_{\|\theta\|<R} P_{n,\theta,\sigma}(|\hat{G}_n(t) - G_{n,\theta,\sigma}(t)| > \delta) \xrightarrow{n\to\infty} 1 \tag{3}$$

for suitable positive constants R and δ that do not depend on the estimator. That is, although the probability in (2) converges to zero for every given θ by consistency, relation (3) shows that it does not do so uniformly in θ. It follows that $\hat{G}_n(t)$ can never be uniformly consistent (not even when restricting consideration to uniform consistency over all compact subsets of the parameter space). Hence, a large sample size does not guarantee a small estimation error with high probability when estimating the distribution function of a post-model-selection estimator. In this sense, reliably assessing the precision of post-model-selection estimators is an intrinsically hard problem. Apart from (3), we also provide minimax lower bounds for arbitrary (not necessarily consistent) estimators of the cumulative distribution function $G_{n,\theta,\sigma}(t)$. For example, we provide results that imply that

$$\liminf_{n\to\infty} \inf_{\hat{G}_n(t)} \sup_{\|\theta\|<R} P_{n,\theta,\sigma}(|\hat{G}_n(t) - G_{n,\theta,\sigma}(t)| > \delta) > 0 \tag{4}$$

holds for suitable positive constants R and δ, where the infimum extends over *all* estimators of $G_{n,\theta,\sigma}(t)$. The results in Section 2.2.2 in fact show that the balls $\|\theta\| < R$ in (3) and (4) can be replaced by suitable balls (not necessarily centered at the origin) shrinking at the rate $n^{-1/2}$. This shows that the nonuniformity phenomenon described in (3) and (4) is a local, rather than a global, phenomenon. In Section 2.2.2 we further show that the nonuniformity phenomenon expressed in (3) and (4) typically also arises when the parameter of interest is not χ but some other linear transformation of $\theta = (\chi', \psi')'$. As discussed in Remark 4.3, the results also hold for randomized estimators of the unconditional distribution function $G_{n,\theta,\sigma}(t)$. Hence no resampling procedure whatsoever can alleviate the problem. This explains the anecdotal evidence in the literature that resampling methods are often unsuccessful in approximating distributional properties of post-model-selection estimators (e.g., Dijkstra and Veldkamp, 1988; Freedman, Navidi, and Peters, 1988). See also the discussion on resampling in Section 6.

The results outlined previously are presented in Section 2.2 for the particular "general-to-specific" model-selection procedure described at the beginning of Section 2. Analogous results for a large class of model-selection procedures, including the AIC and thresholding procedures, are then given in Section 3, based on the results in Section 2.2. In fact, the nonuniformity phenomenon

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

expressed in (3) and (4) is not specific to the model-selection procedures discussed in Sections 2 and 3 of the present paper but will occur for most (if not all) model-selection procedures, including consistent ones; cf. Sections 5 and 6 for more discussion. Section 5 also shows that the results are—as is to be expected—by no means limited to the linear regression model.

We focus on the unconditional distributions of post-model-selection estimators in the present paper. One can, however, also envisage a situation where one is more interested in the conditional distribution given the outcome of the model-selection procedure. In line with the literature on conditional inference (see, e.g., Robinson, 1979; Lehmann and Casella, 1998, p. 421), one may argue that, given the outcome of the model-selection step, the relevant object of interest is the conditional rather than the unconditional distribution of the post-model-selection estimator. In this case similar results can be obtained and are reported in Leeb and Pötscher (2006b). We note that on a technical level the results in Leeb and Pötscher (2006b) and in the present paper require separate treatment.

The plan of the paper is as follows: Post-model-selection estimators based on a "general-to-specific" model-selection procedure are the subject of Section 2. After introducing the basic framework and some notation, such as the family of models M_p from which the "general-to-specific" model-selection procedure $\hat{p}$ selects, and also the post-model-selection estimator $\tilde{\theta}$, the unconditional cdf $G_{n,\theta,\sigma}(t)$ of (a linear function of) the post-model-selection estimator $\tilde{\theta}$ is discussed in Section 2.1. Consistent estimators of $G_{n,\theta,\sigma}(t)$ are given in Section 2.2.1. The main results of the paper are contained in Sections 2.2.2 and 3: In Section 2.2.2 we provide a detailed analysis of the nonuniformity phenomenon encountered in (3) and (4). In Section 3 the "impossibility" result from Section 2.2.2 is extended to a large class of model-selection procedures including the AIC and to selection from a nonnested collection of models. Some remarks are collected in Section 4, and Section 5 discusses extensions and the scope of the results of the paper. Conclusions are drawn in Section 6. All proofs and some auxiliary results are collected in Appendixes. Finally a word on notation: The euclidean norm is denoted by $\|\cdot\|$, and $\lambda_{\max}(E)$ denotes the largest eigenvalue of a symmetric matrix E. A prime denotes transposition of a matrix. For vectors x and y the relation $x \leq y$ ($x < y$, respectively) denotes $x_i \leq y_i$ ($x_i < y_i$, respectively) for all i. As usual, Φ denotes the standard normal distribution function.

2. RESULTS FOR POST-MODEL-SELECTION ESTIMATORS BASED ON A "GENERAL-TO-SPECIFIC" MODEL SELECTION PROCEDURE

Consider the linear regression model

$$Y = X\theta + u, \tag{5}$$

where X is a nonstochastic $n \times P$ matrix with $rank(X) = P$ and $u \sim N(0, \sigma^2 I_n)$, $\sigma^2 > 0$. Here n denotes the sample size, and we assume $n > P \geq 1$. In addi-

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

tion, we assume that $Q = \lim_{n\to\infty} X'X/n$ exists and is nonsingular. In this section we shall—in a similar way as in Pötscher (1991)—consider model selection from the collection of nested models $M_{\mathcal{O}} \subseteq M_{\mathcal{O}+1} \subseteq \cdots \subseteq M_P$, where $\mathcal{O}$ is specified by the user and where for $0 \le p \le P$ the model M_p is given by

$$M_p = \{(\theta_1, \ldots, \theta_P)' \in \mathbf{R}^P : \theta_{p+1} = \cdots = \theta_P = 0\}.$$

(In Section 3 also general nonnested families of models will be considered.) Clearly, the model M_p corresponds to the situation where only the first p regressors in (5) are included. For the most parsimonious model under consideration, i.e., for $M_{\mathcal{O}}$, we assume that $\mathcal{O}$ satisfies $0 \le \mathcal{O} < P$; if $\mathcal{O} > 0$, this model contains as free parameters only those components of the parameter vector θ that are not subject to model selection. (In the notation used in connection with (1) we then have $\chi = (\theta_1, \ldots, \theta_{\mathcal{O}})'$ and $\psi = (\theta_{\mathcal{O}+1}, \ldots, \theta_P)'$.) Furthermore, note that $M_0 = \{(0, \ldots, 0)'\}$ and that $M_P = \mathbf{R}^P$. We call M_p the regression model of order p.

The following notation will prove useful. For matrices B and C of the same row dimension, the column-wise concatenation of B and C is denoted by $(B : C)$. If D is an $m \times P$ matrix, let $D[p]$ denote the $m \times p$ matrix consisting of the first p columns of D. Similarly, let $D[\neg p]$ denote the $m \times (P - p)$ matrix consisting of the last $P - p$ columns of D. If x is a $P \times 1$ vector, we write in abuse of notation $x[p]$ and $x[\neg p]$ for $(x'[p])'$ and $(x'[\neg p])'$, respectively. (We shall use the preceding notation also in the "boundary" cases $p = 0$ and $p = P$. It will always be clear from the context how expressions containing symbols such as $D[0]$, $D[\neg P]$, $x[0]$, or $x[\neg P]$ are to be interpreted.) As usual, the ith component of a vector x is denoted by x_i, and the entry in the ith row and jth column of a matrix B is denoted by $B_{i,j}$.

The restricted least-squares estimator of θ under the restriction $\theta[\neg p] = 0$, i.e., under $\theta_{p+1} = \cdots = \theta_P = 0$, will be denoted by $\tilde{\theta}(p)$, $0 \le p \le P$ (in the case $p = P$ the restriction being void). Note that $\tilde{\theta}(p)$ is given by the $P \times 1$ vector

$$\tilde{\theta}(p) = \begin{pmatrix} (X[p]'X[p])^{-1}X[p]'Y \\ (0, \ldots, 0)' \end{pmatrix},$$

where the expressions $\tilde{\theta}(0)$ and $\tilde{\theta}(P)$, respectively, are to be interpreted as the zero vector in $\mathbf{R}^P$ and as the unrestricted least-squares estimator of θ. Given a parameter vector θ in $\mathbf{R}^P$, the order of θ (relative to the nested sequence of models M_p) is defined as

$$p_0(\theta) = \min\{p : 0 \le p \le P, \qquad \theta \in M_p\}.$$

Hence, if θ is the true parameter vector, a model M_p is a correct model if and only if $p \ge p_0(\theta)$. We stress that $p_0(\theta)$ is a property of a *single parameter* and hence needs to be distinguished from the notion of the order of the model M_p introduced earlier, which is a property of the *set of parameters* M_p.

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

A model-selection procedure is now nothing else than a data-driven (measurable) rule $\hat{p}$ that selects a value from $\{\mathcal{O}, \ldots, P\}$ and thus selects a model from the list of candidate models $M_{\mathcal{O}}, \ldots, M_P$. In this section we shall consider as an important leading case a "general-to-specific" model-selection procedure based on a sequence of hypothesis tests. (Results for a larger class of model-selection procedures, including the AIC, are provided in Sect. 3.) This procedure is given as follows: The sequence of hypotheses $H_0^p : p_0(\theta) < p$ is tested against the alternatives $H_1^p : p_0(\theta) = p$ in decreasing order starting at $p = P$. If, for some $p > \mathcal{O}$, H_0^p is the first hypothesis in the process that is rejected, we set $\hat{p} = p$. If no rejection occurs until even $H_0^{\mathcal{O}+1}$ is not rejected, we set $\hat{p} = \mathcal{O}$. Each hypothesis in this sequence is tested by a kind of t-test where the error variance is always estimated from the overall model (but see the discussion following Theorem 3.1 in Sect. 3 for other choices of estimators of the error variance). More formally, we have

$$\hat{p} = \max\{p : |T_p| \geq c_p, \qquad 0 \leq p \leq P\} \tag{6}$$

with $c_{\mathcal{O}} = 0$ to ensure a well-defined $\hat{p}$ in the range $\{\mathcal{O}, \mathcal{O}+1, \ldots, P\}$. For $\mathcal{O} < p \leq P$, the critical values c_p satisfy $0 < c_p < \infty$ and are independent of sample size (but see also Remark 4.2). The test statistics are given by

$$T_p = \frac{\sqrt{n}\tilde{\theta}_p(p)}{\hat{\sigma}\xi_{n,p}} \qquad (0 < p \leq P)$$

with the convention that $T_0 = 0$. Furthermore,

$$\xi_{n,p} = \left(\left[\left(\frac{X[p]'X[p]}{n} \right)^{-1} \right]_{p,p} \right)^{1/2} \qquad (0 < p \leq P)$$

denotes the nonnegative square root of the pth diagonal element of the matrix indicated, and $\hat{\sigma}^2$ is given by

$$\hat{\sigma}^2 = (n - P)^{-1}(Y - X\tilde{\theta}(P))'(Y - X\tilde{\theta}(P)).$$

Note that under the hypothesis H_0^p the statistic T_p is t-distributed with $n - P$ degrees of freedom for $0 < p \leq P$. It is also easy to see that the so-defined model-selection procedure $\hat{p}$ is conservative: The probability of selecting an incorrect model, i.e., the probability of the event $\{\hat{p} < p_0(\theta)\}$, converges to zero as the sample size increases. In contrast, the probability of the event $\{\hat{p} = p\}$, for p satisfying $\max\{p_0(\theta), \mathcal{O}\} \leq p \leq P$, converges to a positive limit; cf., e.g., Proposition 5.4 and equation (5.6) in Leeb (2006).

The post-model-selection estimator $\tilde{\theta}$ can now be defined as follows: On the event $\hat{p} = p$, $\tilde{\theta}$ is given by the restricted least-squares estimator $\tilde{\theta}(p)$, i.e.,

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

$$\tilde{\theta} = \sum_{p=\mathcal{O}}^{P} \tilde{\theta}(p)\mathbf{1}(\hat{p} = p), \tag{7}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function of the event shown in the argument.

2.1. The Distribution of the Post-Model-Selection Estimator

We now introduce the distribution function of a linear transformation of $\tilde{\theta}$ and summarize some of its properties that will be needed in the subsequent development. To this end, let A be a nonstochastic $k \times P$ matrix of rank k, $1 \leq k \leq P$, and consider the cdf

$$G_{n,\theta,\sigma}(t) = P_{n,\theta,\sigma}(\sqrt{n}A(\tilde{\theta} - \theta) \leq t) \qquad (t \in \mathbf{R}^k). \tag{8}$$

Here $P_{n,\theta,\sigma}(\cdot)$ denotes the probability measure corresponding to a sample of size n from (5).

Depending on the choice of the matrix A, several important scenarios are covered by (8): The cdf of $\sqrt{n}(\tilde{\theta} - \theta)$ is obtained by setting A equal to the $P \times P$ identity matrix I_P. In the case $\mathcal{O} > 0$, the cdf of those components of $\sqrt{n}(\tilde{\theta} - \theta)$ that correspond to the parameter of interest χ in (1) can be studied by setting A to the $\mathcal{O} \times P$ matrix $(I_{\mathcal{O}} : 0)$ as we then have $A\theta = (\theta_1, \ldots, \theta_{\mathcal{O}})' = \chi$. Finally, if $A \neq 0$ is a $1 \times P$ vector, we obtain the distribution of a linear predictor based on the post-model-selection estimator. See the examples at the end of Section 2.2.2 for more discussion.

The cdf $G_{n,\theta,\sigma}$ and its properties have been analyzed in detail in Leeb and Pötscher (2003) and Leeb (2006). To be able to access these results we need some further notation. Note that on the event $\hat{p} = p$ the expression $A(\tilde{\theta} - \theta)$ equals $A(\tilde{\theta}(p) - \theta)$ in view of (7). The expected value of the restricted least-squares estimator $\tilde{\theta}(p)$ will be denoted by $\eta_n(p)$ and is given by the $P \times 1$ vector

$$\eta_n(p) = \begin{pmatrix} \theta[p] + (X[p]'X[p])^{-1}X[p]'X[\neg p]\theta[\neg p] \\ (0, \ldots, 0)' \end{pmatrix} \tag{9}$$

with the conventions that $\eta_n(0) = (0, \ldots, 0)' \in \mathbf{R}^P$ and that $\eta_n(P) = \theta$. Furthermore, let $\Phi_{n,p}$ denote the cdf of $\sqrt{n}A(\tilde{\theta}(p) - \eta_n(p))$, i.e., the cdf of $\sqrt{n}A$ times the restricted least-squares estimator based on model M_p centered at its mean. Hence, $\Phi_{n,p}$ is the cdf of a k-variate Gaussian random vector with mean zero and variance-covariance matrix $\sigma^2 A[p](X[p]'X[p]/n)^{-1}A[p]'$ in the case $p > 0$, and it is the cdf of point mass at zero in $\mathbf{R}^k$ in the case $p = 0$. If $p > 0$ and if the matrix $A[p]$ has full row rank k, then $\Phi_{n,p}$ has a density with respect to Lebesgue measure, and we shall denote this density by $\phi_{n,p}$. We note that $\eta_n(p)$ depends on θ and that $\Phi_{n,p}$ depends on σ (in case $p > 0$), although these dependencies are not shown explicitly in the notation.

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

For $p > 0$ we introduce

$$b_{n,p} = C_n^{(p)'}(A[p](X[p]'X[p]/n)^{-1}A[p]')^{-} \tag{10}$$

and

$$\zeta_{n,p}^2 = \xi_{n,p}^2 - C_n^{(p)'}(A[p](X[p]'X[p]/n)^{-1}A[p]')^{-}C_n^{(p)} \tag{11}$$

with $\zeta_{n,p} \geq 0$. Here $C_n^{(p)} = A[p](X[p]'X[p]/n)^{-1}e_p$, where e_p denotes the pth standard basis vector in $\mathbf{R}^p$ and B^- denotes a generalized inverse of a matrix B. (Observe that $\zeta_{n,p}^2$ is invariant under the choice of the generalized inverse. The same is not necessarily true for $b_{n,p}$ but is true for $b_{n,p}z$ for all z in the column space of $A[p]$. Also note that (12), which follows, depends on $b_{n,p}$ only through $b_{n,p}z$ with z in the column space of $A[p]$.) We observe that the vector of covariances between $A\tilde{\theta}(p)$ and $\tilde{\theta}_p(p)$ is precisely given by $\sigma^2 n^{-1} C_n^{(p)}$ (and hence does *not* depend on θ). Furthermore, observe that $A\tilde{\theta}(p)$ and $\tilde{\theta}_p(p)$ are uncorrelated if and only if $\zeta_{n,p}^2 = \xi_{n,p}^2$ if and only if $b_{n,p}z = 0$ for all z in the column space of $A[p]$; cf. Lemma A.2 in Leeb (2005).

Finally, for a univariate Gaussian random variable $\mathfrak{N}$ with zero mean and variance s^2, $s \geq 0$, we write $\Delta_s(a,b)$ for $\mathbb{P}(|\mathfrak{N} - a| < b)$, $a \in \mathbf{R} \cup \{-\infty, \infty\}$, $b \in \mathbf{R}$. Note that $\Delta_s(\cdot,\cdot)$ is symmetric around zero in its first argument and that $\Delta_s(-\infty, b) = \Delta_s(\infty, b) = 0$ holds. In the case $s = 0$, $\mathfrak{N}$ is to be interpreted as being equal to zero; hence $a \mapsto \Delta_0(a,b)$ reduces to the indicator function of the interval $(-b, b)$.

We are now in a position to present the explicit formula for $G_{n,\theta,\sigma}(t)$ derived in Leeb (2006):

$$\begin{aligned} G_{n,\theta,\sigma}(t) &= \Phi_{n,\mathcal{O}}(t - \sqrt{n}A(\eta_n(\mathcal{O}) - \theta)) \\ &\quad \times \int_0^\infty \prod_{q=\mathcal{O}+1}^{P} \Delta_{\sigma\xi_{n,q}}(\sqrt{n}\eta_{n,q}(q), sc_q\sigma\xi_{n,q})h(s)\,ds \\ &\quad + \sum_{p=\mathcal{O}+1}^{P} \int_{z \leq t - \sqrt{n}A(\eta_n(p)-\theta)} \\ &\quad \times \Bigg[\int_0^\infty (1 - \Delta_{\sigma\zeta_{n,p}}(\sqrt{n}\eta_{n,p}(p) + b_{n,p}z, sc_p\sigma\xi_{n,p})) \\ &\qquad \times \prod_{q=p+1}^{P} \Delta_{\sigma\xi_{n,q}}(\sqrt{n}\eta_{n,q}(q), sc_q\sigma\xi_{n,q})h(s)\,ds \Bigg] \Phi_{n,p}(dz). \end{aligned} \tag{12}$$

In the preceding display, $\Phi_{n,p}(dz)$ denotes integration with respect to the measure induced by the normal cdf $\Phi_{n,p}$ on $\mathbf{R}^k$, and h denotes the density of $\hat{\sigma}/\sigma$; i.e., h is the density of $(n-P)^{-1/2}$ times the square root of a chi-square distributed random variable with $n - P$ degrees of freedom. The finite-sample distribution of the post-model-selection estimator given in (12) is in general

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

not normal; e.g., it can be bimodal; see Figure 2 in Leeb and Pötscher (2005a) and Figure 1 in Leeb (2006). (An exception where (12) is normal is the somewhat trivial case where $C_n^{(p)} = 0$, i.e., where $A\tilde{\theta}(p)$ and $\tilde{\theta}_p(p)$ are uncorrelated, for $p = \mathcal{O} + 1, \ldots, P$; for more discussion, see Leeb, 2006, Sect. 3.3.) We note for later use that $G_{n,\theta,\sigma}(t) = \sum_{p=\mathcal{O}}^{P} G_{n,\theta,\sigma}(t|p)\pi_{n,\theta,\sigma}(p)$ where $G_{n,\theta,\sigma}(t|p)$ represents the cdf of $\sqrt{n}A(\tilde{\theta} - \theta)$ conditional on the event $\{\hat{p} = p\}$ and where $\pi_{n,\theta,\sigma}(p)$ is the probability of this event under $P_{n,\theta,\sigma}$. Note that $\pi_{n,\theta,\sigma}(p)$ is always positive for $\mathcal{O} \le p \le P$; cf. Leeb (2006, Sect. 3.2).

To describe the large-sample limit of $G_{n,\theta,\sigma}$, some further notation is necessary. For p satisfying $0 < p \le P$, partition the matrix $Q = \lim_{n\to\infty} X'X/n$ as

$$Q = \begin{pmatrix} Q[p:p] & Q[p:\neg p] \\ Q[\neg p:p] & Q[\neg p:\neg p] \end{pmatrix},$$

where $Q[p:p]$ is a $p \times p$ matrix. Let $\Phi_{\infty,p}$ be the cdf of a k-variate Gaussian random vector with mean zero and variance-covariance matrix $\sigma^2 A[p]Q[p:p]^{-1}A[p]'$, $0 < p \le P$, and let $\Phi_{\infty,0}$ denote the cdf of point mass at zero in $\mathbf{R}^k$. Note that $\Phi_{\infty,p}$ has a Lebesgue density if $p > 0$ and the matrix $A[p]$ has full row rank k; in this case, we denote the Lebesgue density of $\Phi_{\infty,p}$ by $\phi_{\infty,p}$. Finally, for $p = 1, \ldots, P$, define

$$\begin{aligned} \xi^2_{\infty,p} &= (Q[p:p]^{-1})_{p,p}, \\ \zeta^2_{\infty,p} &= \xi^2_{\infty,p} - C_\infty^{(p)\prime}(A[p]Q[p:p]^{-1}A[p]')^{-}C_\infty^{(p)}, \\ b_{\infty,p} &= C_\infty^{(p)\prime}(A[p]Q[p:p]^{-1}A[p]')^{-}, \end{aligned} \tag{13}$$

where $C_\infty^{(p)} = A[p]Q[p:p]^{-1}e_p$, with e_p denoting the pth standard basis vector in $\mathbf{R}^p$; furthermore, take $\zeta_{\infty,p}$ and $\xi_{\infty,p}$ as the nonnegative square roots of $\zeta^2_{\infty,p}$ and $\xi^2_{\infty,p}$, respectively. As the notation suggests, $\Phi_{\infty,p}$ is the large-sample limit of $\Phi_{n,p}$, and $C_\infty^{(p)}$, $\xi^2_{\infty,p}$, and $\zeta^2_{\infty,p}$ are the limits of $C_n^{(p)}$, $\xi^2_{n,p}$, and $\zeta^2_{n,p}$, respectively; moreover, $b_{n,p}z$ converges to $b_{\infty,p}z$ for each z in the column space of $A[p]$. See Lemma A.2 in Leeb (2005).

The next result describes the large-sample limit of the cdf under local alternatives to θ and is taken from Leeb (2006, Cor. 5.6). Recall that the total variation distance between two cdfs G and G^* on $\mathbf{R}^k$ is defined as $\|G - G^*\|_{TV} = \sup_E |G(E) - G^*(E)|$, where the supremum is taken over all Borel sets E. Clearly, the relation $|G(t) - G^*(t)| \le \|G - G^*\|_{TV}$ holds for all $t \in \mathbf{R}^k$. Thus, if G and G^* are close with respect to the total variation distance, then $G(t)$ is close to $G^*(t)$, uniformly in t.

PROPOSITION 2.1. *Suppose $\theta \in \mathbf{R}^P$ and $\gamma \in \mathbf{R}^P$ and let $\sigma^{(n)}$ be a sequence of positive real numbers that converges to a (finite) limit $\sigma > 0$ as $n \to \infty$. Then the cdf $G_{n,\theta+\gamma/\sqrt{n},\sigma^{(n)}}$ converges to a limit $G_{\infty,\theta,\sigma,\gamma}$ in total variation, i.e.,*

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

$$\|G_{n,\theta+\gamma/\sqrt{n},\sigma^{(n)}} - G_{\infty,\theta,\sigma,\gamma}\|_{TV} \xrightarrow{n\to\infty} 0. \tag{14}$$

The large-sample limit cdf $G_{\infty,\theta,\sigma,\gamma}(t)$ is given by

$$\begin{aligned}
&\Phi_{\infty,p_*}(t-\beta^{(p_*)}) \prod_{q=p_*+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(\nu_q, c_q\sigma\xi_{\infty,q}) \\
&\quad + \sum_{p=p_*+1}^{P} \int_{z\le t-\beta^{(p)}} (1 - \Delta_{\sigma\zeta_{\infty,p}}(\nu_p + b_{\infty,p}z, c_p\sigma\xi_{\infty,p}))\Phi_{\infty,p}(dz) \\
&\quad \times \prod_{q=p+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(\nu_q, c_q\sigma\xi_{\infty,q}),
\end{aligned} \tag{15}$$

where $p_* = \max\{p_0(\theta), \mathcal{O}\}$. Here for $0 \le p \le P$

$$\beta^{(p)} = A\begin{pmatrix} Q[p:p]^{-1}Q[p:\neg p]\gamma[\neg p] \\ -\gamma[\neg p] \end{pmatrix}$$

with the convention that $\beta^{(p)} = -A\gamma$ if $p = 0$ and that $\beta^{(p)} = (0,\ldots,0)'$ if $p = P$. Furthermore, we have set $\nu_p = \gamma_p + (Q[p:p]^{-1}Q[p:\neg p]\gamma[\neg p])_p$ for $p > 0$. (Note that $\beta^{(p)} = \lim_{n\to\infty}\sqrt{n}A(\eta_n(p) - \theta - \gamma/\sqrt{n})$ for $p \ge p_0(\theta)$ and that $\nu_p = \lim_{n\to\infty}\sqrt{n}\,\eta_{n,p}(p)$ for $p > p_0(\theta)$. Here $\eta_n(p)$ is defined as in (9) but with $\theta + \gamma/\sqrt{n}$ replacing θ.)

If $p_* > 0$ and if the matrix $A[p_*]$ has full row rank k, then the Lebesgue density $\phi_{\infty,p}$ of $\Phi_{\infty,p}$ exists for all $p \ge p_*$, and hence the density of (15) exists and is given by

$$\begin{aligned}
&\phi_{\infty,p_*}(t-\beta^{(p_*)}) \prod_{q=p_*+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(\nu_q, c_q\sigma\xi_{\infty,q}) \\
&\quad + \sum_{p=p_*+1}^{P} (1 - \Delta_{\sigma\zeta_{\infty,p}}(\nu_p + b_{\infty,p}(t-\beta^{(p)}), c_p\sigma\xi_{\infty,p}))\phi_{\infty,p}(t-\beta^{(p)}) \\
&\quad \times \prod_{q=p+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(\nu_q, c_q\sigma\xi_{\infty,q}).
\end{aligned}$$

As with the finite-sample distribution, the limiting distribution of the post-model-selection estimator given in (15) is in general not normal. An exception is the case where $C_\infty^{(p)} = 0$ for $p > p_*$, in which case $G_{\infty,\theta,\sigma,\gamma}$ reduces to $\Phi_{\infty,P}$; see Remark A.1 in Appendix A. If $\gamma = 0$, we write $G_{\infty,\theta,\sigma}(t)$ as shorthand for $G_{\infty,\theta,\sigma,0}(t)$ in the following discussion.

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

2.2. Estimators of the Finite-Sample Distribution

For the purpose of inference after model selection the finite-sample distribution of the post-model-selection estimator is an object of particular interest. As we have seen, it depends on unknown parameters in a complicated manner, and hence one will have to be satisfied with estimators of this cdf. As we shall see, it is not difficult to construct consistent estimators of $G_{n,\theta,\sigma}(t)$. However, despite this consistency result, we shall find in Section 2.2.2 that *any* estimator of $G_{n,\theta,\sigma}(t)$ typically performs unsatisfactorily in that the estimation error cannot become small uniformly over (subsets of) the parameter space even as sample size goes to infinity. In particular, no uniformly consistent estimators exist, not even locally.

2.2.1. Consistent Estimators. We construct a consistent estimator of $G_{n,\theta,\sigma}(t)$ by commencing from the asymptotic distribution. Specializing to the case $\gamma = 0$ and $\sigma^{(n)} = \sigma$ in Proposition 2.1, the large-sample limit of $G_{n,\theta,\sigma}(t)$ is given by

$$G_{\infty,\theta,\sigma}(t) = \Phi_{\infty,p_*}(t) \prod_{q=p_*+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(0, c_q\sigma\xi_{\infty,q})$$
$$+ \sum_{p=p_*+1}^{P} \int_{z\le t} (1 - \Delta_{\sigma\zeta_{\infty,p}}(b_{\infty,p}z, c_p\sigma\xi_{\infty,p}))\Phi_{\infty,p}(dz)$$
$$\times \prod_{q=p+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(0, c_q\sigma\xi_{\infty,q}) \qquad \textbf{(16)}$$

with $p_* = \max\{p_0(\theta), \mathcal{O}\}$. Note that $G_{\infty,\theta,\sigma}(t)$ depends on θ only through p_*. Let $\hat{\Phi}_{n,p}$ denote the cdf of a k-variate Gaussian random vector with mean zero and variance-covariance matrix $\hat{\sigma}^2 A[p](X[p]'X[p]/n)^{-1}A[p]'$, $0 < p \le P$; we also adopt the convention that $\hat{\Phi}_{n,0}$ denotes the cdf of point mass at zero in $\mathbf{R}^k$. (We use the same convention for $\hat{\Phi}_{n,p}$ in the case $\hat{\sigma} = 0$, which is a probability zero event.) An estimator $\check{G}_n(t)$ of $G_{n,\theta,\sigma}(t)$ is now defined as follows: We first employ an auxiliary procedure $\bar{p}$ that consistently estimates $p_0(\theta)$ (e.g., $\bar{p}$ could be obtained from the Bayesian information criterion [BIC] or from a "general-to-specific" hypothesis testing procedure employing critical values that go to infinity but are $o(n^{1/2})$ as $n \to \infty$). The estimator $\check{G}_n(t)$ is now given by the expression in (16) but with p_*, σ, $b_{\infty,p}$, $\zeta_{\infty,p}$, $\xi_{\infty,p}$, and $\Phi_{\infty,p}$ replaced by $\max\{\bar{p}, \mathcal{O}\}$, $\hat{\sigma}$, $b_{n,p}$, $\zeta_{n,p}$, $\xi_{n,p}$, and $\hat{\Phi}_{n,p}$, respectively. A little reflection shows that $\check{G}_n$ is again a cdf. We have the following consistency results.

PROPOSITION 2.2. *The estimator $\check{G}_n$ is consistent (in the total variation distance) for $G_{n,\theta,\sigma}$ and $G_{\infty,\theta,\sigma}$. That is, for every $\delta > 0$*

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

$$P_{n,\theta,\sigma}(\|\check{G}_n(\cdot) - G_{n,\theta,\sigma}(\cdot)\|_{TV} > \delta) \xrightarrow{n\to\infty} 0, \tag{17}$$

$$P_{n,\theta,\sigma}(\|\check{G}_n(\cdot) - G_{\infty,\theta,\sigma}(\cdot)\|_{TV} > \delta) \xrightarrow{n\to\infty} 0 \tag{18}$$

for all $\theta \in \mathbf{R}^P$ *and all* $\sigma > 0$.

Although the estimator constructed previously on the basis of the formula for $G_{\infty,\theta,\sigma}$ is consistent, it can be expected to perform poorly in finite samples because convergence of $G_{n,\theta,\sigma}$ to $G_{\infty,\theta,\sigma}$ is typically not uniform in θ (cf. Appendix B) and because in the case where the true θ is "close" to $M_{p_0(\theta)-1}$ the auxiliary decision procedure $\bar{p}$ (although being consistent for $p_0(\theta)$) will then have difficulties making the correct decision in finite samples. In the next section we show that this poor performance is not particular to the estimator $\check{G}_n$ constructed earlier but is a genuine feature of the estimation problem under consideration.

2.2.2. Performance Limits and Impossibility Results. We now provide lower bounds for the performance of estimators of the cdf $G_{n,\theta,\sigma}(t)$ of the post-model-selection estimator $A\tilde{\theta}$; i.e., we give lower bounds on the worst-case probability that the estimation error exceeds a certain threshold. These lower bounds are large, being 1 or $\frac{1}{2}$, depending on the situation considered; furthermore, they remain lower bounds even if one restricts attention only to certain subsets of the parameter space that shrink at the rate $n^{-1/2}$. In this sense the "impossibility" results are of a local nature. In particular, the lower bounds imply that no uniformly consistent estimator of the cdf $G_{n,\theta,\sigma}(t)$ exists, not even locally.

In the following discussion the asymptotic "correlation" between $A\tilde{\theta}(p)$ and $\tilde{\theta}_p(p)$ as measured by $C_\infty^{(p)} = \lim_{n\to\infty} C_n^{(p)}$ will play an important rôle. (Recall that $\tilde{\theta}(p)$ denotes the least-squares estimator of θ based on model M_p and that $A\theta$ is the parameter vector of interest. Furthermore, the vector of covariances between $A\tilde{\theta}(p)$ and $\tilde{\theta}_p(p)$ is given by $\sigma^2 n^{-1} C_n^{(p)}$ with $C_n^{(p)} = A[p](X[p]'X[p]/n)^{-1}e_p$.) Note that $C_\infty^{(p)}$ equals $A[p]Q[p:p]^{-1}e_p$ and hence does *not* depend on the unknown parameters θ or σ. In the important special case discussed in the Introduction, cf. (1), the matrix A equals the $\mathcal{O} \times P$ matrix $(I_\mathcal{O}:0)$, and the condition $C_\infty^{(p)} \neq 0$ reduces to the condition that the regressor corresponding to the pth column of $(V:W)$ is asymptotically correlated with at least one of the regressors corresponding to the columns of V. See Example 1, which follows, for more discussion.

In the result to follow we shall consider performance limits for estimators of $G_{n,\theta,\sigma}(t)$ at a *fixed* value of the argument t. An estimator of $G_{n,\theta,\sigma}(t)$ is now nothing else than a real-valued random variable $\Gamma_n = \Gamma_n(Y, X)$. For mnemonic reasons we shall, however, use the symbol $\hat{G}_n(t)$ instead of Γ_n to denote an arbitrary estimator of $G_{n,\theta,\sigma}(t)$. This notation should not be taken as implying that the estimator is obtained by evaluating an estimated cdf at the argument t or that it is a priori constrained to lie between zero and one. We shall use this

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

notational convention mutatis mutandis also in subsequent sections. Regarding the nonuniformity phenomenon, we then have a dichotomy that is described in the following two results.

THEOREM 2.3. *Suppose that $A\tilde{\theta}(q)$ and $\tilde{\theta}_q(q)$ are asymptotically correlated, i.e., $C_\infty^{(q)} \neq 0$, for some q satisfying $\mathcal{O} < q \leq P$ and let q^* denote the largest q with this property. Then the following holds for every $\theta \in M_{q^*-1}$, every σ, $0 < \sigma < \infty$, and every $t \in \mathbf{R}^k$: There exist $\delta_0 > 0$ and ρ_0, $0 < \rho_0 < \infty$, such that any estimator $\hat{G}_n(t)$ of $G_{n,\theta,\sigma}(t)$ satisfying*

$$P_{n,\theta,\sigma}(|\hat{G}_n(t) - G_{n,\theta,\sigma}(t)| > \delta) \xrightarrow{n\to\infty} 0 \tag{19}$$

for each $\delta > 0$ (in particular, every estimator that is consistent) also satisfies

$$\sup_{\substack{\vartheta \in M_{q^*} \\ \|\vartheta-\theta\| < \rho_0/\sqrt{n}}} P_{n,\vartheta,\sigma}(|\hat{G}_n(t) - G_{n,\vartheta,\sigma}(t)| > \delta_0) \xrightarrow{n\to\infty} 1. \tag{20}$$

The constants δ_0 and ρ_0 may be chosen in such a way that they depend only on t, Q, A, σ, and the critical values c_p for $\mathcal{O} < p \leq P$. Moreover,

$$\liminf_{n\to\infty} \inf_{\hat{G}_n(t)} \sup_{\substack{\vartheta \in M_{q^*} \\ \|\vartheta-\theta\| < \rho_0/\sqrt{n}}} P_{n,\vartheta,\sigma}(|\hat{G}_n(t) - G_{n,\vartheta,\sigma}(t)| > \delta_0) > 0 \tag{21}$$

and

$$\sup_{\delta>0} \liminf_{n\to\infty} \inf_{\hat{G}_n(t)} \sup_{\substack{\vartheta \in M_{q^*} \\ \|\vartheta-\theta\| < \rho_0/\sqrt{n}}} P_{n,\vartheta,\sigma}(|\hat{G}_n(t) - G_{n,\vartheta,\sigma}(t)| > \delta) \geq \tfrac{1}{2}, \tag{22}$$

where the infima in (21) and (22) extend over all *estimators $\hat{G}_n(t)$ of $G_{n,\theta,\sigma}(t)$.*

Remark 2.1. Assume that the conditions of the preceding theorem are satisfied. Suppose further that $p_\odot$, $\mathcal{O} \leq p_\odot < q^*$, is such that either $p_\odot > 0$ and some row of $A[\,p_\odot]$ equals zero or such that $p_\odot = 0$. Then there exist $\delta_0 > 0$ and $0 < \rho_0 < \infty$ such that the left-hand side of (21) is not less than $\frac{1}{2}$ for each $\theta \in M_{p_\odot}$.

Theorem 2.3 a fortiori implies a corresponding "impossibility" result for estimation of the function $G_{n,\theta,\sigma}(\cdot)$ when the estimation error is measured in the total variation distance or the sup-norm; cf. also Section 5.

It remains to consider the—quite exceptional—case where the assumption of Theorem 2.3 is not satisfied, i.e., where $C_\infty^{(q)} = 0$, for all q in the range $\mathcal{O} < q \leq P$. Under this "uncorrelatedness" condition it is indeed possible to construct an estimator of $G_{n,\theta,\sigma}$ that is uniformly consistent: It is not difficult to

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

see that the asymptotic distribution corresponding to $G_{n,\theta,\sigma}$ reduces to $\Phi_{\infty,P}$ under this "uncorrelatedness" condition. Furthermore, the second half of Proposition B.1 in Appendix B shows that then the convergence of $G_{n,\theta,\sigma}$ to its large-sample limit is uniform with respect to θ, suggesting $\hat{\Phi}_{n,P}$, an estimated version of $\Phi_{\infty,P}$, as an estimator for $G_{n,\theta,\sigma}$.

PROPOSITION 2.4. *Suppose that $A\tilde{\theta}(q)$ and $\tilde{\theta}_q(q)$ are asymptotically uncorrelated, i.e., $C_\infty^{(q)} = 0$, for all q satisfying $\mathcal{O} < q \leq P$. Then*

$$\sup_{\theta \in \mathbf{R}^P} \sup_{\substack{\sigma \in \mathbf{R} \\ \sigma_* \leq \sigma \leq \sigma^*}} P_{n,\theta,\sigma}(\|\hat{\Phi}_{n,P} - G_{n,\theta,\sigma}\|_{TV} > \delta) \xrightarrow{n\to\infty} 0 \tag{23}$$

holds for each $\delta > 0$ and for any constants σ_* and σ^* satisfying $0 < \sigma_* \leq \sigma^* < \infty$.

Inspection of the proof of Proposition 2.4 in Appendix C shows that (23) continues to hold if the estimator $\hat{\Phi}_{n,P}$ is replaced by any of the estimators $\hat{\Phi}_{n,p}$ for $\mathcal{O} \leq p \leq P$. We also note that in the case $\mathcal{O} = 0$ the assumption of Proposition 2.4 is never satisfied in view of Proposition 4.4 in Leeb and Pötscher (2006b), and hence Theorem 2.3 always applies in that case. Another consequence of Proposition 4.4 in Leeb and Pötscher (2006b) is that—under the "uncorrelatedness" assumption of Proposition 2.4—the restricted least-squares estimators $A\tilde{\theta}(q)$ for $q \geq \mathcal{O}$ perform asymptotically as well as the unrestricted estimator $A\tilde{\theta}(P)$; this clearly shows that the case covered by Proposition 2.4 is highly exceptional.

In summary we see that it is typically impossible to construct an estimator of $G_{n,\theta,\sigma}(t)$ that performs reasonably well even asymptotically. Whenever Theorem 2.3 applies, any estimator of $G_{n,\theta,\sigma}(t)$ suffers from a nonuniformity defect that is caused by parameters belonging to shrinking "tubes" surrounding M_{q^*-1}. For the sake of completeness, we remark that outside a "tube" of fixed positive radius that surrounds M_{q^*-1} the nonuniformity need not be present: Let q^* be as in Theorem 2.3 and define the set U as $U = \{\theta \in \mathbf{R}^P : |\theta_{q^*}| \geq r\}$ for some fixed $r > 0$. Then $\hat{\Phi}_{n,P}(t)$ is an estimator of $G_{n,\theta,\sigma}(t)$ that is uniformly consistent over $\theta \in U$; more generally, it can be shown that then the relation (23) holds if the supremum over θ on the left-hand side is restricted to $\theta \in U$.

We conclude this section by illustrating the preceding results with some important examples.

Example 1 ***(The distribution of $\tilde{\chi}$)***

Consider the model given in (1) with χ representing the parameter of interest. Using the general notation of Section 2, this corresponds to the case $A\theta = (\theta_1, \ldots, \theta_{\mathcal{O}})' = \chi$ with A representing the $\mathcal{O} \times P$ matrix $(I_{\mathcal{O}} : 0)$. Here $k = \mathcal{O} > 0$. The cdf $G_{n,\theta,\sigma}$ then represents the cdf of $\sqrt{n}(\tilde{\chi} - \chi)$. Assume first

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

that $\lim_{n\to\infty} V'W/n \neq 0$. Then $C_\infty^{(q)} \neq 0$ holds for some $q > \mathcal{O}$. Consequently, the "impossibility" results for the estimation of $G_{n,\theta,\sigma}$ given in Theorem 2.3 always apply. Next assume that $\lim_{n\to\infty} V'W/n = 0$. Then $C_\infty^{(q)} = 0$ for every $q > \mathcal{O}$. In this case Proposition 2.4 applies, and a uniformly consistent estimator of $G_{n,\theta,\sigma}$ indeed exists. Summarizing we note that any estimator of $G_{n,\theta,\sigma}$ suffers from the nonuniformity phenomenon except in the special case where the columns of V and W are asymptotically orthogonal in the sense that $\lim_{n\to\infty} V'W/n = 0$. But this is precisely the situation where inclusion or exclusion of the regressors in W has no effect on the distribution of the estimator $\tilde{\chi}$ asymptotically; hence it is not surprising that also the model-selection procedure does not have an effect on the estimation of the cdf of the post-model-selection estimator $\tilde{\chi}$. This observation may tempt one to enforce orthogonality between the columns of V and W by either replacing the columns of V by their residuals from the projection on the column space of W or vice versa. However, this is not helpful for the following reasons: In the first case one then in fact avoids model selection as all the restricted least-squares estimators for χ under consideration (and hence also the post-model-selection estimator $\tilde{\chi}$) in the reparameterized model coincide with the unrestricted least-squares estimator. In the second case the coefficients of the columns of V in the reparameterized model no longer coincide with the parameter of interest χ (and again are estimated by one and the same estimator regardless of inclusion/exclusion of columns of the transformed W matrix).

Example 2 ***(The distribution of $\tilde{\theta}$)***

For A equal to I_P, the cdf $G_{n,\theta,\sigma}$ is the cdf of $\sqrt{n}(\tilde{\theta} - \theta)$. Here, $A\tilde{\theta}(q)$ reduces to $\tilde{\theta}(q)$, and hence $A\tilde{\theta}(q)$ and $\tilde{\theta}_q(q)$ are perfectly correlated for every $q > \mathcal{O}$. Consequently, the "impossibility" result for estimation of $G_{n,\theta,\sigma}$ given in Theorem 2.3 applies. (In fact, the slightly stronger result mentioned in Remark 2.1 always applies here.) We therefore see that estimation of the distribution of the post-model-selection estimator of the entire parameter vector is always plagued by the nonuniformity phenomenon.

Example 3 ***(The distribution of a linear predictor)***

Suppose $A \neq 0$ is a $1 \times P$ vector and one is interested in estimating the cdf $G_{n,\theta,\sigma}$ of the linear predictor $A\tilde{\theta}$. Then Theorem 2.3 and the discussion following Proposition 2.4 show that the nonuniformity phenomenon always arises in this estimation problem in the case $\mathcal{O} = 0$. In the case $\mathcal{O} > 0$, the nonuniformity problem is generically also present, except in the degenerate case where $C_\infty^{(q)} = 0$, for all q satisfying $\mathcal{O} < q \leq P$ (in which case Proposition 4.4 in Leeb and Pötscher, 2006b, shows that the least-squares predictors from all models M_p, $\mathcal{O} \leq p \leq P$, perform asymptotically equally well).

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

3. EXTENSIONS TO OTHER MODEL-SELECTION PROCEDURES INCLUDING AIC

In this section we show that the "impossibility" result obtained in the previous section for a "general-to-specific" model-selection procedure carries over to a large class of model-selection procedures, including the widely used AIC. Again consider the linear regression model (5) with the same assumptions on the regressors and the errors as in Section 2. Let $\{0,1\}^P$ denote the set of all 0–1 sequences of length P. For each $\mathfrak{r} \in \{0,1\}^P$ let $M_{\mathfrak{r}}$ denote the set $\{\theta \in \mathbf{R}^P : \theta_i(1 - \mathfrak{r}_i) = 0$ for $1 \le i \le P\}$ where $\mathfrak{r}_i$ represents the ith component of $\mathfrak{r}$; i.e., $M_{\mathfrak{r}}$ describes a linear submodel with those parameters θ_i restricted to zero for which $\mathfrak{r}_i = 0$. Now let $\mathfrak{R}$ be a user-supplied subset of $\{0,1\}^P$. We consider model selection procedures that select from the set $\mathfrak{R}$ or equivalently from the set of models $\{M_{\mathfrak{r}} : \mathfrak{r} \in \mathfrak{R}\}$. Note that there is now no assumption that the candidate models are nested (e.g., if $\mathfrak{R} = \{0,1\}^P$ all possible submodels are candidates for selection). Also cases where the inclusion of a subset of regressors is undisputed on a priori grounds are obviously covered by this framework upon suitable choice of $\mathfrak{R}$.

We shall assume throughout this section that $\mathfrak{R}$ contains $\mathfrak{r}_{full} = (1, \ldots, 1)$ and also at least one element $\mathfrak{r}_*$ satisfying $|\mathfrak{r}_*| = P - 1$, where $|\mathfrak{r}_*|$ represents the number of nonzero coordinates of $\mathfrak{r}_*$. Let $\hat{\mathfrak{r}}$ be an arbitrary model-selection procedure; i.e., $\hat{\mathfrak{r}} = \hat{\mathfrak{r}}(Y, X)$ is a random variable taking its values in $\mathfrak{R}$. We furthermore assume throughout this section that the model-selection procedure $\hat{\mathfrak{r}}$ satisfies the following mild condition: For every $\mathfrak{r}_* \in \mathfrak{R}$ with $|\mathfrak{r}_*| = P - 1$ there exists a positive finite constant c (possibly depending on $\mathfrak{r}_*$) such that for every $\theta \in M_{\mathfrak{r}_*}$ that has exactly $P - 1$ nonzero coordinates

$$\lim_{n\to\infty} P_{n,\theta,\sigma}(\{\hat{\mathfrak{r}} = \mathfrak{r}_{full}\} \blacktriangle \{|T_{\mathfrak{r}_*}| \ge c\}) = \lim_{n\to\infty} P_{n,\theta,\sigma}(\{\hat{\mathfrak{r}} = \mathfrak{r}_*\} \blacktriangle \{|T_{\mathfrak{r}_*}| < c\}) = 0 \tag{24}$$

holds for every $0 < \sigma < \infty$. Here $\blacktriangle$ denotes the symmetric difference operator and $T_{\mathfrak{r}_*}$ represents the usual t-statistic for testing the hypothesis $\theta_{i(\mathfrak{r}_*)} = 0$ in the full model, where $i(\mathfrak{r}_*)$ denotes the index of the unique coordinate of $\mathfrak{r}_*$ that equals zero.

The preceding condition is quite natural for the following reason: For $\theta \in M_{\mathfrak{r}_*}$ with exactly $P - 1$ nonzero coordinates, every reasonable model-selection procedure will—with probability approaching unity—decide only between $M_{\mathfrak{r}_*}$ and $M_{\mathfrak{r}_{full}}$; it is then quite natural that this decision will be based (at least asymptotically) on the likelihood ratio between these two models, which in turn boils down to the t-statistic. As will be shown subsequently, condition (24) holds in particular for AIC-like procedures.

Let A be a nonstochastic $k \times P$ matrix of full row rank k, $1 \le k \le P$, as in Section 2.1. We then consider the cdf

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

$$K_{n,\theta,\sigma}(t) = P_{n,\theta,\sigma}(\sqrt{n}A(\bar{\theta} - \theta) \le t) \qquad (t \in \mathbf{R}^k) \tag{25}$$

of a linear transformation of the post-model-selection estimator $\bar{\theta}$ obtained from the model-selection procedure $\hat{\mathfrak{r}}$, i.e.,

$$\bar{\theta} = \sum_{\mathfrak{r} \in \mathfrak{R}} \tilde{\theta}(\mathfrak{r}) \mathbf{1}(\hat{\mathfrak{r}} = \mathfrak{r}),$$

where the $P \times 1$ vector $\tilde{\theta}(\mathfrak{r})$ represents the restricted least-squares estimator obtained from model $M_{\mathfrak{r}}$, with the convention that $\tilde{\theta}(\mathfrak{r}) = 0 \in \mathbf{R}^P$ in the case $\mathfrak{r} = (0, \ldots, 0)$. We then obtain the following result for estimation of $K_{n,\theta,\sigma}(t)$ at a *fixed* value of the argument t that parallels the corresponding "impossibility" result in Theorem 2.3.

THEOREM 3.1. *Let $\mathfrak{r}_* \in \mathfrak{R}$ satisfy $|\mathfrak{r}_*| = P - 1$ and let $i(\mathfrak{r}_*)$ denote the index of the unique coordinate of $\mathfrak{r}_*$ that equals zero; furthermore, let c be the constant in (24) corresponding to $\mathfrak{r}_*$. Suppose that $A\tilde{\theta}(\mathfrak{r}_{full})$ and $\tilde{\theta}_{i(\mathfrak{r}_*)}(\mathfrak{r}_{full})$ are asymptotically correlated, i.e., $AQ^{-1}e_{i(\mathfrak{r}_*)} \neq 0$, where $e_{i(\mathfrak{r}_*)}$ denotes the $i(\mathfrak{r}_*)$th standard basis vector in $\mathbf{R}^P$. Then for every $\theta \in M_{\mathfrak{r}_*}$ that has exactly $P - 1$ nonzero coordinates, for every σ, $0 < \sigma < \infty$, and for every $t \in \mathbf{R}^k$ the following holds: There exist $\delta_0 > 0$ and ρ_0, $0 < \rho_0 < \infty$, such that any estimator $\hat{K}_n(t)$ of $K_{n,\theta,\sigma}(t)$ satisfying*

$$P_{n,\theta,\sigma}(|\hat{K}_n(t) - K_{n,\theta,\sigma}(t)| > \delta) \xrightarrow{n\to\infty} 0 \tag{26}$$

for each $\delta > 0$ (in particular, every estimator that is consistent) also satisfies

$$\sup_{\substack{\vartheta \in \mathbf{R}^P \\ \|\vartheta - \theta\| < \rho_0/\sqrt{n}}} P_{n,\vartheta,\sigma}(|\hat{K}_n(t) - K_{n,\vartheta,\sigma}(t)| > \delta_0) \xrightarrow{n\to\infty} 1. \tag{27}$$

The constants δ_0 and ρ_0 may be chosen in such a way that they depend only on t, Q, A, σ, and c. Moreover,

$$\liminf_{n\to\infty} \inf_{\hat{K}_n(t)} \sup_{\substack{\vartheta \in \mathbf{R}^P \\ \|\vartheta - \theta\| < \rho_0/\sqrt{n}}} P_{n,\vartheta,\sigma}(|\hat{K}_n(t) - K_{n,\vartheta,\sigma}(t)| > \delta_0) > 0 \tag{28}$$

and

$$\sup_{\delta > 0} \liminf_{n\to\infty} \inf_{\hat{K}_n(t)} \sup_{\substack{\vartheta \in \mathbf{R}^P \\ \|\vartheta - \theta\| < \rho_0/\sqrt{n}}} P_{n,\vartheta,\sigma}(|\hat{K}_n(t) - K_{n,\vartheta,\sigma}(t)| > \delta) \ge \frac{1}{2} \tag{29}$$

hold, where the infima in (28) and (29) extend over all *estimators $\hat{K}_n(t)$ of $K_{n,\theta,\sigma}(t)$.*

The basic condition (24) on the model-selection procedure employed in the preceding result will certainly hold for any hypothesis testing procedure that

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

(i) asymptotically selects only correct models, (ii) employs a likelihood ratio test (or an asymptotically equivalent test) for testing $M_{\mathfrak{r}_{full}}$ versus smaller models (at least versus the models $M_{\mathfrak{r}_*}$ with $\mathfrak{r}_*$ as in condition (24)), and (iii) uses a critical value for the likelihood ratio test that converges to a finite positive constant. In particular, this applies to usual thresholding procedures and also to a variant of the "general-to-specific" procedure discussed in Section 2 where the error variance in the construction of the test statistic for hypothesis H_0^p is estimated from the fitted model M_p rather than from the overall model. We next verify condition (24) for AIC-like procedures. Let $RSS(\mathfrak{r})$ denote the residual sum of squares from the regression employing model $M_{\mathfrak{r}}$ and set

$$IC(\mathfrak{r}) = \log(RSS(\mathfrak{r})) + |\mathfrak{r}|\,\Upsilon_n/n, \tag{30}$$

where $\Upsilon_n \geq 0$ denotes a sequence of real numbers satisfying $\lim_{n\to\infty} \Upsilon_n = \Upsilon$ and Υ is a positive real number. Of course, $IC(\mathfrak{r}) = AIC(\mathfrak{r})$ if $\Upsilon_n = 2$. The model-selection procedure $\hat{\mathfrak{r}}_{IC}$ is then defined as a minimizer (more precisely, as a measurable selection from the set of minimizers) of $IC(\mathfrak{r})$ over $\mathfrak{R}$. It is well known that the probability that $\hat{\mathfrak{r}}_{IC}$ selects an incorrect model converges to zero. Hence, elementary calculations show that condition (24) is satisfied for $c = \Upsilon^{1/2}$.

The analysis of post-model-selection estimators based on AIC-like model-selection procedures given in this section proceeded by bringing this case under the umbrella of the results obtained in Section 2. Verification of condition (24) is the key that enables this approach. A complete analysis of post-model-selection estimators based on AIC-like model-selection procedures, similar to the analysis in Section 2 for the "general-to-specific" model-selection procedure, is certainly possible but requires a direct and detailed analysis of the distribution of this post-model-selection estimator. (Even the mild condition that $\mathfrak{R}$ contains $\mathfrak{r}_{full}$ and also at least one element $\mathfrak{r}_*$ satisfying $|\mathfrak{r}_*| = P - 1$ can then be relaxed in such an analysis.) We furthermore note that in the special case where $\mathfrak{R} = \{\mathfrak{r}_{full}, \mathfrak{r}_*\}$ and an AIC-like model-selection procedure as in (30) is used, the results in the preceding theorem in fact hold for all $\theta \in M_{\mathfrak{r}_*}$.

4. REMARKS AND EXTENSIONS

Remark 4.1. Although not emphasized in the notation, all results in the paper also hold if the elements of the design matrix X depend on sample size. Furthermore, all results are expressed solely in terms of the distributions $P_{n,\theta,\sigma}(\cdot)$ of Y, and hence they also apply if the elements of Y depend on sample size, including the case where the random vectors Y are defined on different probability spaces for different sample sizes.

Remark 4.2. The model-selection procedure considered in Section 2 is based on a sequence of tests that use critical values c_p that do not depend on sample size and satisfy $0 < c_p < \infty$ for $\mathcal{O} < p \leq P$. If these critical values are allowed

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

to depend on sample size such that they now satisfy $c_{n,p} \to c_{\infty,p}$ as $n \to \infty$ with $0 < c_{\infty,p} < \infty$ for $\mathcal{O} < p \leq P$, the results in Leeb and Pötscher (2003) and in Leeb (2005, 2006) continue to hold; see Remark 6.2(i) in Leeb and Pötscher (2003) and Remark 6.1(ii) in Leeb (2005). As a consequence, the results in the present paper can also be extended to this case quite easily.

Remark 4.3. The "impossibility" results given in Theorems 2.3 and 3.1 (and also the variants thereof discussed in Remarks 4.4–4.7, which follow) also hold for the class of all randomized estimators (with $P^*_{n,\theta,\sigma}$ replacing $P_{n,\theta,\sigma}$ in those results, where $P^*_{n,\theta,\sigma}$ denotes the distribution of the randomized sample). This follows immediately from Lemma 3.6 and the attending discussion in Leeb and Pötscher (2006a).

Remark 4.4.

(i) Let $\psi_{n,\theta,\sigma}$ denote the expectation of $\tilde{\theta}$ under $P_{n,\theta,\sigma}$ and consider the cdf $H_{n,\theta,\sigma}(t) = P_{n,\theta,\sigma}(\sqrt{n}A(\tilde{\theta} - \psi_{n,\theta,\sigma}) \leq t)$. Results for the cdf $H_{n,\theta,\sigma}$ quite similar to the results for $G_{n,\theta,\sigma}$ obtained in the present paper can be established. A similar remark applies to the post-model-selection estimator $\bar{\theta}$ considered in Section 3.
(ii) In Leeb (2006) the cdf $G^*_{n,\theta,\sigma}$ also is analyzed, which corresponds to a (typically infeasible) model-selection procedure that makes use of knowledge of σ. Results completely analogous to the ones in the present paper can also be obtained for this cdf.

Remark 4.5. Results similar to the ones in Section 2.2.2 can also be obtained for estimation of the asymptotic cdf $G_{\infty,\theta,\sigma}(t)$ (or of the asymptotic cdfs corresponding to the variants discussed in the previous remark). Because these results are of limited interest, we omit them. In particular, note that an "impossibility" result for estimation of $G_{\infty,\theta,\sigma}(t)$ per se does *not* imply a corresponding "impossibility" result for estimation of $G_{n,\theta,\sigma}(t)$, because $G_{n,\theta,\sigma}(t)$ in general does not converge uniformly to $G_{\infty,\theta,\sigma}(t)$ over the relevant subsets in the parameter space; cf. Appendix B. (An analogous remark applies to the model-selection procedures considered in Section 3.)

Remark 4.6. Let $\pi_{n,\theta,\sigma}(p)$ denote the model-selection probability $P_{n,\theta,\sigma}(\hat{p} = p)$, $\mathcal{O} \leq p \leq P$ corresponding to the model-selection procedure discussed in Section 2. The finite-sample properties and the large-sample limit behavior of these quantities are thoroughly analyzed in Leeb (2006); cf. also Leeb and Pötscher (2003). For these model-selection probabilities the following results can be established, which we discuss here only briefly:

(i) The model-selection probabilities $\pi_{n,\theta,\sigma}(p)$ converge to well-defined large-sample limits that we denote by $\pi_{\infty,\theta,\sigma}(p)$. In a similar way as in Proposition B.1 in Appendix B, the convergence of $\pi_{n,\theta,\sigma}(p)$ to $\pi_{\infty,\theta,\sigma}(p)$

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

is nonuniform with respect to θ. (For the case $\mathcal{O} = 0$, this phenomenon is described in Corollary 5.6 of Leeb and Pötscher, 2003.)

(ii) The model-selection probabilities $\pi_{n,\theta,\sigma}(p)$ can be estimated consistently. However, uniformly consistent estimation is again not possible. A similar remark applies to the large-sample limits $\pi_{\infty,\theta,\sigma}(p)$.

Remark 4.7. "Impossibility" results similar to the ones given in Theorems 2.3 and 3.1 for the cdf can also be obtained for other characteristics of the distribution of a linear function of a post-model-selection estimator such as the mean squared error or the bias of $\sqrt{n}A\tilde{\theta}$.

5. ON THE SCOPE OF THE "IMPOSSIBILITY" RESULTS

The nonuniformity phenomenon described, e.g., in (20) of Theorem 2.3 is caused by a mechanism that can informally be described as follows. Under the assumptions of that theorem, one can find an appropriate θ and an appropriate sequence $\vartheta_n = \theta + \gamma/\sqrt{n}$ exhibiting two crucial properties:

a. The probability measures $P_{n,\vartheta_n,\sigma}$ corresponding to ϑ_n are "close" to the measures $P_{n,\theta,\sigma}$ corresponding to θ, in the sense of contiguity. This entails that an estimator, which converges to some limit in probability under $P_{n,\theta,\sigma}$, converges to the *same limit* also under $P_{n,\vartheta_n,\sigma}$.
b. For given t, the estimands $G_{n,\vartheta_n,\sigma}(t)$ corresponding to ϑ_n are "far away" from the estimands $G_{n,\theta,\sigma}(t)$ corresponding to θ, in the sense that $G_{n,\vartheta_n,\sigma}(t)$ and $G_{n,\theta,\sigma}(t)$ converge to *different limits*; i.e., $G_{\infty,\theta,\sigma,0}(t)$ is different from $G_{\infty,\theta,\sigma,\gamma}(t)$.

In view of property a, an estimator $\hat{G}_n(t)$ satisfying $\hat{G}_n(t) - G_{n,\theta,\sigma}(t) \to 0$ in probability under $P_{n,\theta,\sigma}$ also satisfies $\hat{G}_n(t) - G_{n,\theta,\sigma}(t) \to 0$ in probability under $P_{n,\vartheta_n,\sigma}$. In view of property b, such an estimator $\hat{G}_n(t)$ is hence "far away" from the estimand $G_{n,\vartheta_n,\sigma}(t)$ with high probability under $P_{n,\vartheta_n,\sigma}$. In other words, an estimator that is close to $G_{n,\theta,\sigma}(t)$ under $P_{n,\theta,\sigma}$ must be far away from $G_{n,\vartheta_n,\sigma}(t)$ under $P_{n,\vartheta_n,\sigma}$. Formalized and refined, this argument leads to (20) and, as a consequence, to the nonexistence of uniformly consistent estimators for $G_{n,\theta,\sigma}(t)$. (There are a number of technical details in this formalization process that need careful attention to obtain the results in their full strength as given in Sects. 2 and 3.)

The preceding informal argument that derives (20) from properties a and b can be refined and formalized in a much more general and abstract framework; see Section 3 of Leeb and Pötscher (2006a) and the references therein. That paper also provides a general framework for deriving results like (21) and (22) of Theorem 2.3. The mechanism leading to such lower bounds is similar to the one outlined earlier, where for some of the results the concept of contiguity of the probability measures involved has to be replaced by closeness of these measures in total variation distance. We use the results in Section 3 of Leeb and

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

Pötscher (2006a) to formally convert properties a and b into the "impossibility" results of the present paper; cf. Appendix C.

Verifying the aforementioned property a in the context of the present paper is straightforward because we consider a Gaussian linear model. What is technically more challenging and requires some work is the verification of property b; this is done in Appendix A inter alia and rests on results of Leeb (2002, 2005, 2006).

Two important observations on properties a and b are in order: First, property a is typically satisfied in general parametric models under standard regularity conditions; e.g., it is satisfied whenever the model is locally asymptotically normal. Second, property b relies on limiting properties only and not on the finite-sample structure of the underlying statistical model. Now, the limit distributions of post-model-selection estimators in sufficiently regular parametric or semiparametric models are typically the same as the limiting distributions of the corresponding post-model-selection estimators in a Gaussian linear model (see, e.g., Sen, 1979; Pötscher, 1991; Nickl, 2003; Hjort and Claeskens, 2003). Hence, establishing property b for the Gaussian linear model then typically establishes the same result for a large class of general parametric or semiparametric models.[1] For example, property b can be verified for a large class of pretest estimators in sufficiently regular parametric models by arguing as in Appendix A and using the results of Nickl (2003) to reduce to the Gaussian linear case. Hence, the "impossibility" result given in Theorem 2.3 can be extended to more general parametric and semiparametric models with ease. The fact that we use a Gaussian linear model for the analysis in the present paper is a matter of convenience rather than a necessity.

The nonuniformity results in Theorem 2.3 are for (conservative) "general-to-specific" model selection from a nested family of models. Theorem 3.1 extends this to more general (conservative) model-selection procedures (including AIC and related procedures) and to more general families of models. The proof of Theorem 3.1 proceeds by reducing the problem to one where only two nested models are considered and then appealing to the results of Theorem 2.3. The condition on the model-selection procedures that enables this reduction is condition (24). It is apparent from the discussion in Section 3 that this condition is satisfied for many model-selection procedures. Furthermore, for the same reasons as given in the preceding paragraph, Theorem 3.1 also can easily be extended to sufficiently regular parametric and semiparametric models.

The "impossibility" results in the present paper are formulated for estimating $G_{n,\theta,\sigma}(t)$ for a given value of t. Suppose that we are now asking the question whether the cdf $G_{n,\theta,\sigma}(\cdot)$ viewed as a *function* can be estimated uniformly consistently, where consistency is relative to a metric that metrizes weak convergence.[2] Using similar reasoning as before (which can again be made formal by using, e.g., Lemma 3.1 in Leeb and Pötscher, 2006a) the key step now is to show that the *function* $G_{\infty,\theta,\sigma,0}(\cdot)$ is different from the *function* $G_{\infty,\theta,\sigma,\gamma}(\cdot)$. Obviously, it is a much simpler problem to find a γ such that the *functions* $G_{\infty,\theta,\sigma,0}(\cdot)$

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

and $G_{\infty,\theta,\sigma,\gamma}(\cdot)$ differ than to find a γ such that the *values* $G_{\infty,\theta,\sigma,0}(t)$ and $G_{\infty,\theta,\sigma,\gamma}(t)$ for a given t differ. Certainly, because we have solved the latter problem in Appendix A, this also provides an answer to the former. This then immediately delivers the desired "impossibility" result. (We note that in some special cases simpler arguments than the ones used in Appendix A can be employed to solve the former problem: For example, in the case $A = I$ the functions $G_{\infty,\theta,\sigma,0}(\cdot)$ and $G_{\infty,\theta,\sigma,\gamma}(\cdot)$ can both be shown to be convex combinations of cdfs that are concentrated on subspaces of different dimensions. This can be exploited to establish without much difficulty that the *functions* $G_{\infty,\theta,\sigma,0}(\cdot)$ and $G_{\infty,\theta,\sigma,\gamma}(\cdot)$ differ. For purpose of comparison we note that for general A the distributions $G_{\infty,\theta,\sigma,0}$ and $G_{\infty,\theta,\sigma,\gamma}$ can both be absolutely continuous with respect to Lebesgue measure, not allowing one to use this simple argument.) Again the discussion in this paragraph extends to more general parametric and semiparametric models without difficulty.

The present paper, including the discussion in this section, has focused on conservative model-selection procedures. However, the discussion should make it clear that similar "impossibility" results plague consistent model selection. Section 2.3 in Leeb and Pötscher (2006a) in fact gives such an "impossibility" result in a simple case.

We close with the following observations. Verification of property b, whether it is for $G_{\infty,\theta,\sigma,0}(t)$ and $G_{\infty,\theta,\sigma,\gamma}(t)$ (for given t) or for $G_{\infty,\theta,\sigma,0}(\cdot)$ and $G_{\infty,\theta,\sigma,\gamma}(\cdot)$, shows that the post-model-selection estimator $A\tilde{\theta}$ is a so-called nonregular estimator for $A\theta$: Consider an estimator $\tilde{\beta}$ in a parametric model $\{P_{n,\beta}: \beta \in B\}$ where the parameter space B is an open subset of euclidean space $\mathbb{R}^d$. Suppose $\tilde{\beta}$, properly scaled and centered, has a limit distribution under local alternatives, in the sense that $\sqrt{n}(\tilde{\beta} - (\beta + \gamma/\sqrt{n}))$ converges in law under $P_{n,\beta+\gamma/\sqrt{n}}$ to a limit distribution $L_{\infty,\beta,\gamma}(\cdot)$ for every γ. The estimator $\tilde{\beta}$ is called *regular* if for every β the limit distribution $L_{\infty,\beta,\gamma}(\cdot)$ does not depend on γ; cf. van der Vaart (1998, Sect. 8.5). Suppose now that the model is, e.g., locally asymptotically normal (hence the contiguity property in property a is satisfied). The informal argument outlined at the beginning of this section (and which is formalized in Lemma 3.1 of Leeb and Pötscher, 2006a) then in fact shows that the cdf of *any* nonregular estimator cannot be estimated uniformly consistently (where consistency is relative to any metric that metrizes weak convergence).

6. CONCLUSIONS

Despite the fact that we have shown that consistent estimators for the distribution of a post-model-selection estimator can be constructed with relative ease, we have also demonstrated that no estimator of this distribution can have satisfactory performance (locally) uniformly in the parameter space, even asymptotically. In particular, no (locally) uniformly consistent estimator of this distribution exists. Hence, the answer to the question posed in the title has to

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

be negative. The results in the present paper also cover the case of linear functions (e.g., predictors) of the post-model-selection estimator.

We stress here that resampling procedures such as, e.g., the bootstrap or subsampling do not solve the problem at all. First note that standard bootstrap techniques will typically not even provide consistent estimators of the finite-sample distribution of the post-model-selection estimator, as the bootstrap can be shown to stay random in the limit (Kulperger and Ahmed, 1992; Knight, 1999, Exp. 3).[3] Basically the only way one can coerce the bootstrap into delivering a consistent estimator is to resample from a model that has been selected by an *auxiliary* consistent model-selection procedure. The consistent estimator constructed in Section 2.2.1 is in fact of this form. In contrast to the standard bootstrap, subsampling will typically deliver consistent estimators. However, the "impossibility" results given in this paper apply to *any* estimator (including randomized estimators) of the cdf of a post-model-selection estimator. Hence, also *any* resampling-based estimator suffers from the nonuniformity defects described in Theorems 2.3 and 3.1; cf. also Remark 4.3.

The "impossibility" results in Theorems 2.3 and 3.1 are derived in the framework of a normal linear regression model (and a fortiori these results continue to hold in any model that includes the normal linear regression model as a special case), but this is more a matter of convenience than anything else: As discussed in Section 5, similar results can be obtained in general statistical models allowing for nonlinearity or dependent data, e.g., as long as standard regularity conditions for maximum likelihood theory are satisfied.

The results in the present paper are derived for a large class of conservative model-selection procedures (i.e., procedures that select overparameterized models with positive probability asymptotically) including the AIC and typical "general-to-specific" hypothesis testing procedures. For consistent model-selection procedures—such as BIC or testing procedures with suitably diverging critical values c_p (cf. Bauer, Pötscher, and Hackl, 1988)—the (pointwise) asymptotic distribution is always normal. (This is elementary; cf. Lemma 1 in Pötscher, 1991.) However, as discussed at length in Leeb and Pötscher (2005a), this asymptotic normality result paints a misleading picture of the finite-sample distribution, which can be far from normal, the convergence of the finite-sample distribution to the asymptotic normal distribution not being uniform. "Impossibility" results similar to the ones presented here can also be obtained for post-model-selection estimators based on consistent model-selection procedures. These will be discussed in detail elsewhere. For a simple special case such an "impossibility" result is given in Section 2.3 of Leeb and Pötscher (2006a).

The "impossibility" of estimating the distribution of the post-model-selection estimator does not per se preclude the possibility of conducting valid inference after model selection, a topic that deserves further study. However, it certainly makes this a more challenging task.

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

NOTES

1. Some care has to be taken here. In the Gaussian linear case the finite-sample cdfs converge at every value of the argument t, cf. Proposition 2.1. In a general parametric model, sometimes the asymptotic results (e.g., Hjort and Claeskens, 2003, Thm. 4.1) only guarantee weak convergence. Hence, to ensure convergence of the relevant cdfs at a given argument t as required in property b, additional considerations have to be employed. (This is, however, of no concern in the context discussed in the seventh paragraph in Sect. 5.)

2. Or, in fact, any metric with respect to which the relevant cdfs converge.

3. Brownstone (1990) claims the validity of a bootstrap procedure that is based on a conservative model-selection procedure in a linear regression model. Kilian (1998) makes a similar claim in the context of autoregressive models selected by a conservative model-selection procedure. Also the paper by Hansen (2003) contains such a claim for a stationary bootstrap procedure based on a conservative model-selection procedure. The preceding discussion intimates that these claims are at least unsubstantiated.

REFERENCES

Ahmed, S.E. & A.K. Basu (2000) Least squares, preliminary test and Stein-type estimation in general vector $AR(p)$ models. *Statistica Neerlandica* 54, 47–66.

Bauer, P., B.M. Pötscher, & P. Hackl (1988) Model selection by multiple test procedures. *Statistics* 19, 39–44.

Billingsley, P. (1995) *Probability and Measure*, 3rd ed. Wiley.

Brownstone, D. (1990) Bootstrapping improved estimators for linear regression models. *Journal of Econometrics* 44, 171–187.

Danilov, D.L. & J.R. Magnus (2004) On the harm that ignoring pre-testing can cause. *Journal of Econometrics* 122, 27–46.

Dijkstra, T.K. & J.H. Veldkamp (1988) Data-driven selection of regressors and the bootstrap. *Lecture Notes in Economics and Mathematical Systems* 307, 17–38.

Dukić, V.M. & E.A. Peña (2005) Variance estimation in a model with Gaussian submodel. *Journal of the American Statistical Association* 100, 296–309.

Freedman, D.A., W. Navidi, & S.C. Peters (1988) On the impact of variable selection in fitting regression equations. *Lecture Notes in Economics and Mathematical Systems* 307, 1–16.

Hansen, P.R. (2003) Regression Analysis with Many Specifications: A Bootstrap Method for Robust Inference. Working paper, Department of Economics, Brown University.

Hjort, N.L. & G. Claeskens (2003) Frequentist model average estimators. *Journal of the American Statistical Association* 98, 879–899.

Kabaila, P. (1995) The effect of model selection on confidence regions and prediction regions. *Econometric Theory* 11, 537–549.

Kapetanios, G. (2001) Incorporating lag order selection uncertainty in parameter inference for AR models. *Economics Letters* 72, 137–144.

Kilian, L. (1998) Accounting for lag order uncertainty in autoregressions: The endogenous lag order bootstrap algorithm. *Journal of Time Series Analysis* 19, 531–548.

Knight, K. (1999) Epi-convergence in Distribution and Stochastic Equi-semicontinuity. Working paper, Department of Statistics, University of Toronto.

Kulperger, R.J. & S.E. Ahmed (1992) A bootstrap theorem for a preliminary test estimator. *Communications in Statistics: Theory and Methods* 21, 2071–2082.

Leeb, H. (2002) On a differential equation with advanced and retarded arguments. *Communications on Applied Nonlinear Analysis* 9, 77–86.

Leeb, H. (2005) The distribution of a linear predictor after model selection: Conditional finite-sample distributions and asymptotic approximations. *Journal of Statistical Planning and Inference* 134, 64–89.

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

Leeb, H. (2006) The distribution of a linear predictor after model selection: Unconditional finite-sample distributions and asymptotic approximations. *IMS Lecture Notes-Monograph Series* 49, 291–311.
Leeb, H. & B.M. Pötscher (2003) The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* 19, 100–142.
Leeb, H. & B.M. Pötscher (2005a) Model selection and inference: Facts and fiction. *Econometric Theory* 21, 21–59.
Leeb, H. & B.M. Pötscher (2005b) Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators? Working paper, Department of Statistics, University of Vienna.
Leeb, H. & B.M. Pötscher (2006a) Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk bound results. *Econometric Theory* 22, 69–97. (Corrigendum, *Econometric Theory*, forthcoming.)
Leeb, H. & B.M. Pötscher (2006b) Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 34, 2554–2591.
Lehmann, E.L. & G. Casella (1998) *Theory of Point Estimation*, 2nd ed. Springer Texts in Statistics. Springer-Verlag.
Nickl, R. (2003) Asymptotic distribution theory of post-model-selection maximum likelihood estimators. Master's thesis, Department of Statistics, University of Vienna.
Pötscher, B.M. (1991) Effects of model selection on inference. *Econometric Theory* 7, 163–185.
Pötscher, B.M. (1995) Comment on "The effect of model selection on confidence regions and prediction regions" by P. Kabaila. *Econometric Theory* 11, 550–559.
Pötscher, B.M. & A.J. Novak (1998) The distribution of estimators after model selection: Large and small sample results. *Journal of Statistical Computation and Simulation* 60, 19–56.
Rao, C.R. & Y. Wu (2001) On model selection. *IMS Lecture Notes-Monograph Series* 38, 1–57.
Robinson, G.K. (1979) Conditional properties of statistical procedures. *Annals of Statistics* 7, 742–755.
Sen, P.K. (1979) Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics* 7, 1019–1033.
Sen, P.K. & A.K.M.E. Saleh (1987) On preliminary test and shrinkage *M*-estimation in linear models. *Annals of Statistics* 15, 1580–1592.
van der Vaart, A.W. (1998) *Asymptotic Statistics.* Cambridge University Press.

APPENDIX A: Auxiliary Lemmas

LEMMA A.1. *Let Z be a random vector with values in $\mathbf{R}^k$ and let W be a univariate standard Gaussian random variable independent of Z. Furthermore, let $C \in \mathbf{R}^k$ and $\tau > 0$. Then*

$$\mathbb{P}(Z \leq Cx)\mathbb{P}(|W - x| < \tau) + \mathbb{P}(Z \leq CW, |W - x| \geq \tau) \tag{A.1}$$

is constant as a function of $x \in \mathbf{R}$ if and only if $C = 0$ or $\mathbb{P}(Z \leq Cx) = 0$ for each $x \in \mathbf{R}$.

Proof of Lemma A.1. Suppose $C = 0$ holds. Using independence of Z and W it is then easy to see that (A.1) reduces to $\mathbb{P}(Z \leq 0)$, which is constant in x. If $\mathbb{P}(Z \leq Cx) = 0$ for every $x \in \mathbf{R}$, then $\mathbb{P}(Z \leq CW) = 0$, and hence (A.1) is again constant, namely, equal to zero.

To prove the converse, assume that (A.1) is constant in $x \in \mathbf{R}$. Letting $x \to \infty$, we see that (A.1) must be equal to $\mathbb{P}(Z \leq CW)$. This entails that

$$\mathbb{P}(Z \leq Cx)\mathbb{P}(|W - x| < \tau) = \mathbb{P}(Z \leq CW, |W - x| < \tau)$$

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

holds for every $x \in \mathbf{R}$. Write $F(x)$ as shorthand for $\mathbb{P}(Z \leq Cx)$ and let $\Phi(z)$ and $\phi(z)$ denote the cdf and density of W, respectively. Then the expression in the preceding display can be written as

$$F(x)(\Phi(x+\tau) - \Phi(x-\tau)) = \int_{x-\tau}^{x+\tau} F(z)\phi(z)\,dz \qquad (x \in \mathbf{R}). \tag{A.2}$$

We now further assume that $C \neq 0$ and that $F(x) \neq 0$ for at least one $x \in \mathbf{R}$ and show that this leads to a contradiction.

Consider first the case where all components of C are nonnegative. Because F is not identically zero, it is then, up to a scale factor, the cdf of a random variable on the real line. But then (A.2) cannot hold for all $x \in \mathbf{R}$ as shown in Example 7 in Leeb (2002) (cf. also equation (7) in that paper). The case where all components of C are nonpositive follows similarly by applying the preceding argument to $F(-x)$ and upon observing that both $\Phi(x+\tau) - \Phi(x-\tau)$ and $\phi(x)$ are symmetric around $x = 0$.

Finally, consider the case where C has at least one positive and one negative component. In this case clearly $\lim_{x\to-\infty} F(x) = \lim_{x\to\infty} F(x) = 0$ holds. Because $F(x)$ is continuous in view of (A.2), we see that $F(x)$ attains its (positive) maximum at some point $x_1 \in \mathbf{R}$. Now note that (A.2) with x_1 replacing x can be written as

$$\int_{x_1-\tau}^{x_1+\tau} (F(x_1) - F(z))\phi(z)\,dz = 0.$$

This immediately entails that $F(x) = F(x_1)$ for each $x \in [x_1 - \tau, x_1 + \tau]$ (because $F(x)$ is continuous and because of the definition of x_1). Repeating this argument with $x_1 - \tau$ replacing x_1 and proceeding inductively, we obtain that $F(x) = F(x_1)$ for each x satisfying $x \leq x_1 + \tau$, a contradiction with $\lim_{x\to-\infty} F(x) = 0$. ■

LEMMA A.2. *Let M and N be matrices of dimension $k \times p$ and $k \times q$, respectively, such that the matrix $(M:N)$ has rank k ($k \geq 1$, $p \geq 1$, $q \geq 1$). Let $t \in \mathbf{R}^k$ and let V be a random vector with values in $\mathbf{R}^p$ whose distribution assigns positive mass to every (nonempty) open subset of $\mathbf{R}^p$ (e.g., it possesses an almost everywhere positive Lebesgue density). Set $f(x) = \mathbb{P}(MV \leq t + Nx)$, $x \in \mathbf{R}^q$. If one of the rows of M consists of zeros only, then f is discontinuous at some point x_0. More precisely, there exist $x_0 \in \mathbf{R}^q$, $z \in \mathbf{R}^q$, and a constant $c > 0$, such that $f(x_0 + \delta z) \geq c$ and $f(x_0 - \delta z) = 0$ hold for every sufficiently small $\delta > 0$.*

Proof of Lemma A.2. The case where M is the zero matrix is trivial. Otherwise, let I_0 denote the set of indices i, $1 \leq i \leq k$, for which the ith row of M is zero. Let $(M_0:N_0)$ denote the matrix consisting of those rows of $(M:N)$ whose index is in I_0 and let $(M_1:N_1)$ denote the matrix consisting of the remaining rows of $(M:N)$. Clearly, M_0 is then the zero matrix. Furthermore, note that N_0 has full row rank. Moreover, let t_0 denote the vector consisting of those components of t whose index is in I_0 and let t_1 denote the vector containing the remaining components. With this notation, $f(x)$ can be written as $\mathbb{P}(0 \leq t_0 + N_0 x, M_1 V \leq t_1 + N_1 x)$.

For vectors $\mu \in \mathbf{R}^p$ and $\eta \in \mathbf{R}^q$ to be specified in a moment, set $t^* = t + M\mu + N\eta$ and let t_0^* and t_1^* be defined similarly to t_0 and t_1. Because the matrix $(M:N)$ has full rank k, we can choose μ and η such that $t_0^* = 0$ and $t_1^* > 0$. Choose $z \in \mathbf{R}^q$ such that

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

$N_0 z > 0$, which is possible because N_0 has full row rank. Set $x_0 = \eta$. Then for every $\epsilon \in \mathbb{R}$ we have

$$\begin{aligned} f(x_0 + \epsilon z) &= f(\eta + \epsilon z) = \mathbb{P}(MV \le t + N(\eta + \epsilon z)) \\ &= \mathbb{P}(0 \le t_0 + N_0(\eta + \epsilon z), M_1 V \le t_1 + N_1(\eta + \epsilon z)) \\ &= \mathbb{P}(0 \le t_0^* + \epsilon N_0 z, M_1(V + \mu) \le t_1^* + \epsilon N_1 z) \\ &= \mathbb{P}(0 \le \epsilon N_0 z, M_1(V + \mu) \le t_1^* + \epsilon N_1 z). \end{aligned}$$

Because $t_1^* > 0$, we can find a t_1^{**} such that $0 < t_1^{**} < t_1^* + \epsilon N_1 z$ holds for every ϵ with $|\epsilon|$ small enough. If now $\epsilon > 0$ then

$$f(x_0 + \epsilon z) = \mathbb{P}(M_1(V + \mu) \le t_1^* + \epsilon N_1 z) \ge \mathbb{P}(M_1(V + \mu) \le t_1^{**}).$$

The right-hand side in the preceding display is positive because $t_1^{**} > 0$ and because the distribution of $M_1(V + \mu)$ assigns positive mass to any neighborhood of the origin, because the same is true for the distribution of $V + \mu$ and because M_1 maps neighborhoods of zero into neighborhoods of zero. Setting $c = \mathbb{P}(M_1(V + \mu) \le t_1^{**})/2$, we have $f(x_0 + \epsilon z) \ge c > 0$ for each sufficiently small $\epsilon > 0$. Furthermore, for $\epsilon < 0$ we have $f(x_0 + \epsilon z) = 0$, because $f(x_0 + \epsilon z) \le \mathbb{P}(0 \le \epsilon N_0 z) = 0$ in view of $N_0 z > 0$. ■

LEMMA A.3. *Let Z be a random vector with values in $\mathbf{R}^p$, $p \ge 1$, with a distribution that is absolutely continuous with respect to Lebesgue measure on $\mathbf{R}^p$. Let B be a $k \times p$ matrix, $k \ge 1$. Then the cdf $\mathbb{P}(BZ \le \cdot)$ of BZ is discontinuous at $t \in \mathbf{R}^k$ if and only if $\mathbb{P}(BZ \le t) > 0$ and if for some i_0, $1 \le i_0 \le k$, the i_0th row of B and the i_0th component of t are both zero, i.e., $B_{i_0,\cdot} = (0, \ldots, 0)$ and $t_{i_0} = 0$.*

Proof of Lemma A.3. To establish sufficiency of the preceding condition, let $\mathbb{P}(BZ \le t) > 0$, $t_{i_0} = 0$, and $B_{i_0,\cdot} = (0, \ldots, 0)$ for some i_0, $1 \le i_0 \le k$. Then, of course, $\mathbb{P}(B_{i_0,\cdot} Z = 0) = 1$. For $t_n = t - n^{-1} e_{i_0}$, where e_{i_0} denotes the i_0th unit vector in $\mathbf{R}^k$, we have $\mathbb{P}(BZ \le t_n) \le \mathbb{P}(B_{i_0,\cdot} Z \le t_{n,i_0}) = \mathbb{P}(B_{i_0,\cdot} Z \le -1/n) = 0$ for every n. Consequently, $\mathbb{P}(BZ \le t)$ is discontinuous at t.

To establish necessity, we first show that the following holds: If $t_n \in \mathbf{R}^k$ is a sequence converging to $t \in \mathbf{R}^k$ as $n \to \infty$, then every accumulation point of the sequence $\mathbb{P}(BZ \le t_n)$ has the form

$$\mathbb{P}(B_{i_1,\cdot} Z \le t_{i_1}, \ldots, B_{i_m,\cdot} Z \le t_{i_m}, B_{i_{m+1},\cdot} Z < t_{i_{m+1}}, \ldots, B_{i_k,\cdot} Z < t_{i_k}) \tag{A.3}$$

for some m, $0 \le m \le k$, and for some permutation $(i_1, \ldots, i_k)$ of $(1, \ldots, k)$. This can be seen as follows: Let α be an accumulation point of $\mathbb{P}(BZ \le t_n)$. Then we may find a subsequence such that $\mathbb{P}(BZ \le t_n)$ converges to α along this subsequence. From this subsequence we may even extract a further subsequence along which each component of the $k \times 1$ vector t_n converges to the corresponding component of t monotonously, i.e., either from above or from below. Without loss of generality, we may also assume that those components that converge from below are strictly increasing. The resulting subsequence will be denoted by n_j in what follows. Assume that the components of t_{n_j} with indices $i_1, \ldots, i_m$ converge from above and the components with indices $i_{m+1}, \ldots, i_k$ converge from below. Now

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

$$\mathbb{P}(BZ \le t_{n_j}) = \int_{z \in \mathbf{R}^k} \prod_{s=1}^{k} \mathbf{1}_{(-\infty, t_{n_j,s}]}(z_s) \mathbb{P}_{BZ}(dz), \tag{A.4}$$

where $\mathbb{P}_{BZ}$ denotes the distribution of BZ. The integrand in (A.4) now converges to $\prod_{l=1}^{m} 1_{(-\infty, t_{i_l}]}(z_{i_l}) \prod_{l=m+1}^{k} 1_{(-\infty, t_{i_l})}(z_{i_l})$ for all $z \in \mathbf{R}^k$ as $n_j \to \infty$. The right-hand side of (A.4) converges to the expression in (A.3) as $n_j \to \infty$ by the dominated convergence theorem, and the left-hand side of (A.4) converges to α by construction. This establishes the claim regarding (A.3).

Now suppose that $\mathbb{P}(BZ \le t)$ is discontinuous at t; i.e., there exists a sequence t_n converging to t as $n \to \infty$ such that $\mathbb{P}(BZ \le t_n)$ does not converge to $\mathbb{P}(BZ \le t)$ as $n \to \infty$. From the sequence t_n we can extract a subsequence t_{n_s} along which $\mathbb{P}(BZ \le t_{n_s})$ converges to a limit different from $\mathbb{P}(BZ \le t)$ as $n_s \to \infty$. As shown earlier, the limit has to be of the form (A.3), and $m < k$ has to hold. Consequently, the limit of $\mathbb{P}(BZ \le t_{n_s})$ is smaller than $\mathbb{P}(BZ \le t) = \mathbb{P}(B_{i,\cdot} Z \le t_i,\ i = 1, \ldots, k)$. The difference of $\mathbb{P}(BZ \le t)$ and the limit of $\mathbb{P}(BZ \le t_{n_s})$ is positive and because of (A.3) can be written as

$$\mathbb{P}(B_{i_j,\cdot} Z \le t_{i_j} \quad \text{for each } j = 1, \ldots, k, \qquad B_{i_j,\cdot} Z = t_{i_j} \quad \text{for some } j = m+1, \ldots, k) > 0.$$

We thus see that $\mathbb{P}(B_{i_{j_0},\cdot} Z = t_{i_{j_0}}) > 0$ for some j_0 satisfying $m + 1 \le j_0 \le k$. As Z is absolutely continuous with respect to Lebesgue measure on $\mathbf{R}^p$, this can only happen if $B_{i_{j_0},\cdot} = (0, \ldots, 0)$ and $t_{i_{j_0}} = 0$. ■

LEMMA A.4. *Suppose that $A\tilde{\theta}(q)$ and $\tilde{\theta}_q(q)$ are asymptotically correlated, i.e., $C_\infty^{(q)} \neq 0$, for some q satisfying $\mathcal{O} < q \le P$, and let q^* denote the largest q with this property. Moreover let $\theta \in M_{q^*-1}$, let σ satisfy $0 < \sigma < \infty$, and let $t \in \mathbf{R}^k$. Then $G_{\infty,\theta,\sigma,\gamma}(t)$ is nonconstant as a function of $\gamma \in M_{q^*} \backslash M_{q^*-1}$. More precisely, there exist $\delta_0 > 0$ and ρ_0, $0 < \rho_0 < \infty$ such that*

$$\sup_{\substack{\gamma^{(1)}, \gamma^{(2)} \in M_{q^*} \backslash M_{q^*-1} \\ \|\gamma^{(i)}\| < \rho_0,\ i=1,2}} |G_{\infty,\theta,\sigma,\gamma^{(1)}}(t) - G_{\infty,\theta,\sigma,\gamma^{(2)}}(t)| > 2\delta_0 \tag{A.5}$$

holds. The constants δ_0 and ρ_0 can be chosen in such a way that they depend only on t, Q, A, σ, and the critical values c_p for $\mathcal{O} < p \le P$.

LEMMA A.5. *Suppose that $A\tilde{\theta}(q)$ and $\tilde{\theta}_q(q)$ are asymptotically correlated, i.e., $C_\infty^{(q)} \neq 0$, for some q satisfying $\mathcal{O} < q \le P$, and let q^* denote the largest q with this property. Suppose further that for some $p_\odot$ satisfying $\mathcal{O} \le p_\odot < q^*$ either $p_\odot = 0$ holds or $p_\odot > 0$ and $A[p_\odot]$ has a row of zeros. Then, for every $\theta \in M_{p_\odot}$, every σ, $0 < \sigma < \infty$, and every $t \in \mathbf{R}^k$ the quantity $G_{\infty,\theta,\sigma,\gamma}(t)$ is discontinuous as a function of $\gamma \in M_{q^*}$. More precisely, for each $s = \mathcal{O}, \ldots, p_\odot$, there exist vectors β_* and γ_* in M_{q^*} and constants $\delta_* > 0$ and $\epsilon_* > 0$ such that*

$$|G_{\infty,\theta,\sigma,\beta_*+\epsilon\gamma_*}(t) - G_{\infty,\theta,\sigma,\beta_*-\epsilon\gamma_*}(t)| \ge \delta_* \tag{A.6}$$

holds for every θ satisfying $\max\{p_0(\theta), \mathcal{O}\} = s$ and for every ϵ with $0 < \epsilon < \epsilon_*$. The quantities δ_*, ϵ_*, β_*, and γ_* can be chosen in such a way that—besides t, Q, A, σ, and the critical values c_p for $\mathcal{O} < p \le P$—they depend on θ only through $\max\{p_0(\theta), \mathcal{O}\}$.

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

Before we prove the preceding lemmas, we provide a representation of $G_{\infty,\theta,\sigma,\gamma}(t)$ that will be useful in the following discussion: For $0 < p \leq P$ define $Z_p = \sum_{r=1}^{p} \xi_{\infty,r}^{-2} C_\infty^{(r)} W_r$, where $C_\infty^{(r)}$ has been defined after (13) and the random variables W_r are independent normally distributed with mean zero and variances $\sigma^2 \xi_{\infty,r}^2$. For convenience, let Z_0 denote the zero vector in $\mathbf{R}^k$. Observe that Z_p, $p > 0$, is normally distributed with mean zero and variance-covariance matrix $\sigma^2 A[p]Q[p:p]^{-1}A[p]'$ because it has been shown in the proof of Proposition 4.4 in Leeb and Pötscher (2006b) that the asymptotic variance-covariance matrix $\sigma^2 A[p]Q[p:p]^{-1}A[p]'$ of $\sqrt{n}A\tilde{\theta}(p)$ can be expressed as $\sum_{r=1}^{p} \sigma^2 \xi_{\infty,r}^{-2} C_\infty^{(r)} C_\infty^{(r)\prime}$. Also the joint distribution of Z_p and the set of variables W_r, $1 \leq r \leq P$, is normal, with the covariance vector between Z_p and W_r given by $\sigma^2 C_\infty^{(r)}$ in the case $r \leq p$; otherwise Z_p and W_r are independent. Define the constants $\nu_r = \gamma_r + (Q[r:r]^{-1}Q[r:\neg r]\gamma[\neg r])_r$ for $0 < r \leq P$. It is now easy to see that for $p \geq p_* = \max\{p_0(\theta), \mathcal{O}\}$ the quantity $\beta^{(p)}$ defined in Proposition 2.1 equals $-\sum_{r=p+1}^{P} \xi_{\infty,r}^{-2} C_\infty^{(r)} \nu_r$. (This is seen as follows: It was noted in Proposition 2.1 that $\beta^{(p)} = \lim_{n\to\infty} \sqrt{n}A(\eta_n(p) - \theta - \gamma/\sqrt{n})$ for $p \geq p_0(\theta)$, when $\eta_n(p)$ is defined as in (9), but with $\theta + \gamma/\sqrt{n}$ replacing θ. Using the representation (20) of Leeb (2005) and taking limits, the result follows if we observe that $\sqrt{n}\eta_{n,r}(r) \to \nu_r$ for $r > p \geq p_0(\theta)$.) The cdf in (15) can now be written as

$$\begin{aligned}
&\mathbb{P}\left(Z_{p_*} \leq t + \sum_{r=p_*+1}^{P} \xi_{\infty,r}^{-2} C_\infty^{(r)} \nu_r\right) \prod_{q=p_*+1}^{P} \mathbb{P}(|W_q + \nu_q| < c_q \sigma \xi_{\infty,q}) \\
&\quad + \sum_{p=p_*+1}^{P} \mathbb{P}\left(Z_p \leq t + \sum_{r=p+1}^{P} \xi_{\infty,r}^{-2} C_\infty^{(r)} \nu_r, |W_p + \nu_p| \geq c_p \sigma \xi_{\infty,p}\right) \\
&\quad \times \prod_{q=p+1}^{P} \mathbb{P}(|W_q + \nu_q| < c_q \sigma \xi_{\infty,q}). \qquad \textbf{(A.7)}
\end{aligned}$$

That the terms corresponding to $p = p_*$ in (A.7) and (15) agree is obvious. Furthermore, for each $p > p_*$ the terms under the product sign in (A.7) and (15) coincide by definition of the function $\Delta_s(a,b)$. It is also easy to see that the conditional distribution of W_p given $Z_p = z$ is Gaussian with mean $b_{\infty,p} z$ and variance $\sigma^2 \zeta_{\infty,p}^2$. Consequently, the probability of the event $\{|W_p + \nu_p| \geq c_p \sigma \xi_{\infty,p}\}$ conditional on $Z_p = z$ is given by the integrand shown in (15). Because Z_p has distribution $\Phi_{\infty,p}$ as noted before, it follows that (A.7) and (15) agree.

Remark A.1. If $C_\infty^{(p)} = 0$ for all $p > p_*$, then in view of the preceding discussion $Z_{p_*} = Z_p = Z_P$, and hence $\Phi_{\infty,p_*} = \Phi_{\infty,p} = \Phi_{\infty,P}$ holds for all $p > p_*$. Using the independence of W_r, $r > p_*$, from Z_{p_*}, inspection of (A.7) shows that $G_{\infty,\theta,\sigma,\gamma}$ reduces to $\Phi_{\infty,P}$; see also Leeb (2006, Rmk. 5.2).

Proof of Lemma A.4. From (A.7) (or (15)) it follows that the map $\gamma \mapsto G_{\infty,\theta,\sigma,\gamma}(t)$ depends only on t, Q, A, σ, the critical values c_p for $\mathcal{O} < p \leq P$, and also on θ; however, the dependence on θ is only through $p_* = \max\{p_0(\theta), \mathcal{O}\}$. It hence suffices to find, for each possible value of p_* in the range $p_* = \mathcal{O}, \ldots, q^* - 1$, constants $0 < \rho_0 < \infty$ and $\delta_0 > 0$ such that (A.5) is satisfied for some (and hence all) θ returning this particular value of $p_* = \max\{p_0(\theta), \mathcal{O}\}$. For this in turn it is sufficient to show that for every $\theta \in M_{q^*-1}$ the quantity $G_{\infty,\theta,\sigma,\gamma}(t)$ is nonconstant as a function of $\gamma \in M_{q^*} \backslash M_{q^*-1}$.

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

Let $\theta \in M_{q^*-1}$ and assume that $G_{\infty,\theta,\sigma,\gamma}(t)$ is constant in $\gamma \in M_{q^*}\backslash M_{q^*-1}$. Observe that, by assumption, $C_\infty^{(q^*)}$ is nonzero and $C_\infty^{(p)} = 0$ for $p > q^*$. For $\gamma \in M_{q^*}$, we clearly have $\nu_{q^*} = \gamma_{q^*}$ and $\nu_r = 0$ for $r > q^*$. Letting $\gamma_{q^*-1} \to \infty$ while γ_{q^*} is held fixed, we see that $\nu_{q^*-1} \to \infty$; hence,

$$\mathbb{P}(|W_{q^*-1} + \nu_{q^*-1}| < c_{q^*-1}\sigma\xi_{\infty,q^*-1}) \to 0.$$

It follows that (A.7) converges to

$$\begin{aligned}
&\mathbb{P}(Z_{q^*-1} \le t + \xi_{\infty,q^*}^{-2} C_\infty^{(q^*)} \gamma_{q^*}) \mathbb{P}(|W_{q^*} + \gamma_{q^*}| < c_{q^*}\sigma\xi_{\infty,q^*}) \prod_{q=q^*+1}^{P} \mathbb{P}(|W_q| < c_q\sigma\xi_{\infty,q}) \\
&\quad + \mathbb{P}(Z_{q^*} \le t, |W_{q^*} + \gamma_{q^*}| \ge c_{q^*}\sigma\xi_{\infty,q^*}) \prod_{q=q^*+1}^{P} \mathbb{P}(|W_q| < c_q\sigma\xi_{\infty,q}) \\
&\quad + \sum_{p=q^*+1}^{P} \mathbb{P}(Z_p \le t, |W_p| \ge c_p\sigma\xi_{\infty,p}) \prod_{q=p+1}^{P} \mathbb{P}(|W_q| < c_q\sigma\xi_{\infty,q}). \qquad \textbf{(A.8)}
\end{aligned}$$

By assumption, the expression in the preceding display is constant in $\gamma_{q^*} \in \mathbf{R}\backslash\{0\}$. Dropping the terms that do not depend on γ_{q^*} and observing that $\mathbb{P}(|W_q| < c_q\sigma\xi_{\infty,q})$ is never zero for $q > q^* > \mathcal{O}$, we see that

$$\begin{aligned}
&\mathbb{P}(Z_{q^*-1} \le t + \xi_{\infty,q^*}^{-2} C_\infty^{(q^*)} \gamma_{q^*}) \mathbb{P}(|W_{q^*} + \gamma_{q^*}| < c_{q^*}\sigma\xi_{\infty,q^*}) \\
&\quad + \mathbb{P}(Z_{q^*} \le t, |W_{q^*} + \gamma_{q^*}| \ge c_{q^*}\sigma\xi_{\infty,q^*}) \qquad \textbf{(A.9)}
\end{aligned}$$

has to be constant in $\gamma_{q^*} \in \mathbf{R}\backslash\{0\}$. We now show that the expression in (A.9) is in fact constant in $\gamma_{q^*} \in \mathbf{R}$: Observe first that $\mathbb{P}(|W_{q^*} + \gamma_{q^*}| < c_{q^*}\sigma\xi_{\infty,q^*})$ is positive and continuous in $\gamma_{q^*} \in \mathbf{R}$; also the probability $\mathbb{P}(Z_{q^*} \le t, |W_{q^*} + \gamma_{q^*}| \ge c_{q^*}\sigma\xi_{\infty,q^*})$ is continuous in $\gamma_{q^*} \in \mathbf{R}$ because W_{q^*}, being normal with mean zero and positive variance, is absolutely continuously distributed. Concerning the remaining term in (A.9), we note that $Z_{q^*-1} = MV$ where $M = [\xi_{\infty,1}^{-2} C_\infty^{(1)}, \ldots, \xi_{\infty,q^*-1}^{-2} C_\infty^{(q^*-1)}]$ and $V = (W_1, \ldots, W_{q^*-1})'$. In the case where no row of M is identically zero, Lemma A.3 also shows that $\mathbb{P}(Z_{q^*-1} \le t + \xi_{\infty,q^*}^{-2} C_\infty^{(q^*)} \gamma_{q^*})$ is continuous in $\gamma_{q^*} \in \mathbf{R}$. Hence, in this case (A.9) is indeed constant for all $\gamma_{q^*} \in \mathbf{R}$. In the case where a row of M is identically zero, define $N = \xi_{\infty,q^*}^{-2} C_\infty^{(q^*)}$ and rewrite the probability in question as $\mathbb{P}(MV \le t + N\gamma_{q^*})$. Note that $(M:N)$ has full row rank k, because

$$\begin{aligned}
(M:N)\,\mathrm{diag}[\xi_{\infty,1}^2, \ldots, \xi_{\infty,q^*}^2](M:N)' &= \sum_{r=1}^{q^*} \xi_{\infty,r}^{-2} C_\infty^{(r)} C_\infty^{(r)\prime} = \sum_{r=1}^{P} \xi_{\infty,r}^{-2} C_\infty^{(r)} C_\infty^{(r)\prime} \\
&= AQ^{-1}A' \qquad \textbf{(A.10)}
\end{aligned}$$

by definition of q^* and because the latter matrix is nonsingular in view of rank $A = k$. Lemma A.2 then shows that there exists a $\gamma_{q^*}^{(0)} \in \mathbf{R}$, $z \in \{-1, 1\}$, and a constant $c > 0$ such that $\mathbb{P}(MV \le t + N(\gamma_{q^*}^{(0)} - \delta z)) = 0$ and $\mathbb{P}(MV \le t + N(\gamma_{q^*}^{(0)} + \delta z)) \ge c$ holds for

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

arbitrary small $\delta > 0$. Observe that $\gamma_{q^*}^{(0)} - \delta z$ and $\gamma_{q^*}^{(0)} + \delta z$ are nonzero for sufficiently small $\delta > 0$. But then (A.9)—being constant for $\gamma_{q^*} \in \mathbf{R}\backslash\{0\}$—gives the same value for $\gamma_{q^*} = \gamma_{q^*}^{(0)} - \delta z$ and $\gamma_{q^*} = \gamma_{q^*}^{(0)} + \delta z$ and all sufficiently small $\delta > 0$. Letting δ go to zero in this equality and using the continuity properties for the second and third probability in (A.9) noted above we obtain that

$$
\begin{aligned}
&c\mathbb{P}(|W_{q^*} + \gamma_{q^*}^{(0)}| < c_{q^*}\sigma\xi_{\infty,q^*}) + \mathbb{P}(Z_{q^*} \le t, |W_{q^*} + \gamma_{q^*}^{(0)}| \ge c_{q^*}\sigma\xi_{\infty,q^*}) \\
&\quad \le \liminf_{\delta\downarrow 0} \mathbb{P}(Z_{q^*-1} \le t + \xi_{\infty,q^*}^{-2} C_\infty^{(q^*)}(\gamma_{q^*}^{(0)} + \delta z))\mathbb{P}(|W_{q^*} + \gamma_{q^*}^{(0)}| < c_{q^*}\sigma\xi_{\infty,q^*}) \\
&\qquad + \mathbb{P}(Z_{q^*} \le t, |W_{q^*} + \gamma_{q^*}^{(0)}| \ge c_{q^*}\sigma\xi_{\infty,q^*}) \\
&\quad = \liminf_{\delta\downarrow 0} \mathbb{P}(Z_{q^*-1} \le t + \xi_{\infty,q^*}^{-2} C_\infty^{(q^*)}(\gamma_{q^*}^{(0)} - \delta z))\mathbb{P}(|W_{q^*} + \gamma_{q^*}^{(0)}| < c_{q^*}\sigma\xi_{\infty,q^*}) \\
&\qquad + \mathbb{P}(Z_{q^*} \le t, |W_{q^*} + \gamma_{q^*}^{(0)}| \ge c_{q^*}\sigma\xi_{\infty,q^*}) \\
&\quad = \mathbb{P}(Z_{q^*} \le t, |W_{q^*} + \gamma_{q^*}^{(0)}| \ge c_{q^*}\sigma\xi_{\infty,q^*}),
\end{aligned}
$$

which is impossible because $c > 0$ and $\mathbb{P}(|W_{q^*} + \gamma_{q^*}^{(0)}| < c_{q^*}\sigma\xi_{\infty,q^*}) > 0$. Hence we have shown that (A.9) is indeed constant for all $\gamma_{q^*} \in \mathbf{R}$.

Now write Z, W, C, τ, and x for $Z_{q^*-1} - t$, $-W_{q^*}/\sigma\xi_{\infty,q^*}$, $\sigma\xi_{\infty,q^*}^{-1} C_\infty^{(q^*)}$, c_{q^*}, and $\gamma_{q^*}/\sigma\xi_{\infty,q^*}$, respectively. Upon observing that Z_{q^*} equals $Z_{q^*-1} + \xi_{\infty,q^*}^{-2} C_\infty^{(q^*)} W_{q^*}$, it is easy to see that (A.9) can be written as in (A.1). By our assumptions, this expression is constant in $x = \gamma_{q^*}/\sigma\xi_{\infty,q^*} \in \mathbf{R}$. Lemma A.1 then entails that either $C = 0$ or $\mathbb{P}(Z \le Cx) = 0$ for each $x \in \mathbf{R}$. Because C equals $\sigma\xi_{\infty,q^*}^{-1} C_\infty^{(q^*)}$, it is nonzero by assumption. Hence,

$$
\mathbb{P}(Z_{q^*-1} \le t + \xi_{\infty,q^*}^{-2} C_\infty^{(q^*)} \gamma_{q^*}) = 0
$$

must hold for every value of γ_{q^*}. But the preceding probability is just the conditional probability that $Z_{q^*} \le t$ given $W_{q^*} = -\gamma_{q^*}$. It follows that $\mathbb{P}(Z_{q^*} \le t)$ equals zero also. By our assumption $C_\infty^{(p)} = 0$ for $p > q^*$, and hence $Z_{q^*} = Z_P$. We thus obtain $\mathbb{P}(Z_P \le t) = 0$, a contradiction with the fact that Z_P is a Gaussian random variable on $\mathbf{R}^k$ with nonsingular variance-covariance matrix $\sigma^2 A Q^{-1} A'$. ■

Inspection of the preceding proof shows that it can be simplified if the claim of nonconstancy of $G_{\infty,\theta,\sigma,\gamma}(t)$ as a function of $\gamma \in M_{q^*}\backslash M_{q^*-1}$ in Lemma A.4 is weakened to nonconstancy for $\gamma \in M_{q^*}$. The strong form of the lemma as given here is needed in the proof of Proposition B.1.

Proof of Lemma A.5. Let $p_\oplus$ be the largest index p, $\mathcal{O} \le p \le P$, for which $A[p]$ has a row of zeros and set $p_\oplus = 0$ if no such index exists. We first show that $p_\oplus$ satisfies $p_\oplus < q^*$. Suppose $p_\oplus \ge q^*$ would hold. Because $Z_{p_\oplus}$ is a Gaussian random vector with mean zero and variance-covariance matrix $\sigma^2 A[p_\oplus] Q[p_\oplus : p_\oplus]^{-1} A[p_\oplus]'$, at least one component of $Z_{p_\oplus}$ is equal to zero with probability one. However, $Z_{p_\oplus}$ equals Z_P

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

because of $p_{\oplus} \geq q^*$ and the definition of q^*. This leads to a contradiction because Z_P has the nonsingular variance-covariance matrix $\sigma^2 A Q^{-1} A'$. Without loss of generality, we may hence assume that $p_{\odot} = p_{\oplus}$.

In view of the discussion in the first paragraph of the proof of Lemma A.4, it suffices to establish, for each possible value s in the range $\mathcal{O} \leq s \leq p_{\odot}$, the result (A.6) for some θ with $s = \max\{p_0(\theta), \mathcal{O}\} = p_*$. Now fix such an s and θ (and also, of course, t, Q, A, σ, and the critical values c_p for $\mathcal{O} < p \leq P$). Then (A.7) expresses the map $\gamma \mapsto G_{\infty,\theta,\sigma,\gamma}(t)$ in terms of $\nu = (\nu_1, \ldots, \nu_P)'$. It is easy to see that the correspondence between γ and ν is a linear bijection from $\mathbb{R}^P$ onto itself and that $\gamma \in M_{q^*}$ if and only if $\nu \in M_{q^*}$. It is hence sufficient to find a $\delta_* > 0$ and vectors ν and μ in M_{q^*} such that (A.7) with $\nu + \epsilon\mu$ in place of ν and (A.7) with $\nu - \epsilon\mu$ in place of ν differ by at least δ_* for sufficiently small $\epsilon > 0$. Note that (A.7) is the sum of $P - p_* + 1$ terms indexed by $p = p_*, \ldots, P$. We shall now show that ν and μ can be chosen in such a way that, when replacing ν with $\nu + \epsilon\mu$ and $\nu - \epsilon\mu$, respectively, (i) the resulting terms in (A.7) corresponding to $p = p_{\odot}$ differ by some $d > 0$ and (ii) the difference of the other terms becomes arbitrarily small, provided that $\epsilon > 0$ is sufficiently small.

Consider first the case where $s = p_* = p_{\odot}$. Using the shorthand notation

$$g(\nu) = \mathbb{P}\left(Z_{p_{\odot}} \leq t + \sum_{r=p_{\odot}+1}^{q^*} \xi_{\infty,r}^{-2} C_{\infty}^{(r)} \nu_r\right),$$

note that the $p_{\odot}$th term in (A.7) is given by $g(\nu)$ multiplied by a product of positive probabilities that are continuous in ν. To prove property (i) it thus suffices to find a constant $c > 0$ and vectors ν and μ in M_{q^*} such that $|g(\nu + \epsilon\mu) - g(\nu - \epsilon\mu)| \geq c$ holds for each sufficiently small $\epsilon > 0$.

In the subcase $p_{\odot} = 0$ choose $c = 1$ and set

$$\nu = -[C_{\infty}^{(1)}, \ldots, C_{\infty}^{(P)}]' \left[\sum_{r=1}^{P} \xi_{\infty,r}^{-2} C_{\infty}^{(r)} C_{\infty}^{(r)\prime}\right]^{-1} t$$

and

$$\mu = [C_{\infty}^{(1)}, \ldots, C_{\infty}^{(P)}]' \left[\sum_{r=1}^{P} \xi_{\infty,r}^{-2} C_{\infty}^{(r)} C_{\infty}^{(r)\prime}\right]^{-1} (1, \ldots, 1)',$$

observing that the matrix to be inverted is indeed nonsingular because—as discussed after Lemma A.5—it is up to a multiplicative factor σ^2 identical to the variance-covariance matrix $\sigma^2 A Q^{-1} A'$ of Z_P. But then ν and μ satisfy $\sum_{r=p_{\odot}+1}^{q^*} \xi_{\infty,r}^{-2} C_{\infty}^{(r)} \nu_r = -t$ and $\sum_{r=p_{\odot}+1}^{q^*} \xi_{\infty,r}^{-2} C_{\infty}^{(r)} \mu_r = (1, \ldots, 1)'$ if we note that by the definition of q^*

$$\sum_{r=p_{\odot}+1}^{q^*} \xi_{\infty,r}^{-2} C_{\infty}^{(r)} \nu_r = \sum_{r=1}^{P} \xi_{\infty,r}^{-2} C_{\infty}^{(r)} \nu_r$$

holds and that a similar relation holds with μ replacing ν. Because $Z_{p_{\odot}} = Z_0 = 0 \in \mathbb{R}^k$, it is then obvious that $g(\nu + \epsilon\mu)$ and $g(\nu - \epsilon\mu)$ differ by one for each $\epsilon > 0$.

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

In the other subcase $p_\odot > 0$, define $M = [\xi_{\infty,1}^{-2} C_\infty^{(1)}, \ldots, \xi_{\infty,p_\odot}^{-2} C_\infty^{(p_\odot)}]$, $N = [\xi_{\infty,p_\odot+1}^{-2} C_\infty^{(p_\odot+1)}, \ldots, \xi_{\infty,q^*}^{-2} C_\infty^{(q^*)}]$, and $V = (W_1, \ldots, W_{p_\odot})'$. It is then easy to see that $g(\nu)$ equals $f((\nu_{p_\odot+1}, \ldots, \nu_{q^*})')$, with f defined as in Lemma A.2, and that M has a row of zeros. Furthermore, the matrix $(M:N)$ has rank k by the same argument as in the proof of Lemma A.4; cf. (A.10). By Lemma A.2, we thus obtain vectors x_0 and z, and a $c > 0$ such that $|f(x_0 + \epsilon z) - f(x_0 - \epsilon z)| \geq c$ holds for each sufficiently small $\epsilon > 0$. Setting $(\nu_{p_\odot+1}, \ldots, \nu_{q^*})' = x_0$, $(\mu_{p_\odot+1}, \ldots, \mu_{q^*})' = z$, setting $\nu[\neg q^*]$ and $\mu[\neg q^*]$ each equal to zero, and setting $\nu[p_\odot]$ and $\mu[p_\odot]$ to arbitrary values, we see that $g(\nu \pm \epsilon\mu)$ has the desired properties.

To complete the proof in the case $s = p_* = p_\odot$, we need to establish property (ii), for which it suffices to show that, for $p > p_\odot$, the pth term in (A.7) depends continuously on ν. For $p > q^*$, the pth term does not depend on ν, because $C_\infty^{(r)} = 0$ for $r = q^*, \ldots, P$. For p satisfying $p_\odot < p \leq q^*$, it suffices to show that

$$h(\nu_p, \ldots, \nu_{q^*}) = \mathbb{P}\left(Z_p \leq t + \sum_{r=p+1}^{q^*} \xi_{\infty,r}^{-2} C_\infty^{(r)} \nu_r, |W_p + \nu_p| \geq c_p \sigma \xi_{\infty,p}\right)$$

is a continuous function. Suppose that $(\nu_p^{(m)}, \ldots, \nu_{q^*}^{(m)})$ converges to $(\nu_p, \ldots, \nu_{q^*})$ as $m \to \infty$. For arbitrary $\alpha > 0$, $\sum_{r=p+1}^{q^*} \xi_{\infty,r}^{-2} C_\infty^{(r)} \nu_r$ and $\sum_{r=p+1}^{q^*} \xi_{\infty,r}^{-2} C_\infty^{(r)} \nu_r^{(m)}$ differ by less than α in each coordinate, provided that m is sufficiently large. This implies

$$\begin{aligned}
&\limsup_{m\to\infty} h(\nu_p^{(m)}, \ldots, \nu_{q^*}^{(m)}) \\
&\quad \leq \limsup_{m\to\infty} \mathbb{P}\left(Z_p \leq t + \sum_{r=p+1}^{q^*} \xi_{\infty,r}^{-2} C_\infty^{(r)} \nu_r + \alpha(1, \ldots, 1)', |W_p + \nu_p^{(m)}| \geq c_p \sigma \xi_{\infty,p}\right) \\
&\quad = \mathbb{P}\left(Z_p \leq t + \sum_{r=p+1}^{q^*} \xi_{\infty,r}^{-2} C_\infty^{(r)} \nu_r + \alpha(1, \ldots, 1)', |W_p + \nu_p| \geq c_p \sigma \xi_{\infty,p}\right),
\end{aligned}$$

observing that the latter probability is obviously continuous in the single variable ν_p (because W_p has an absolutely continuous distribution). Letting α decrease to zero we obtain $\limsup_{m\to\infty} h(\nu_p^{(m)}, \ldots, \nu_{q^*}^{(m)}) \leq h(\nu_p, \ldots, \nu_{q^*})$. A similar argument establishes $\liminf_{m\to\infty} h(\nu_p^{(m)}, \ldots, \nu_{q^*}^{(m)}) \geq \mathbb{P}(Z_p < t + \sum_{r=p+1}^{q^*} \xi_{\infty,r}^{-2} C_\infty^{(r)} \nu_r, |W_p + \nu_p| \geq c_p \sigma \xi_{\infty,p})$. The proof of the continuity of h is then complete if we can show that $\mathbb{P}(Z_p \leq \cdot, |W_p + \nu_p| \geq c_p \sigma \xi_{\infty,p})$ is continuous or, equivalently, that $\mathbb{P}(Z_p \leq \cdot \,\big|\, |W_p + \nu_p| \geq c_p \sigma \xi_{\infty,p})$ is a continuous cdf. Because $p > p_\odot$, the variance-covariance matrix $\sigma^2 A[p] Q[p:p]^{-1} A[p]'$ of Z_p only has nonzero diagonal elements. Consequently, when representing Z_p as $B(W_1, \ldots, W_p)'$, the matrix B cannot have rows that consist entirely of zeros. The conditional distribution of $(W_1, \ldots, W_p)'$ given the event $\{|W_p + \nu_p| \geq c_p \sigma \xi_{\infty,p}\}$ is clearly absolutely continuous with respect to p-dimensional Lebesgue measure. But then Lemma A.3 delivers the desired result.

The case where $s = p_* < p_\odot$ is reduced to the previously discussed case as follows: It is easy to see that, for $\nu_{p_\odot} \to \infty$, the expression in (A.7) converges to a limit uniformly with respect to all ν_p with $p \neq p_\odot$. Then observe that this limit is again of the form (A.7) but now with $p_\odot$ taking the rôle of p_*. ■

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

APPENDIX B: Nonuniformity of the Convergence of the Finite-Sample cdf to the Large-Sample Limit

PROPOSITION B.1.

(i) Suppose that $A\tilde{\theta}(q)$ and $\tilde{\theta}_q(q)$ are asymptotically correlated, i.e., $C_\infty^{(q)} \neq 0$, for some q satisfying $\mathcal{O} < q \leq P$, and let q^* denote the largest q with this property. Then for every $\theta \in M_{q^*-1}$, every σ, $0 < \sigma < \infty$, and every $t \in \mathbf{R}^k$ there exists a ρ, $0 < \rho < \infty$, such that

$$\liminf_{n\to\infty} \sup_{\substack{\vartheta \in M_{q^*} \\ \|\vartheta-\theta\| < \rho/\sqrt{n}}} |G_{n,\vartheta,\sigma}(t) - G_{\infty,\vartheta,\sigma}(t)| > 0 \tag{B.1}$$

holds. The constant ρ may be chosen in such a way that it depends only on t, Q, A, σ, and the critical values c_p for $\mathcal{O} < p \leq P$.

(ii) Suppose that $A\tilde{\theta}(q)$ and $\tilde{\theta}_q(q)$ are asymptotically uncorrelated, i.e., $C_\infty^{(q)} = 0$, for all q satisfying $\mathcal{O} < q \leq P$. Then $G_{n,\theta,\sigma}$ converges to $\Phi_{\infty,P}$ in total variation uniformly in $\theta \in \mathbf{R}^P$; more precisely

$$\sup_{\theta\in\mathbf{R}^P} \sup_{\substack{\sigma\in\mathbf{R} \\ \sigma_* \leq \sigma \leq \sigma^*}} \|G_{n,\theta,\sigma} - \Phi_{\infty,P}\|_{TV} \xrightarrow{n\to\infty} 0$$

holds for any constants σ_* and σ^* satisfying $0 < \sigma_* \leq \sigma^* < \infty$.

Under the assumptions of Proposition B.1(i), we see that convergence of $G_{n,\theta,\sigma}(t)$ to $G_{\infty,\theta,\sigma}(t)$ is nonuniform over shrinking "tubes" around M_{q^*-1} that are contained in M_{q^*}. (On the complement of a "tube" with a *fixed* positive radius, i.e., on the set $U = \{\theta \in \mathbf{R}^P : |\theta_{q^*}| \geq r\}$ with fixed $r > 0$, convergence of $G_{n,\theta,\sigma}(t)$ to $G_{\infty,\theta,\sigma}(t)$ is in fact uniform [even with respect to the total variation distance], as can be shown. Note that for $\theta \in U$ the cdf $G_{\infty,\theta,\sigma}(t)$ reduces to the Gaussian cdf $\Phi_{\infty,P}(t)$, i.e., to the asymptotic distribution of the least-squares estimator based on the overall model; cf. Remark A.1.) A precursor to Proposition B.1(i) is Corollary 5.5 of Leeb and Pötscher (2003), which establishes (B.1) in the special case where $\mathcal{O} = 0$ and where A is the $P \times P$ identity matrix. Proposition B.1(ii) describes an exceptional case where convergence is uniform. (In this case $G_{\infty,\theta,\sigma}$ reduces to the Gaussian cdf $\Phi_{\infty,P}$ for all θ and $\Phi_{\infty,P} = \Phi_{\infty,p}$, $\mathcal{O} \leq p \leq P$, holds; cf. Remark A.1.) Recall that under the assumptions of part (ii) of Proposition B.1 we necessarily always have (i) $\mathcal{O} > 0$ and (ii) rank $A[\mathcal{O}] = k$; cf. Proposition 4.4 in Leeb and Pötscher (2006b).

Proof of Proposition B.1. We first prove part (i). As noted at the beginning of the proof of Lemma A.4, the map $\gamma \mapsto G_{\infty,\theta,\sigma,\gamma}(t)$ depends only on t, Q, A, σ, the critical values c_p for $\mathcal{O} < p \leq P$, and also on θ, but the dependence on θ is only through $p_* = \max\{p_0(\theta), \mathcal{O}\}$. It hence suffices to find, for each possible value of p_* in the range $p_* = \mathcal{O}, \ldots, q^* - 1$, a constant $0 < \rho < \infty$ such that (B.1) is satisfied for some (and hence all) θ returning this particular value of $p_* = \max\{p_0(\theta), \mathcal{O}\}$. For this in turn it is sufficient to show that given such a θ we can find a $\gamma \in M_{q^*}$ such that

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

$$\liminf_{n\to\infty} |G_{n,\theta+\gamma/\sqrt{n},\sigma}(t) - G_{\infty,\theta+\gamma/\sqrt{n},\sigma}(t)| > 0 \tag{B.2}$$

holds. Note that (B.2) is equivalent to

$$\liminf_{n\to\infty} |G_{\infty,\theta,\sigma,\gamma}(t) - G_{\infty,\theta+\gamma/\sqrt{n},\sigma}(t)| > 0 \tag{B.3}$$

in light of Proposition 2.1. To establish (B.3), we proceed as follows: For each $\gamma \in M_{q^*}$ with $\gamma_{q^*} \neq 0$, $G_{\infty,\theta+\gamma/\sqrt{n},\sigma}(t)$ in (15) reduces to $\Phi_{\infty,q^*}(t)$ as is easily seen from (A.7) because $p_0(\theta + \gamma/\sqrt{n}) = q^*$, which in turn follows from $p_0(\theta) < q^*$ and $\gamma_{q^*} \neq 0$. Furthermore, Lemma A.4 entails that $G_{\infty,\theta,\sigma,\gamma}(t)$ is nonconstant in $\gamma \in M_{q^*} \backslash M_{q^*-1}$. But this shows that (B.3) must hold.

To prove part (ii), we write

$$\begin{aligned}\|G_{n,\theta,\sigma} - \Phi_{\infty,P}\|_{TV} &= \left\| \sum_{p=\mathcal{O}}^{P} G_{n,\theta,\sigma}(\cdot|p)\pi_{n,\theta,\sigma}(p) - \Phi_{\infty,P}(\cdot) \right\|_{TV} \\ &\leq \sum_{p=\mathcal{O}}^{P} \|G_{n,\theta,\sigma}(\cdot|p) - \Phi_{\infty,P}(\cdot)\|_{TV}\,\pi_{n,\theta,\sigma}(p),\end{aligned}$$

where the conditional cdfs $G_{n,\theta,\sigma}(\cdot|p)$ and the model-selection probabilities $\pi_{n,\theta,\sigma}(p)$ have been introduced after (12). By the "uncorrelatedness" assumption, we have that $\Phi_{\infty,p} = \Phi_{\infty,P}$ for all p in the range $\mathcal{O} \leq p \leq P$; cf. Remark A.1. We hence obtain

$$\begin{aligned}&\sup_{\theta\in\mathbf{R}^P} \sup_{\substack{\sigma\in\mathbf{R} \\ \sigma_*\leq\sigma\leq\sigma^*}} \|G_{n,\theta,\sigma} - \Phi_{\infty,P}\|_{TV} \\ &\quad\leq \sum_{p=\mathcal{O}}^{P} \sup_{\theta\in\mathbf{R}^P} \sup_{\substack{\sigma\in\mathbf{R} \\ \sigma_*\leq\sigma\leq\sigma^*}} \|G_{n,\theta,\sigma}(\cdot|p) - \Phi_{\infty,p}(\cdot)\|_{TV}\,\pi_{n,\theta,\sigma}(p).\end{aligned} \tag{B.4}$$

Now for every p with $\mathcal{O} \leq p \leq P$ and for every ρ, $0 < \rho < \infty$, we can write

$$\begin{aligned}&\sup_{\theta\in\mathbf{R}^P} \sup_{\substack{\sigma\in\mathbf{R} \\ \sigma_*\leq\sigma\leq\sigma^*}} \|G_{n,\theta,\sigma}(\cdot|p) - \Phi_{\infty,p}(\cdot)\|_{TV}\,\pi_{n,\theta,\sigma}(p) \\ &\quad\leq \max\left\{ \sup_{\substack{\theta\in\mathbf{R}^P \\ \|\theta[\neg p]\|<\rho/\sqrt{n}}} \sup_{\substack{\sigma\in\mathbf{R} \\ \sigma_*\leq\sigma\leq\sigma^*}} \|G_{n,\theta,\sigma}(\cdot|p) - \Phi_{\infty,p}(\cdot)\|_{TV},\ \sup_{\substack{\theta\in\mathbf{R}^P \\ \|\theta[\neg p]\|\geq\rho/\sqrt{n}}} \sup_{\substack{\sigma\in\mathbf{R} \\ \sigma_*\leq\sigma\leq\sigma^*}} \pi_{n,\theta,\sigma}(p) \right\}.\end{aligned} \tag{B.5}$$

In the case $p = P$, we use the convention that the second term in the maximum is absent and that the first supremum in the first term in the maximum extends over all of $\mathbb{R}^P$. Letting first n and then ρ go to infinity in (B.5), we may apply Lemmas C.2 and C.3 in Leeb and Pötscher (2005b) to conclude that the left-hand side of (B.5), and hence the left-hand side of (B.4), goes to zero as $n \to \infty$. ■

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

APPENDIX C: Proofs for Sections 2.1–2.2.2

In the proofs that follow it will be convenient to show the dependence of $\Phi_{n,p}$ and $\Phi_{\infty,p}$ on σ in the notation. Thus, in the following discussion we shall write $\Phi_{n,p,\sigma}$ and $\Phi_{\infty,p,\sigma}$, respectively, for the cdf of a k-variate Gaussian random vector with mean zero and variance-covariance matrix $\sigma^2 A[p](X[p]'X[p]/n)^{-1}A[p]'$ and $\sigma^2 A[p]Q[p:p]^{-1}A[p]'$, respectively. For convenience, let $\Phi_{n,0,\sigma}$ and $\Phi_{\infty,0,\sigma}$ denote the cdf of point mass at zero in $\mathbf{R}^k$.

The following lemma is elementary to prove if we recall that $b_{n,p}z$ converges to $b_{\infty,p}z$ as $n \to \infty$ for every $z \in \operatorname{Im} A[p]$, the column space of $A[p]$.

LEMMA C.1. *Suppose $p > \mathcal{O}$. Define $R_{n,p}(z,\sigma) = 1 - \Delta_{\sigma\zeta_{n,p}}(b_{n,p}z, c_p\sigma\xi_{n,p})$ and $R_{\infty,p}(z,\sigma) = 1 - \Delta_{\sigma\zeta_{\infty,p}}(b_{\infty,p}z, c_p\sigma\xi_{\infty,p})$ for $z \in \operatorname{Im} A[p]$, $0 < \sigma < \infty$. Let $\sigma^{(n)}$ converge to σ, $0 < \sigma < \infty$. If $\zeta_{\infty,p} \neq 0$, then $R_{n,p}(z,\sigma^{(n)})$ converges to $R_{\infty,p}(z,\sigma)$ for every $z \in \operatorname{Im} A[p]$; if $\zeta_{\infty,p} = 0$, then convergence holds for every $z \in \operatorname{Im} A[p]$, except possibly for $z \in \operatorname{Im} A[p]$ satisfying $|b_{\infty,p}z| = c_p\sigma\xi_{\infty,p}$. (This exceptional subset of $\operatorname{Im} A[p]$ has rank$(A[p])$-dimensional Lebesgue measure zero because $c_p\sigma\xi_{\infty,p} > 0$.)*

The following observation is useful in the proof of Proposition 2.2, which follows: Because the proposition depends on Y only through its distribution (cf. Remark 4.1), we may assume without loss of generality that the errors in (5) are given by $u_t = \sigma\varepsilon_t$, $t \in \mathbf{N}$, with i.i.d. ε_t that are standard normal. In particular, all random variables involved are then defined on the same probability space.

Proof of Proposition 2.2. Because $P_{n,\theta,\sigma}(\bar{p} = p_0(\theta)) \to 1$ by consistency, we may replace $\max\{\bar{p}, \mathcal{O}\}$ by $p_* = \max\{p_0(\theta), \mathcal{O}\}$ in the formula for $\breve{G}_n$ for the remainder of the proof. Furthermore, because $\hat{\sigma} \to \sigma$ in $P_{n,\theta,\sigma}$ probability, each subsequence contains a further subsequence along which $\hat{\sigma} \to \sigma$ almost surely (with respect to the probability measure on the common probability space supporting all random variables involved), and we restrict ourselves to such a further subsequence for the moment. In particular, we write $\{\hat{\sigma} \to \sigma\}$ for the event that $\hat{\sigma}$ converges to σ along the subsequence under consideration; clearly, the event $\{\hat{\sigma} \to \sigma\}$ has probability one. Also note that we can assume without loss of generality that $\hat{\sigma} > 0$ holds on this event (at least from some data-dependent n onward), because $\sigma > 0$ holds. But then obviously $\prod_{q=p_*+1}^{P} \Delta_{\hat{\sigma}\xi_{n,q}}(0, c_q\hat{\sigma}\xi_{n,q})$ converges to $\prod_{q=p_*+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(0, c_q\sigma\xi_{\infty,q})$, and $\hat{\Phi}_{n,p_*}(t)$ converges to $\Phi_{\infty,p_*,\sigma}(t)$ in total variation by Lemma A.3 of Leeb (2005) in the case $p_* > 0$ and trivially so in the case $p_* = 0$. This proves that the first term in the formula for $\breve{G}_n$ converges to the corresponding term in the formula for $G_{\infty,\theta,\sigma}$ in total variation.

Next, consider the term in $\breve{G}_n$ that carries the index $p > p_*$. By Lemma A.3 in Leeb (2005), $\hat{\Phi}_{n,p} = \Phi_{n,p,\hat{\sigma}}$ has a density $d\Phi_{n,p,\hat{\sigma}}/d\Phi_{\infty,p,\sigma}$ with respect to $\Phi_{\infty,p,\sigma}$, which converges to one except on a set that has measure zero under $\Phi_{\infty,p,\sigma}$. By Scheffé's lemma (Billingsley, 1995, Thm. 16.12), $d\Phi_{n,p,\hat{\sigma}}/d\Phi_{\infty,p,\sigma}$ converges to one also in the $L^1(\Phi_{\infty,p,\sigma})$ sense. By Lemma C.1, $R_{n,p}(z,\hat{\sigma})$ converges to $R_{\infty,p}(z,\sigma)$ except possibly on a set that has measure zero under $\Phi_{\infty,p,\sigma}$. (Recall that $\Phi_{\infty,p,\sigma}$ is concentrated on $\operatorname{Im} A[p]$ and is not degenerate there.) Observing that $|R_{n,p}(z,\hat{\sigma})|$ is uniformly bounded by one, we obtain that $R_{n,p}(z,\hat{\sigma})$ also converges to $R_{\infty,p}(z,\sigma)$ in the $L^1(\Phi_{\infty,p,\sigma})$ sense. Hence,

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

$$\left\| R_{n,p}(z,\hat{\sigma}) \frac{d\Phi_{n,p,\hat{\sigma}}}{d\Phi_{\infty,p,\sigma}}(z) - R_{\infty,p}(z,\sigma) \right\|$$

$$\leq \left\| R_{n,p}(z,\hat{\sigma}) \frac{d\Phi_{n,p,\hat{\sigma}}}{d\Phi_{\infty,p,\sigma}}(z) - R_{n,p}(z,\hat{\sigma}) \right\| + \|R_{n,p}(z,\hat{\sigma}) - R_{\infty,p}(z,\sigma)\|$$

$$\leq \left\| \frac{d\Phi_{n,p,\hat{\sigma}}}{d\Phi_{\infty,p,\sigma}}(z) - 1 \right\| + \|R_{n,p}(z,\hat{\sigma}) - R_{\infty,p}(z,\sigma)\| \xrightarrow{n\to\infty} 0, \tag{C.1}$$

where $\|\cdot\|$ denotes the $L^1(\Phi_{\infty,p,\sigma})$ norm. Because $\prod_{q=p+1}^{P} \Delta_{\hat{\sigma}\xi_{n,q}}(0, c_q\hat{\sigma}\xi_{n,q})$ obviously converges to $\prod_{q=p+1}^{P} \Delta_{\sigma\xi_{\infty,q}}(0, c_q\sigma\xi_{\infty,q})$, the relation (C.1) shows that the term in $\check{G}_n$ carrying the index p converges to the corresponding term in $G_{\infty,\theta,\sigma}$ in the total variation sense. This proves (18) along the subsequence under consideration. However, because any subsequence contains such a further subsequence, this establishes (18). Because $G_{n,\theta,\sigma}$ converges to $G_{\infty,\theta,\sigma}$ in total variation by Proposition 2.1, the claim in (17) also follows. ■

Before we prove the main result we observe that the total variation distance between $P_{n,\theta,\sigma}$ and $P_{n,\vartheta,\sigma}$ satisfies $\|P_{n,\theta,\sigma} - P_{n,\vartheta,\sigma}\|_{TV} \leq 2\Phi(\|\theta - \vartheta\|\lambda_{\max}^{1/2}(X'X)/2\sigma) - 1$; furthermore, if $\theta^{(n)}$ and $\vartheta^{(n)}$ satisfy $\|\theta^{(n)} - \vartheta^{(n)}\| = O(n^{-1/2})$, the sequence $P_{n,\vartheta^{(n)},\sigma}$ is contiguous with respect to the sequence $P_{n,\theta^{(n)},\sigma}$ (and vice versa). This follows exactly in the same way as Lemma A.1 in Leeb and Pötscher (2006a).

Proof of Theorem 2.3. We first prove (20) and (21). For this purpose we make use of Lemma 3.1 in Leeb and Pötscher (2006a) with $\alpha = \theta \in M_{q^*-1}$, $B = M_{q^*}$, $B_n = \{\vartheta \in M_{q^*} : \|\vartheta - \theta\| < \rho_0 n^{-1/2}\}$, $\beta = \vartheta$, $\varphi_n(\beta) = G_{n,\vartheta,\sigma}(t)$, $\hat{\varphi}_n = \hat{G}_n(t)$, where ρ_0, $0 < \rho_0 < \infty$, will be chosen shortly (and σ is held fixed). The contiguity assumption of this lemma (and also the mutual contiguity assumption used in the corrigendum to Leeb and Pötscher, 2006a) is satisfied in view of the preparatory remark preceding this proof. It hence remains only to show that there exists a value of ρ_0, $0 < \rho_0 < \infty$, such that δ^* in Lemma 3.1 of Leeb and Pötscher (2006a) (which represents the limit inferior of the oscillation of $\varphi_n(\cdot)$ over B_n) is positive. Applying Lemma 3.5(i) of Leeb and Pötscher (2006a) with $\zeta_n = \rho_0 n^{-1/2}$ and the set G_0 equal to the set G, it remains, in light of Proposition 2.1, to show that there exists a ρ_0, $0 < \rho_0 < \infty$, such that $G_{\infty,\theta,\sigma,\gamma}(t)$ as a function of γ is nonconstant on the set $\{\gamma \in M_{q^*} : \|\gamma\| < \rho_0\}$. In view of Lemma 3.1 of Leeb and Pötscher (2006a), the corresponding δ_0 can then be chosen as any positive number less than one-half of the oscillation of $G_{\infty,\theta,\sigma,\gamma}(t)$ over this set. That such a ρ_0 indeed exists now follows from Lemma A.4 in Appendix A, where it is also shown that ρ_0 and δ_0 can be chosen such that they depend only on t, Q, A, σ, and c_p for $\mathcal{O} < p \leq P$. This completes the proof of (20) and (21).

To prove (22) we use Corollary 3.4 in Leeb and Pötscher (2006a) with the same identification of notation as before, with $\zeta_n = \rho_0 n^{-1/2}$ and with $V = M_{q^*}$ (viewed as a vector space isomorphic to $\mathbf{R}^{q^*}$). The asymptotic uniform equicontinuity condition in that corollary is then satisfied in view of $\|P_{n,\theta,\sigma} - P_{n,\vartheta,\sigma}\|_{TV} \leq 2\Phi(\|\theta - \vartheta\|\lambda_{\max}^{1/2}(X'X)/2\sigma) - 1$. Given that the positivity of δ^* has already been established in the previous paragraph, applying Corollary 3.4(i) of Leeb and Pötscher (2006a) then establishes (22). ■

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press

Proof of Remark 2.1. The proof is similar to the proof of (22) just given, except for using Corollary 3.4(ii) and Lemma 3.5(ii) in Leeb and Pötscher (2006a) instead of Corollary 3.4(i) and Lemma 3.5(i) from that paper. Furthermore, Lemma A.5 in Appendix A instead of Lemma A.4 is used. ■

Proof of Proposition 2.4. In view of Proposition B.1(ii) and the fact that $\hat{\Phi}_{n,P}(\cdot) = \Phi_{n,P,\hat{\sigma}}(\cdot)$ holds (in the case $\hat{\sigma} > 0$), it suffices to show that

$$\sup_{\substack{\sigma \in \mathbf{R} \\ \sigma_* \le \sigma \le \sigma^*}} \|\Phi_{n,P,\sigma}(\cdot) - \Phi_{\infty,P,\sigma}(\cdot)\|_{TV} \xrightarrow{n\to\infty} 0 \tag{C.2}$$

$$\sup_{\substack{\sigma \in \mathbf{R} \\ \sigma_* \le \sigma \le \sigma^*}} P_{n,\theta,\sigma}(\|\Phi_{n,P,\hat{\sigma}}(\cdot) - \Phi_{n,P,\sigma}(\cdot)\|_{TV} > \delta) \xrightarrow{n\to\infty} 0 \tag{C.3}$$

hold for each $\delta > 0$ and for any constants σ_* and σ^* satisfying $0 < \sigma_* \le \sigma^* < \infty$. (Note that the probability in (C.3) in fact does not depend on θ.) But this has already been established in the proof of Proposition 4.3 of Leeb and Pötscher (2005b). ■

APPENDIX D: Proof for Section 3

Proof of Theorem 3.1. After rearranging the elements of θ (and hence the regressors) if necessary and then correspondingly rearranging the rows of the matrix A, we may assume without loss of generality that $\mathfrak{r}_* = (1, \ldots, 1, 0)$ and hence that $i(\mathfrak{r}_*) = P$. That is, $M_{\mathfrak{r}_*} = M_{P-1}$ and $M_{\mathfrak{r}_{full}} = M_P$. Furthermore, note that after this arrangement $C_\infty^{(P)} \neq 0$. Let $\hat{p}$ be the model-selection procedure introduced in Section 2 with $\mathcal{O} = P - 1$, $c_P = c$, and $c_{\mathcal{O}} = 0$. Let $\tilde{\theta}$ be the corresponding post-model-selection estimator and let $G_{n,\theta,\sigma}(t)$ be as defined in Section 2.1. Condition (24) now implies that for every $\theta \in M_{P-1}$ that has exactly $P - 1$ nonzero coordinates

$$\lim_{n\to\infty} P_{n,\theta,\sigma}(\{\hat{\mathfrak{r}} = \mathfrak{r}_{full}\} \blacktriangle \{\hat{p} = P\}) = \lim_{n\to\infty} P_{n,\theta,\sigma}(\{\hat{\mathfrak{r}} = \mathfrak{r}_*\} \blacktriangle \{\hat{p} = P - 1\}) = 0 \tag{D.1}$$

holds for every $0 < \sigma < \infty$. Because the sequences $P_{n,\vartheta^{(n)},\sigma}$ and $P_{n,\theta,\sigma}$ are contiguous for $\vartheta^{(n)}$ satisfying $\|\theta - \vartheta^{(n)}\| = O(n^{-1/2})$ as remarked prior to the proof of Theorem 2.3 in Appendix C, it follows that condition (D.1) continues to hold with $P_{n,\vartheta^{(n)},\sigma}$ replacing $P_{n,\theta,\sigma}$. This implies that for every sequence of positive real numbers s_n with $s_n = O(n^{-1/2})$, for every σ, $0 < \sigma < \infty$, and for every $\theta \in M_{P-1}$ that has exactly $P - 1$ nonzero coordinates

$$\sup_{\substack{\vartheta \in \mathbf{R}^P \\ \|\vartheta - \theta\| < s_n}} \|K_{n,\vartheta,\sigma} - G_{n,\vartheta,\sigma}\|_{TV} \to 0 \tag{D.2}$$

holds as $n \to \infty$. From (D.2) we conclude that the limit of $K_{n,\theta+\gamma/\sqrt{n},\sigma}$ (with respect to total variation distance) exists and coincides with $G_{\infty,\theta,\sigma,\gamma}$. Repeating the proof of Theorem 2.3 with $q^* = P$, with $K_{n,\vartheta,\sigma}(t)$ replacing $G_{n,\vartheta,\sigma}(t)$ and with $\hat{K}_n(t)$ replacing $\hat{G}_n(t)$ gives the desired result. ■

https://doi.org/10.1017/S0266466608080158 Published online by Cambridge University Press