Econometric Theory, **24**, 2008, 323–337. Printed in the United States of America.
DOI: 10.1017/S0266466608080146

SHRINKAGE ESTIMATION FOR NEARLY SINGULAR DESIGNS

KEITH KNIGHT
University of Toronto

Shrinkage estimation procedures such as ridge regression and the lasso have been proposed for stabilizing estimation in linear models when high collinearity exists in the design. In this paper, we consider asymptotic properties of shrinkage estimators in the case of "nearly singular" designs.

1. INTRODUCTION

Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$
$$= \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (i.i.d.) random variables with mean 0 and variance σ^2. For simplicity, we will assume that the predictors are centered to have mean 0 and that the intercept β_0 is always estimated by $\overline{Y}$. This assumption allows us to focus on estimation of $\beta_1, \ldots, \beta_p$, but it is not essential.

Throughout this paper, we will assume that the $\boldsymbol{x}_i$'s are nearly collinear in the sense that the matrix

$$C_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{2}$$

is nonsingular for each n but that

$$C_n \to C, \tag{3}$$

where C is singular; we will refer to such designs as "nearly singular."

I thank Hannes Leeb and Benedikt Pötscher and also the referees for their valuable comments. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Address correspondence to Keith Knight, Department of Statistics, University of Toronto, 100 St. George St., Toronto, ON M5S 3G3, Canada; e-mail: keith@utstat.toronto.edu.

 0266-4666/08 $15.00

The exact definition of near singularity (which will be given in the next section) is an asymptotic one, but in practice, a nearly singular design might be characterized as one where the smallest eigenvalue (or eigenvalues) of C_n is small compared to the trace of C_n. In some cases, the near singularity is, in fact, a consequence of the model; see, for example, Phillips (2001). It is well known that ordinary least squares (OLS) estimation, although unbiased, leads to parameter estimates with large variance. Several alternative methods, which trade bias for variance, have been proposed to deal with this problem; these methods include ridge regression (Hoerl and Kennard, 1970), partial least squares (Wold, 1984; Lorber, Wanger, and Kowalski, 1987), continuum regression (Stone and Brooks, 1990), the "lasso" (Tibshirani, 1996; Radchenko, 2004), and the smooth clipped absolute deviation (SCAD) penalty of Fan and Li (2001).

Under the condition

$$\max_{1\le i\le n} \boldsymbol{x}_i^T \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \boldsymbol{x}_i \to 0 \quad \text{as } n \to \infty, \tag{4}$$

the OLS estimator, which we will denote by $\hat{\boldsymbol{\beta}}_n^{(0)}$, is asymptotically normal; more precisely, we have

$$\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{1/2} (\hat{\boldsymbol{\beta}}_n^{(0)} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 I) \tag{5}$$

(Srivastava, 1971). Note that the condition (4) can be rewritten as

$$\frac{1}{n} \max_{1\le i\le n} \boldsymbol{x}_i C_n^{-1} \boldsymbol{x}_i \to 0,$$

which if C_n tends to a nonsingular matrix C is equivalent to

$$\frac{1}{n} \max_{1\le i\le n} \boldsymbol{x}_i^T \boldsymbol{x}_i \to 0;$$

moreover, if C is nonsingular then the asymptotic normality in (5) can be expressed as

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(0)} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 C^{-1}). \tag{6}$$

The convergence in (5) is very general and quite useful in practice even in the case of nearly singular designs; on the other hand, generalizing (5) to estimators obtained after shrinkage procedures (such as ridge regression or the lasso) or automatic model selection procedures (such as the Akaike information criterion [AIC]) is difficult. Results such as (6) (where the normalization is by a

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

sequence of constants rather than a sequence of matrices) turn out to be easier to obtain and can give considerable insight into the properties (from a large-sample perspective) of the particular estimator.

2. PENALIZED LEAST SQUARES ESTIMATION

Regularization is a frequently used technique in statistics for obtaining estimators in situations where standard estimators are unstable or otherwise poorly defined.

We will consider estimating $\boldsymbol{\beta}$ by minimizing the penalized least squares (LS) criterion

$$\sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi})^2 + \lambda_n \sum_{j=1}^{p} |\phi_j|^\gamma \tag{7}$$

for a given λ_n where $\gamma > 0$; the resulting estimator will be denoted throughout by $\hat{\boldsymbol{\beta}}_n$, thereby suppressing its dependence on both γ and λ_n with $\hat{\boldsymbol{\beta}}_n^{(0)}$ denoting the OLS estimator (with $\lambda_n = 0$). These so-called Bridge estimators were introduced by Frank and Friedman (1993) as a generalization of ridge regression (which occurs for $\gamma = 2$). The special case when $\gamma = 1$ corresponds to the lasso (Tibshirani, 1996). Properties of these estimators have been studied by, among others, Fu (1998), Knight and Fu (2000), Radchenko (2004), and Leeb and Pötscher (2006). For $\gamma \leq 1$, the estimators minimizing (7) have the potentially attractive feature of being exactly 0 if λ_n is sufficiently large, thus combining parameter estimation and model selection; indeed model selection methods such as AIC and the Bayesian information criterion (BIC) can be viewed as limiting cases as $\gamma \to 0$. Also note that when $\gamma < 1$, the objective function (7) is not convex and the estimator $\hat{\boldsymbol{\beta}}_n$ can be quite sensitive to the choice of λ_n; more precisely, when $\gamma < 1$, the mapping from λ_n to $\hat{\boldsymbol{\beta}}_n$ will have jump discontinuities. The SCAD penalty of Fan and Li (2001) is a nonconvex penalty (indexed by two parameters) that combines the features of a lasso-type penalty (for small parameter values) with an AIC-type penalty (for larger parameter values).

We could also replace (7) by a penalized LS criterion that allows us a separate tuning parameter for each coefficient:

$$\sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi})^2 + \sum_{j=1}^{p} \lambda_n^{(j)} |\phi_j|^\gamma. \tag{8}$$

It is straightforward to generalize the results of this paper to estimators obtained by minimizing (8). The objective function (7) is more common in practice; typically, the predictors are scaled to have a variance of 1.

In this section, we will consider the asymptotic behavior of Bridge estimators when the design is nearly singular. More precisely, suppose that C_n (as

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

defined in (3)) is nonsingular but tends to a singular matrix C. In particular, we will assume that

$$a_n(C_n - C) \to D_0 \tag{9}$$

for some sequence $\{a_n\}$ tending to infinity where D_0 is positive definite on the null space of C (i.e., $\boldsymbol{v}^T D_0 \boldsymbol{v} > 0$ for nonzero $\boldsymbol{v}$ with $C\boldsymbol{v} = \mathbf{0}$). Note that D_0 is necessarily nonnegative definite on the null space of C so that it is not too stringent to require it to be positive definite on this null space. If D_0 is not positive definite on the null space of C then we can modify (9) to obtain appropriate limiting distributions; this will be considered in the next section. We are also assuming (at least implicitly) that the near singularity affects all the predictors in the model. Applications where the condition (9) holds are given in Phillips (2001) and Gabaix and Ibragimov (2006). A referee has also pointed out a possible connection to the problem of weak instruments (cf. Stock, Wright, and Yogo, 2002), for example, in two-stage least squares estimation. Caner (2004, 2006) considers nearly singular designs in the context of generalized method of moments (GMM) estimation.

To obtain consistency and limiting distributions for $\hat{\boldsymbol{\beta}}_n$ minimizing (7), we need to impose conditions on the sequence $\{\lambda_n\}$ so that it does not grow too quickly. In the case where C is nonsingular, Knight and Fu (2000) showed that to obtain nondegenerate limiting distributions for $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$, we require $\lambda_n/\sqrt{n} \to \lambda_0$ for $\gamma \geq 1$ and $\lambda_n/n^{\gamma/2} \to \lambda_0$ for $\gamma < 1$; for nearly singular designs, the growth criterion for $\{\lambda_n\}$ will be somewhat more stringent.

It is worth mentioning that asymptotic results tend to undersell the value of shrinkage estimation in practice. The reason for this is simple. Shrinkage is used in practice to reduce the variability in the estimation of parameters that are "small" by forcing their estimates toward 0 (or setting them to 0). However, from an asymptotic perspective, the only parameters that are "small" are those that are exactly 0 as all other parameters can be (with probability tending to 1 as $n \to \infty$) distinguished as different than 0. Thus for a parameter β_k whose value is nonzero, shrinkage generally produces bias in the resulting estimator that may or may not vanish asymptotically, and the resulting asymptotic bias is typically not compensated by a reduction in the asymptotic variance. On the other hand, if $\beta_k = 0$ then shrinkage will typically reduce the asymptotic variance of the estimator (without any asymptotic bias), which leads to a sort of superefficiency in these cases. Obviously, it is desirable to produce estimators that have no asymptotic bias when $\beta_k \neq 0$ and are superefficient when $\beta_k = 0$; such estimators exist but can be extremely sensitive to small perturbations in the data or changes in the choice of tuning parameters. The asymptotic results are very useful in giving insight into how sensitive a given methodology is to the choice of tuning parameters.

A useful tool in the development of the asymptotic distribution of the penalized LS estimators is the notion of epi-convergence in distribution, which is

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

discussed in Pflug (1995), Geyer (1994, 1996), and Knight (1999). A sequence of random lower semicontinuous functions $\{Z_n\}$ on $\Re^p$ (taking values in $[-\infty,\infty]$) epi-converges in distribution to Z ($Z_n \xrightarrow{e-d} Z$) if for closed rectangles $R_1,\ldots,R_k$ with open interiors $R_1^o,\ldots,R_k^o$, we have

$$
\begin{aligned}
&P\left[\inf_{\mathbf{u}\in R_1} Z(\boldsymbol{u}) > a_1,\ldots,\inf_{\mathbf{u}\in R_k} Z(\boldsymbol{u}) > a_k\right] \\
&\quad\le \liminf_{n\to\infty} P\left[\inf_{\mathbf{u}\in R_1} Z_n(\boldsymbol{u}) > a_1,\ldots,\inf_{\mathbf{u}\in R_k} Z_n(\boldsymbol{u}) > a_k\right] \\
&\quad\le \limsup_{n\to\infty} P\left[\inf_{\mathbf{u}\in R_1^o} Z_n(\boldsymbol{u}) > a_1,\ldots,\inf_{\mathbf{u}\in R_k^o} Z_n(\boldsymbol{u}) > a_k\right] \\
&\quad\le P\left[\inf_{\mathbf{u}\in R_1^o} Z(\boldsymbol{u}) > a_1,\ldots,\inf_{\mathbf{u}\in R_k^o} Z(\boldsymbol{u}) > a_k\right]
\end{aligned}
$$

for all real $a_1,\ldots,a_k$. Epi-convergence in distribution is particularly useful for studying estimators that minimize (or maximize) objective functions subject to constraints and also estimators that minimize discontinuous (but lower semicontinuous) objective functions; the best known weak convergence for functions, which is based on uniform convergence on compact sets (van der Vaart and Wellner, 1996), is poorly suited to these types of objective functions. However, this type of weak convergence, when applicable, does imply epi-convergence in distribution.

The limiting distributions of "argmin" estimators can often be determined via epi-convergence of the associated objective functions; in particular, if

$$Z_n(\boldsymbol{U}_n) = \inf_{\mathbf{u}} Z_n(\boldsymbol{u}) + o_p(1)$$

and $Z_n \xrightarrow{e-d} Z$ where Z has a unique minimizer $\boldsymbol{U}$ then $\boldsymbol{U}_n \xrightarrow{d} \boldsymbol{U}$ provided that $\boldsymbol{U}_n = O_p(1)$. For an application of epi-convergence in distribution in the context of estimation in nonregular econometric models, see Chernozhukov and Hong (2004) and Chernozhukov (2005).

In the case where $\{Z_n\}$ are convex with $Z_n \xrightarrow{e-d} Z$ (where the minimizer of Z, $\boldsymbol{U}$, is unique) then the condition $\boldsymbol{U}_n = O_p(1)$ is guaranteed, and so $\boldsymbol{U}_n \xrightarrow{d} \boldsymbol{U}$. Moreover, in the case of convexity, finite-dimensional weak convergence of $\{Z_n\}$ to Z is sufficient for $Z_n \xrightarrow{e-d} Z$ provided that Z is finite (with probability 1) on an open set (Geyer, 1996); however, for nearly singular designs, this latter condition is not satisfied as the appropriate limiting objective function is finite only on a lower dimensional subspace of $\Re^p$. Finite-dimensional weak convergence implies epi-convergence in distribution if $\{Z_n\}$ is stochastically equi–lower semicontinuous as defined in Knight (1999).

We will now consider the asymptotic behavior of nearly singular designs under fairly weak conditions. We will assume that C_n is nonsingular for all n and satisfies (9) for some sequence $\{a_n\}$. Define $b_n = (n/a_n)^{1/2}$ and define Z_n to be

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

$$Z_n(\boldsymbol{u}) = \sum_{i=1}^{n} [(\varepsilon_i - \boldsymbol{u}^T \boldsymbol{x}_i / b_n)^2 - \varepsilon_i^2] + \lambda_n \sum_{j=1}^{p} (|\beta_j + u_j / b_n|^\gamma - |\beta_j|^\gamma). \quad \textbf{(10)}$$

If $\hat{\boldsymbol{\beta}}_n$ minimizes (7) then the minimizer of (10) is simply $b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$; the objective function Z_n in (10) is simply a rescaled version of the objective function (7) with constants subtracted to ensure convergence. Note that because $b_n = o(\sqrt{n})$, the estimators will have a slower rate of convergence than when C is nonsingular.

The following result was given in Knight and Fu (2000).

THEOREM 1. *Assume the linear model (1) where C_n in (2) satisfies (3), (4), and (9) where C is singular and D_0 is positive definite on the null space of C. Define $\boldsymbol{W}$ to be a 0 mean multivariate normal random vector such that $\mathrm{Var}(\boldsymbol{u}^T \boldsymbol{W}) = \sigma^2 \boldsymbol{u}^T D_0 \boldsymbol{u} > 0$ for each nonzero $\boldsymbol{u}$ satisfying $C\boldsymbol{u} = \boldsymbol{0}$. Let $\hat{\boldsymbol{\beta}}_n$ minimize (7) for some $\gamma > 0$ and $\lambda_n \geq 0$.*

(i) If $\gamma > 1$ and $\lambda_n / b_n \to \lambda_0 \geq 0$ then

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \mathrm{argmin}\{Z(\boldsymbol{u}) : C\boldsymbol{u} = \boldsymbol{0}\},$$

where

$$Z(\boldsymbol{u}) = -2\boldsymbol{u}^T \boldsymbol{W} + \boldsymbol{u}^T D_0 \boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} u_j \,\mathrm{sgn}(\beta_j) |\beta_j|^{\gamma - 1}.$$

(ii) If $\gamma = 1$ and $\lambda_n / b_n \to \lambda_0 \geq 0$ then

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \mathrm{argmin}\{Z(\boldsymbol{u}) : C\boldsymbol{u} = \boldsymbol{0}\},$$

where

$$Z(\boldsymbol{u}) = -2\boldsymbol{u}^T \boldsymbol{W} + \boldsymbol{u}^T D_0 \boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} \{u_j \,\mathrm{sgn}(\beta_j) + |u_j| I(\beta_j = 0)\}.$$

(iii) If $\gamma < 1$ and $\lambda_n / b_n^\gamma \to \lambda_0 \geq 0$ then

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \mathrm{argmin}\{Z(\boldsymbol{u}) : C\boldsymbol{u} = \boldsymbol{0}\},$$

where

$$Z(\boldsymbol{u}) = -2\boldsymbol{u}^T \boldsymbol{W} + \boldsymbol{u}^T D_0 \boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} |u_j|^\gamma I(\beta_j = 0).$$

Proof. Define Z_n as in (10). First of all, we must show in each case that $Z_n \xrightarrow{e-d} Z_0$ where $Z_0(\boldsymbol{u}) = Z(\boldsymbol{u})$ for $\boldsymbol{u}$ satisfying $C\boldsymbol{u} = \boldsymbol{0}$ and $Z_0(\boldsymbol{u}) = \infty$ otherwise. This follows by first showing finite-dimensional weak convergence of

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

$\{Z_n\}$ to Z_0 ($Z_n \xrightarrow{f-d} Z_0$) and then stochastic equi–lower semicontinuity (e-1-sc) (Knight, 1999) of $\{Z_n\}$; note that, because Z_0 is not finite on an open set, $Z_n \xrightarrow{f-d} Z_0$ is not sufficient for $Z_n \xrightarrow{e-d} Z_0$ even when Z_n is convex (i.e., when $\gamma \geq 1$). Finally, we must show that

$$\operatorname{argmin}_{\mathbf{u}} Z_n(\boldsymbol{u}) = b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = O_p(1).$$

When $\gamma \geq 1$, this holds automatically from the convexity of the Z_n's; for $\gamma < 1$, it can be established by noting that the quadratic part of Z_n is growing faster (in $\|\boldsymbol{u}\|$) than the nonconvex penalty. ■

Note that for $\gamma \geq 1$, the limiting distribution will typically depend on λ_0 whereas for $\gamma < 1$, this is only true if at least one of the β_j's is 0. However, if $\gamma < 1$ and at least one β_j is 0 then the mapping from λ_0 to the limiting distribution will have discontinuities; this mapping is continuous for $\gamma \geq 1$ because of the convexity (in $\boldsymbol{u}$) of the limiting objective function V for any λ_0.

The condition on λ_n in part (iii) of Theorem 1 can be modified to achieve the "best of both worlds" for $\gamma < 1$, that is, no asymptotic bias for estimators of nonzero parameters and superefficiency for estimators of zero parameters. We do this by assuming that $\lambda_n / b_n^{\tau} \to \lambda_0 > 0$ where $\gamma < \tau < 1$. Although this seems attractive, it should be noted that this is an asymptotic condition and does not really give much insight regarding the choice of λ_n for fixed n.

Example 1

Consider a design with p predictors with common mutual correlation ρ_n. Assuming the predictors are normalized to have variance 1, we have

$$C_n = \begin{pmatrix} 1 & \rho_n & \cdots & \rho_n \\ \rho_n & 1 & \cdots & \rho_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_n & \cdots & \rho_n & 1 \end{pmatrix};$$

we will assume that $\rho_n \to 1$ and $a_n(1 - \rho_n) \to \psi > 0$. In this case, $\{C_n\}$ converges to a matrix C (of all 1's) and $a_n(C_n - C) \to D_0$ where

$$D_0 = \begin{pmatrix} 0 & -\psi & \cdots & -\psi \\ -\psi & 0 & \cdots & -\psi \\ \vdots & \vdots & \ddots & \vdots \\ -\psi & \cdots & -\psi & 0 \end{pmatrix}.$$

(In this example, the form of D_0 is not particularly important.) If the matrices are $p \times p$ then the null space of C is the space of vectors $\boldsymbol{u}$ with $u_1 + \cdots +$

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

$u_p = 0$. For the sake of illustration, let us suppose that $\beta_1, \ldots, \beta_p$ are all nonzero and take $\gamma \geq 1$. Then the limiting objective function Z in Theorem 1 is

$$Z(\boldsymbol{u}) = -2\boldsymbol{u}^T\boldsymbol{W} + \boldsymbol{u}^T D_0 \boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j)|\beta_j|^{\gamma-1} \quad \text{for } u_1 + \cdots + u_p = 0, \quad \textbf{(11)}$$

where

$$\lambda_0 = \lim_{n\to\infty} \lambda_n \left(\frac{a_n}{n}\right)^{1/2}.$$

By Theorem 1, we have (setting $b_n = (n/a_n)^{1/2}$)

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \operatorname{argmin}\{Z(\boldsymbol{u}) : u_1 + \cdots + u_p = 0\}.$$

It is interesting to compare this limiting distribution to the limiting distribution of the OLS estimator:

$$b_n(\hat{\boldsymbol{\beta}}_n^{(0)} - \boldsymbol{\beta}) \xrightarrow{d} \operatorname{argmin}\{Z_0(\boldsymbol{u}) : u_1 + \cdots + u_p = 0\},$$

where Z_0 is simply Z in (11) setting $\lambda_0 = 0$. The size of the asymptotic bias of $\hat{\boldsymbol{\beta}}_n$ relative to $\hat{\boldsymbol{\beta}}_n^{(0)}$ (which is unbiased) depends on the coefficients of $u_1, \ldots, u_p$ in the penalty

$$\lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j)|\beta_j|^{\gamma-1}.$$

Note that these coefficients are bounded (in $\boldsymbol{\beta}$) only if $\gamma = 1$ (the lasso) and that the bias vanishes for $\gamma > 1$ (i.e., $\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n^{(0)} = o_p(b_n^{-1})$) if, and only if, $\beta_1 = \cdots = \beta_p$ whereas for $\gamma = 1$, this same condition holds under the weaker condition $\operatorname{sgn}(\beta_1) = \cdots = \operatorname{sgn}(\beta_p)$. It should be noted also that the preceding discussion does not depend on the form of the matrix D_0.

Next suppose that $\beta_1 \neq 0$ and $\beta_2 = \cdots = \beta_p = 0$. If $\gamma \leq 1$ and $\lambda_0 > 0$ then the joint limiting distribution of the estimators of $\beta_2, \ldots, \beta_p$ will have positive probability mass at $\mathbf{0}$ and because the limiting distribution lies in the null space of C, this implies that the limiting distribution of $b_n(\hat{\beta}_{n1} - \beta_1)$ has positive probability mass at 0.

Theorem 1 can be extended to model selection methods such as AIC and BIC. Suppose that $\hat{\boldsymbol{\beta}}_n$ minimizes

$$n \ln\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\phi})^2\right] + \lambda_n \sum_{j=1}^{p} I(\phi_k \neq 0); \quad \textbf{(12)}$$

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

for AIC, $\lambda_n = 2$ whereas for BIC, we have $\lambda_n = \ln(n)$. The following result gives the limiting distribution in AIC-like situations where $\lambda_n \to \lambda_0 \in (0,\infty)$.

THEOREM 2. *Assume the linear model (1) where C_n in (2) satisfies (3), (4), and (9) where C is singular and D_0 is positive definite on the null space of C. Suppose that $\hat{\boldsymbol{\beta}}_n$ minimizes (12) and $\lambda_n \to \lambda_0 \geq 0$. Then*

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \operatorname{argmin}\{Z(\boldsymbol{u}) : C\boldsymbol{u} = \mathbf{0}\},$$

where

$$Z(\boldsymbol{u}) = \frac{1}{\sigma^2}(\boldsymbol{u}^T D_0 \boldsymbol{u} - 2\boldsymbol{u}^T \boldsymbol{W}) + \lambda_0 \sum_{j=1}^{p} [I(\beta_j \neq 0) + I(u_j \neq 0) I(\beta_j = 0)]$$

with $\boldsymbol{u}^T \boldsymbol{W} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{u}^T D_0 \boldsymbol{u})$ for $\boldsymbol{u}$ in the null space of C.

Proof. Define the objective function

$$Z_n(\boldsymbol{u}) = n \ln\left[\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \boldsymbol{x}_i^T \boldsymbol{u}/b_n)^2\right] - n \ln\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right) + \lambda_n \sum_{j=1}^{p} I(\beta_j + u_j/b_n \neq 0) \tag{13}$$

and note that it is minimized at $\boldsymbol{u} = b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. First of all, outside of the null space of C, it is easy to see that $Z_n(\boldsymbol{u}) \xrightarrow{p} \infty$. For $\boldsymbol{u}$ in the null space of C, we have

$$\begin{aligned} & n \ln\left[\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \boldsymbol{x}_i^T \boldsymbol{u}/b_n)^2\right] - n \ln\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right) \\ &\quad = \frac{1}{\sigma^2}\left[\frac{1}{b_n^2}\sum_{i=1}^{n}(\boldsymbol{x}_i^T \boldsymbol{u})^2 - \frac{2}{b_n}\sum_{i=1}^{n}\boldsymbol{x}_i^T \boldsymbol{u}\varepsilon_i\right] + o_p(1) \\ &\quad \xrightarrow{f-d} \frac{1}{\sigma^2}(\boldsymbol{u}^T D_0 \boldsymbol{u} - 2\boldsymbol{u}^T \boldsymbol{W}). \end{aligned}$$

For the penalty term,

$$\lambda_n \sum_{j=1}^{p} I(\beta_j + u_j/b_n \neq 0) \to \lambda_0 \sum_{j=1}^{p} [I(\beta_j \neq 0) + I(u_j \neq 0) I(\beta_j = 0)].$$

Thus we have $Z_n \xrightarrow{f-d} Z$. Epi-convergence in distribution follows by establishing e-l-sc (Knight, 1999); we need to show that for each bounded set B, $\epsilon > 0$, and $\delta > 0$, there exist $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m \in B$ and open neighborhoods $O(\boldsymbol{u}_1), \ldots, O(\boldsymbol{u}_m)$ such that

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

$$B \subset \bigcup_{i=1}^{m} O(\boldsymbol{u}_i)$$

and

$$\limsup_{n\to\infty} P\left\{\bigcup_{i=1}^{m}\left[\inf_{\mathbf{u}\in O(\mathbf{u}_i)} Z_n(\boldsymbol{u}) \le \min(\epsilon^{-1}, Z_n(\boldsymbol{u}_i) - \epsilon)\right]\right\} < \delta.$$

First of all, note that Z_n is finite for each n with discontinuities at points that do not depend on n. If B does not intersect the null space of C then e-l-sc is straightforward; for each n, Z_n is approximately a quadratic function that is tending to $+\infty$. On the other hand, if B does intersect the null space of C then we can take $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ to lie in this null space and obtain the desired inequality. It remains only to establish that $b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = O_p(1)$; this follows because for $n >$ some n_0, we have $Z_n(\mathbf{0}) \le (\lambda_0 + \epsilon)p$, and there exists a compact set K_ϵ such that

$$\limsup_{n\to\infty} P\left[\inf_{\mathbf{u}\notin K_\epsilon} Z_n(\boldsymbol{u}) \le (\lambda_0 + \epsilon)p\right] < \epsilon$$

for each $\epsilon > 0$. Thus $b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = O_p(1)$. ■

As noted previously, AIC corresponds to the case where $\lambda_0 = 2$; Theorem 2 confirms the well-known fact that AIC is not a consistent model selection method in the sense that if $\beta_{r+1} = \cdots = \beta_p = 0$ then asymptotically AIC gives positive probability to models with at least one of $\beta_{r+1}, \ldots, \beta_p$ nonzero. Note however that the parameter estimators computed by minimizing AIC are themselves consistent (in this case b_n-consistent). So-called consistent model selection procedures such as BIC have $\lambda_n \to \infty$ at some (usually slow) rate.

THEOREM 3. *Assume the linear model (1) where C_n in (2) satisfies (3), (4), and (9) where C is singular and D_0 is positive definite on the null space of C. Suppose that $\hat{\boldsymbol{\beta}}_n$ minimizes (12) where $\lambda_n \to +\infty$ with $\lambda_n = o(b_n^2)$. Then*

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \operatorname{argmin}\{Z(\boldsymbol{u}) : C\boldsymbol{u} = \mathbf{0}\},$$

where

$$Z(\boldsymbol{u}) = \begin{cases} \dfrac{1}{\sigma^2}(\boldsymbol{u}^T D_0 \boldsymbol{u} - 2\boldsymbol{u}^T \boldsymbol{W}) & \text{if } \beta_j \neq 0 \text{ or } \beta_j = 0,\ u_j = 0 \text{ for all } j = 1, \ldots, p \\ +\infty & \text{if } \beta_j = 0 \text{ and } u_j \neq 0 \text{ for some } j \end{cases}$$

with $\boldsymbol{u}^T \boldsymbol{W} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{u}^T D_0 \boldsymbol{u})$ for $\boldsymbol{u}$ in the null space of C.

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

Proof. When $\lambda_n \to \infty$, we can rewrite Z_n in (13) as

$$Z_n(\boldsymbol{u}) = n \ln\left[\frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \boldsymbol{x}_i^T \boldsymbol{u}/b_n)^2 \right] - n \ln\left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \right)$$

$$+ \lambda_n \sum_{j=1}^{p} [I(\beta_j + u_j/b_n \neq 0) - I(\beta_j \neq 0)].$$

Then for $\beta_j \neq 0$,

$$\lambda_n [I(\beta_j + u_j/b_n \neq 0) - I(\beta_j \neq 0)] = -\lambda_n I(\beta_j = -u_j/b_n)$$

$$\to 0$$

uniformly over compact sets (and thus this convergence is also epi-convergence). On the other hand if $\beta_j = 0$ then

$$\lambda_n [I(\beta_j + u_j/b_n \neq 0) - I(\beta_j \neq 0)] = \lambda_n I(u_j/b_n \neq 0)$$

$$= \lambda_n I(u_j \neq 0)$$

$$\to \begin{cases} 0 & \text{for } u_j = 0 \\ \infty & \text{for } u_j \neq 0, \end{cases}$$

where this pointwise convergence can be extended to epi-convergence. Now note that if λ_n grows too quickly then Z_n may be minimized at some $\hat{\boldsymbol{u}}_n$ having $\|\hat{\boldsymbol{u}}_n\| = O(b_n)$; this possibility is ruled out by the assumption that $\lambda_n = o(b_n^2)$ and so $\operatorname{argmin}(Z_n) = O_p(1)$. ■

The form of the penalty in the asymptotic objective effectively forces the limiting distribution of $b_n \hat{\beta}_{nj}$ to be a point mass at 0 when $\beta_j = 0$.

Example 2

Consider a design with C_n defined as in Example 1 with $\beta_1 = \cdots = \beta_p = 0$. When $0 \leq 0 < \rho_n = \rho < 1$, then Table 1 gives the limiting distribution of the estimated model size for $p = 5$ and $p = 10$ for $\rho = 0, 0.5, 0.9$; given that we have a "null" model here, the correct model size is 0. Table 1 suggests that as $\rho \to 1$, the probability of selecting a model of size 1 decreases and the probabilities of selecting a model of size 0 or 2 increase. The result of Theorem 2 suggests that if $a_n(1 - \rho_n) \to \psi > 0$ then the probability of AIC selecting a model of size 1 tends to 0; this is somewhat misleading as the mapping $\boldsymbol{u} \mapsto \sum_{j=1}^{p} I(u_j \neq 0)$ is not continuous at any $\boldsymbol{u}$ having at least one 0 component.

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

Table 1. Limiting distributions of estimated model size for AIC with $p = 5$ and 10 predictors, and mutual interpredictor correlations of $\rho = 0$, 0.5, and 0.9. The probability estimates are based on 10,000 replications and have a standard error of at most 0.005.

Est.	$p = 5$			$p = 10$		
size	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
0	0.43	0.46	0.52	0.18	0.22	0.23
1	0.40	0.31	0.15	0.36	0.24	0.07
2	0.14	0.19	0.31	0.28	0.35	0.54
3	0.03	0.03	0.02	0.13	0.12	0.07
>3	0.00	0.00	0.00	0.06	0.07	0.08

3. OTHER POINTS OF INTEREST

3.1. Higher Order Near Singularity

Theorems 1 and 2 require that the matrix D_0 be positive definite on the null space of C. Unfortunately, this is not always true; Phillips (2001) gives an example involving polynomial regression with "slowly varying" predictors where this condition is violated.

The near singularity condition $a_n(C_n - C) \to D_0$ with $\boldsymbol{u}^T D_0 \boldsymbol{u} > 0$ for nonzero $\boldsymbol{u}$ satisfying $C\boldsymbol{u} = \boldsymbol{0}$ can be generalized as follows. We start by recursively defining matrices H_1, D_1, H_2, $D_2, \ldots$ such that

$$H_1 = a_n(C_n - C) - D_0, \tag{14}$$

$$a_n^{(1)} H_1 \to D_1, \tag{15}$$

$$H_k = a_n^{(k-1)} H_{k-1} - D_{k-1} \qquad k = 2,3,\ldots, \tag{16}$$

$$a_n^{(k)} H_k \to D_k \qquad k = 2,3,\ldots. \tag{17}$$

Now define the following subspace of the null space of C:

$$\mathcal{S}_k = \{\boldsymbol{v} : C\boldsymbol{v} = D_0\boldsymbol{v} = \cdots = D_{k-1}\boldsymbol{v} = \boldsymbol{0}, \boldsymbol{v}^T D_k \boldsymbol{v} > 0 \quad \text{for } \boldsymbol{v} \neq 0\}. \tag{18}$$

Note that $\mathcal{S}_k$ is always well defined (if C, $D_0, \ldots, D_k$ in (14)–(17) are well defined) as it contains at least the vector $\boldsymbol{0}$. However, we are most interested in cases where $\mathcal{S}_k$ is larger. We can then redefine b_n in terms of $a_n^{(1)}, \ldots, a_n^{(k)}$ as follows:

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

$$b_n = \left(\frac{n}{a_n^{(1)} \times \cdots \times a_n^{(k)}} \right)^{1/2}. \tag{19}$$

Then it can be shown that the conclusions of Theorems 1 and 2 hold with b_n defined as in (19), $\mathrm{Var}(\boldsymbol{u}^T \boldsymbol{W}) = \sigma^2 \boldsymbol{u}^T D_k \boldsymbol{u} > 0$ for $\boldsymbol{u} \in \mathcal{S}_k$ defined in (18), D_k replacing D_0 in the definition of $Z(\boldsymbol{u})$:

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{argmin}\{Z(\boldsymbol{u}) : \boldsymbol{u} \in \mathcal{S}_k\}.$$

It is also possible to extend the results of this paper to cases where different degrees of near singularity exist in disjoint subsets of variables; in this case, we will obtain different convergent rates for the estimators of the parameters in the different subsets. For example, if $\boldsymbol{x}_i = \mathrm{vec}(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)})$ then

$$C_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$$

$$= \begin{pmatrix} C_n^{(11)} & C_n^{(12)} \\ C_n^{(21)} & C_n^{(22)} \end{pmatrix}.$$

Suppose that $C_n \to C$ where $C_n^{(11)} \to C^{(11)}$ (nonsingular) and $C_n^{(22)} \to C^{(22)}$ (singular) with $a_n(C_n^{(22)} - C^{(22)}) \to D_0^{(22)}$; then it is also reasonable to assume that $a_n^{1/2}(C_n^{(12)} - C^{(12)}) \to D_0^{(12)}$ (and likewise for $C_n^{(21)}$). In this case, writing $\boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$, we would typically (i.e., subject to other regularity conditions) have

$$\begin{pmatrix} \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}^{(1)}) \\ b_n(\hat{\boldsymbol{\beta}}_n^{(2)} - \boldsymbol{\beta}^{(2)}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \boldsymbol{V}^{(1)} \\ \boldsymbol{V}^{(2)} \end{pmatrix},$$

where $b_n = (n/a_n)^{1/2}$. For shrinkage estimation minimizing (7), we would need to choose λ_n to match the slowest rate of convergence to obtain nondegenerate limiting distributions.

3.2. Maximum Likelihood and GMM Estimation

The results of this paper extend naturally to maximum likelihood estimation where the information matrix is nearly singular. In regular models where the log-likelihood function is locally quadratic, it is straightforward to extend Theorems 1 and 2; applications would include model selection and shrinkage estimation for so-called generalized linear models, which include logistic regression and log-linear Poisson regression. As mentioned previously, the notion of near singularity may be very useful in determining the asymptotic behavior of estimation procedures with weak instruments. In the context of GMM and gener-

https://doi.org/10.1017/S0266466608080146 Published online by Cambridge University Press

alized empirical likelihood estimation, Caner (2004, 2006) has investigated similar issues.

It is worth noting that there is a considerable literature on estimation for models where the information matrix is singular; for some recent examples, see Barnabani (2002) and Rotnitzky, Cox, Bottai, and Robins (2000). In such cases, typically the limiting distributions of maximum likelihood estimators are concentrated on a lower dimensional subspace or have a slower rate of convergence than the standard rate.

REFERENCES

Barnabani, M. (2002) Wald-Based Approach with Singular Information Matrix. Working paper 2002/13, Department of Statistics, University of Florence.

Caner, M. (2004) Nearly Singular Design in GMM and Generalized Empirical Likelihood Estimators. Working paper, Department of Economics, University of Pittsburgh.

Caner, M. (2006) Lasso Type GMM Estimator. Working paper, Department of Economics, University of Pittsburgh.

Chernozhukov, V. (2005) Extremal quantile regression. *Annals of Statistics* 33, 806–839.

Chernozhukov, V. & H. Hong (2004) Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica* 72, 1445–1480.

Fan, J. & R. Li (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.

Frank, I.E. & J.H. Friedman (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 35, 109–148.

Fu, W.J. (1998) Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* 7, 397–416.

Gabaix, X. & R. Ibragimov (2006) Log(rank − 1/2): A Simple Way to Improve the OLS Estimation of Tail Exponents. Working paper, Harvard Institute of Economic Research.

Geyer, C.J. (1994) On the asymptotics of constrained M-estimation. *Annals of Statistics* 22, 1993–2010.

Geyer, C.J. (1996) On the Asymptotics of Convex Stochastic Optimization. Technical report, Department of Statistics, University of Minnesota.

Hoerl, A.E. & R.W. Kennard (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.

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

Knight, K. & W. Fu (2000) Asymptotics for lasso-type estimators. *Annals of Statistics* 28, 1356–1378.

Leeb, H. & B. Pötscher (2006) 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.

Lorber, A., L.E. Wanger, & B.R. Kowalski (1987) A theoretical foundation for the PLS algorithm. *Journal of Chemometrics* 1, 19–31.

Pflug, G.C. (1995) Asymptotic stochastic programs. *Mathematics of Operations Research* 20, 769–789.

Phillips, P.C.B. (2001) Regression with Slowly Varying Regressors. Cowles Foundation Discussion paper 1310, Yale University.

Radchenko, P. (2004) Reweighting the Lasso. Unpublished manuscript, Department of Statistics, University of Chicago.

Rotnitzky, A., D.R. Cox, M. Bottai, & J. Robins (2000) Likelihood-based inference with singular information matrix. *Bernoulli* 6, 243–284.

Srivastava, M.S. (1971) On fixed width confidence bounds for regression parameters. *Annals of Mathematical Statistics* 42, 1403–1411.

Stock, J.H., J.H. Wright, & M. Yogo (2002) A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 20, 518–529.

Stone, M. & R.J. Brooks (1990) Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *Journal of the Royal Statistical Society, Series B* 52, 237–269; corrigendum (1992) 54, 906–907.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58, 267–288.

van der Vaart, A.W. & J.A. Wellner (1996) *Weak Convergence and Empirical Processes with Applications to Statistics.* Springer.

Wold, H. (1984) PLS regression. In N.L. Johnson & S. Kotz (eds.), *Encyclopedia of Statistical Sciences*, vol. 6, pp. 581–591. Wiley.