Efficient Shrinkage in Parametric Models

Bruce E. Hansen*
University of Wisconsin†

www.ssc.wisc.edu/~bhansen

This draft: November 2011
Preliminary

Abstract

This paper introduces shrinkage for general parametric models. We show how to shrink maximum likelihood estimators towards parameter subspaces defined by general nonlinear restrictions. We derive the asymptotic distribution and risk of the generalized shrinkage estimator using a local asymptotic framework. We show that if the shrinkage dimension exceeds two, the asymptotic risk of the shrinkage estimator is strictly less than that of the MLE. This reduction holds globally in the parameter space. We show that the reduction in asymptotic risk is substantial, even for moderately large values of the parameters.

The formula simplify in a very convenient way in the context of high dimensional models. We derive a simple bound for the asymptotic risk.

We also provide a new large sample minimax efficiency bound. We use the concept of local asymptotic minimax bounds, a generalization of the conventional asymptotic minimax bounds. The difference is that we consider minimax regions that are defined locally to the parametric restriction, and are thus tighter. We show that our shrinkage estimator asymptotically achieves this local asymptotic minimax bound when the shrinkage dimension is high. This theory is a combination and extension of standard asymptotic efficiency theory (e.g. chapter 8 of Van der Vaart (1998)) and local minimax efficiency theory for Gaussian models (e.g. chapter 7 of Wasserman (2006)).

Our estimators and theory allow for the loss functions which are in the class of weighted-average mean-squared-error, where the user selects the weights. This allows for the division of the parameter space into focus and nuisance parameters.

*Research supported by the National Science Foundation.
†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706.

1 Introduction

In a conventional parametric setting, one where maximum likelihood estimation applies, it is widely understood that the conventional MLE is asymptotically efficient – no other estimator can achieve a smaller mean-squared error. In this paper we show that this understanding is incomplete. We show that a very simple shrinkage modification can achieve substantially smaller asymptotic risk (weighted mean-squared error) and thus the conventional MLE is inefficient. The magnitude of the improvement depends on the distance between the true parameters and a parametric restriction. If the distance is small then the reduction in risk can be quite substantial. Even when the distance is moderately large the reduction in risk can be significant.

Shrinkage was introduced by James and Stein (1961) in the context of exact normal sampling, and spawned an enormous literature. Our goal is to extend their methods to encompass a broad array of conventional parametric econometric models. In subsuquent work we expect to extend these results to semiparametric estimation settings.

To make these extensions we need to develop an asymptotic (large sample) distributional theory for shrinkage estimators. This can be accomplished using the local asymptotic normality approach (e.g., van der Vaart (1998)). We model the parameter vector as being in a $n^{-1/2}$-neighborhood of the specified restriction, so that the asymptotic distributions are continuous in the localizing parameter. This approach has been used successfully for averaging estimators by Hjort and Claeskens (2003) and Liu (2011), and for Stein-type estimators by Saleh (2006).

Given the localized asymptotic parameter structure, the asymptotic distribution of the shrinkage estimator takes a James-Stein form. It follows that the asymptotic risk of the estimator can be analyzed using techniques introduced by Stein (1981). Not surprisingly, the benefits of shrinkage are maximized when the magnitude of the localizing parameter is small. What is surprising (or at least it may be to some readers) is that the numerical magnitude of the reduction in asymptotic risk (weighted mean-squared error) is quite substantial, even for relatively distant values of the localizing parameter. We can be very precise about the nature of this improvement, as we provide simple and interpretable expressions for the asymptotic risk.

We measure estimation efficiency by asymptotic risk – the large sample weighted mean-squared error. The weighted MSE necessarily depends on a weight matrix, and the optimal shrinkage estimator depends on its value. For a generic measure of fit the weight matrix can be set to the inverse of the usual asymptotic covariance, but in other cases a user may wish to select a specific weight matrix which allows focus on selected parameters.

We benefit from the recent theory of efficient high-dimensional Gaussian shrinkage, specifically Pinkser's Theorem (Nussbaum, 1999), which gives a lower minimax bound for estimation of high dimensional normal means. We combine Pinker's Theorem with classic large-sample minimax efficiency theory to provide a new asymptotic local minimax efficiency bound. We provide a minimax lower bound on the asymptotic risk, and show that the asymptotic risk of our shrinkage estimators equals this lower bound when the shrinkage dimension diverges towards infinity. This shows that the proposed shrinkage estimator is minimax efficient.

The literature on shrinkage estimation in enormous, and we only mention a few of the most relevant contributions. Stein (1956) first observed that an unconstrained Gaussian estimator is inadmissible when the dimension exceeds two. James and Stein (1961) introduced the classic shrinkage estimator. Baranchick (1964) showed that the positive part version has reduced risk. Judge and Bock (1978) developed the method for econometric estimators. Stein (1981) provided theory for the analysis of risk. Oman (1982a, 1982b) developed estimators which shrink Gaussian estimators towards linear subspaces. An in-depth treatment of shrinkage theory can be found in Chapter 5 of Lehmann and Casella (1998). Wasserman (2006, chapter 7) reviews the recent minimax efficiency theory for high-dimensional Gaussian inference.

The organization of the paper is as follows. Section 2 presents the general framework and the generalized shrinkage estimator. Section 3 presents the asymptotic distribution of the estimator. Section 4 develops a bound for its asymptotic risk. Section 5 uses a large-parameter approximation to the asymptotic risk, showing that the gains are substantial and broad in the parameters space. Section 6 presents a new local minimax efficiency bound. Mathematical proofs are left ot the appendix..

2 Model

Suppose that we observe a random sample $X_1, ..., X_n$ from a density $f(x, \boldsymbol{\theta})$ indexed by a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$. Furthermore, suppose we have a restricted parameter space $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ defined by a differentiable parametric restriction

$$\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}\} \tag{1}$$

where $\mathbf{r}(\boldsymbol{\theta}) : \mathbb{R}^m \to \mathbb{R}^p$ with $p \geq 3$. Set $\mathbf{R}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{r}(\boldsymbol{\theta})'$.

The restriction (1) is not believed to be true, but represents a plausible simplification, centering, or "prior" about the likely value of $\boldsymbol{\theta}$. An important special case occurs when $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_0\}$ is a singleton (such as the zero vector) in which case $p = m$. We call this situation **full shrinkage**. We call the case $p < m$ **partial shrinkage**. Most commonly, we can think of the unrestricted model $\boldsymbol{\Theta}$ as the "kitchen-sink", and the restricted model $\boldsymbol{\Theta}_0$ as a tight parametric specification. Often $\boldsymbol{\Theta}_0$ will take the form of an exclusion restiction. For example, if we partition

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} \qquad \begin{matrix} m - p \\ p \end{matrix}$$

then an exclusion restriction takes the form $\mathbf{r}(\boldsymbol{\theta}) = \boldsymbol{\theta}_2$. $\boldsymbol{\Theta}_0$ may also be a linear subspace in which case we can write

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{R}'\boldsymbol{\theta} - \mathbf{a} \tag{2}$$

where $\mathbf{R}$ is $(m + p) \times p$ and $\mathbf{a}$ is $p \times 1$. In other cases, $\boldsymbol{\Theta}_0$ may be a nonlinear subspace, for example if $\mathbf{r}(\boldsymbol{\theta}) = \theta_1\theta_2 - 1$.

The log likelihood for the sample is

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}). \tag{3}$$

We consider two standard estimators of $\boldsymbol{\theta}$. The unrestricted maximum likelihood estimator (MLE) maximizes (3) over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}_n(\boldsymbol{\theta}).$$

The restricted MLE maximizes (3) over $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathcal{L}_n(\boldsymbol{\theta}).$$

The information matrix is $\boldsymbol{I}_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}}\left(s(X_i, \boldsymbol{\theta}) s(X_i, \boldsymbol{\theta})'\right)$ where $s(x, \boldsymbol{\theta}) = \dfrac{\partial}{\partial \boldsymbol{\theta}} \ln f(x, \boldsymbol{\theta})$. Set $\mathbf{V}_{\boldsymbol{\theta}} = \boldsymbol{I}_{\boldsymbol{\theta}}^{-1}$ and its estimate

$$\widehat{\mathbf{V}} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \ln f(X_i, \widehat{\boldsymbol{\theta}}) \right)^{-1}.$$

Our goal is to improve upon the MLE $\widehat{\boldsymbol{\theta}}$ by shrinking it towards the restricted estimator $\widetilde{\boldsymbol{\theta}}$. We will measure estimation efficiency by weighted quadratic loss. For some positive semi-definite weight matrix $\mathbf{W} \geq 0$ our loss function is

$$\ell(\mathbf{u}) = \frac{\mathbf{u}'\mathbf{W}\mathbf{u}}{\operatorname{tr}(\mathbf{V}\mathbf{W})}.$$

We have scaled the loss function by $\operatorname{tr}(\mathbf{V}\mathbf{W})$ to normalize the asymptotic risk of the MLE. The ideal weight matrix $\mathbf{W}$ may not be known, but we assume that a consistent estimate $\widehat{\mathbf{W}}$ is available. The weight function need not be positive definite, and indeed this is appropriate when a subset of $\boldsymbol{\theta}$ are nuisance parameters. (We will discuss this situation later.)

In many cases we want a generic measure of fit and so do not have a motivation for selection of the weight matrix $\mathbf{W}$. In this case, we recommend $\mathbf{W} = \mathbf{V}^{-1}$ as this renders the loss invariant to rotations of the parameter space. We call this the **canonical case**. Notice as well that in this case the loss function simplifies to $\ell(\mathbf{u}) = \dfrac{\mathbf{u}'\mathbf{V}^{-1}\mathbf{u}}{p}$ and we have the natural estimator $\widehat{\mathbf{W}} = \widehat{\mathbf{V}}^{-1}$ for $\mathbf{W}$. The canonical case is convenient for practical applications as many formula simplify.

In other cases the economic or statistical problem will suggest a particular choice for the weight matrix $\mathbf{W}$. This includes the situation where a subset of the parameter vector $\boldsymbol{\theta}$ is of particular interest. We call this situation **targeted shrinkage**.

Our generalized shrinkage estimator of $\boldsymbol{\theta}$ is the weighted average of the MLE and restricted MLE

$$\widehat{\boldsymbol{\theta}}^{*} = \hat{w}\widehat{\boldsymbol{\theta}} + (1 - \hat{w})\widetilde{\boldsymbol{\theta}} \tag{4}$$

where

$$\hat{w} = \left(1 - \frac{\tau_n}{D_n}\right)_+ \tag{5}$$

with $(x)_+ = x1(x \geq 0)$ is the "positive-part" function, and

$$D_n = n\left(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\right)' \widehat{\mathbf{W}} \left(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\right), \tag{6}$$

a distance-type statistic for the restriction (1) in $\boldsymbol{\Theta}$. The scalar $\tau_n \geq 0$ controls the degree of shrinkage. We recommend

$$\tau_n = \left(\operatorname{tr}(\widehat{\mathbf{A}}) - 2\lambda_{\max}(\widehat{\mathbf{A}})\right)_+ \tag{7}$$

where $\lambda_{\max}(\widehat{\mathbf{A}})$ is the largest eigenvalue of the matrix $\widehat{\mathbf{A}}$,

$$\widehat{\mathbf{A}} = \left(\widehat{\mathbf{R}}'\widehat{\mathbf{V}}\widehat{\mathbf{R}}\right)^{-1} \widehat{\mathbf{R}}'\widehat{\mathbf{V}}\widehat{\mathbf{W}}\widehat{\mathbf{V}}\widehat{\mathbf{R}} \tag{8}$$

and $\widehat{\mathbf{R}} = \mathbf{R}(\widehat{\boldsymbol{\theta}})$. The recommendation (7) will be justified in Section 4.

Notice that the degree of shrinkage depends on the ratio τ_n/D_n. When $D_n < \tau_n$ then $\hat{w} = 0$ and $\widehat{\boldsymbol{\theta}}^* = \widehat{\boldsymbol{\theta}}$ equals the restricted estimator. When $D_n > \tau_n$ then $\widehat{\boldsymbol{\theta}}^*$ is a weighted average of the usual and restricted estimators, with more weight on the usual estimator as D_n/τ_n is large.

Several simplifications occur in the canonical case ($\mathbf{W} = \mathbf{V}^{-1}$). The full shrinkage estimator is the classic James-Stein estimator, and the partial shrinkage estimator with linear $\mathbf{r}(\boldsymbol{\theta})$ is Oman's (1982ab) shrinkage estimator. If $\tau_n = p$ the partial shrinkage estimator also corresponds to Hansen's (2007) Mallows Model Averaging (MMA) estimator. It is also useful to observe that in the canonical case if τ_n is set as recommended in (7), then $\widehat{\mathbf{A}} = \mathbf{I}_p$ and $\tau_n = p - 2$. Additionally, D_n in (6) is asymptotically equivalent to

$$LR_n = 2\left(\mathcal{L}_n(\widehat{\boldsymbol{\theta}}) - \mathcal{L}_n(\widetilde{\boldsymbol{\theta}})\right),$$

the likelihood ratio statistic for (1). Thus in the canonical case we can substitute LR_n for D_n in formula (5), without a change in the theory.

3 Asymptotic Distribution

To obtain a useful approximation we derive the asymptotic distribution along parameter sequences $\boldsymbol{\theta}_n$ approaching the restricted set $\boldsymbol{\Theta}_0$. In particular we consider sequences of the form

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h} \tag{9}$$

where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$ and $\boldsymbol{h} \in \mathbb{R}^m$. In this framework, the true value of the parameter is $\boldsymbol{\theta}_n$, and $n^{-1/2}\boldsymbol{h}$ is the magnitude of the distance of the parameter from the restricted set. For any fixed $\boldsymbol{h}$ this distance shrinks as the sample size increases, but as we do not restrict the magnitude of $\boldsymbol{h}$ this does not meaningfully limit the application of our theory. We will use the symbol "$\xrightarrow[\boldsymbol{\theta}_n]{}$" to denote

convergence in distribution along the parameter sequences $\boldsymbol{\theta}_n$ as defined in (9).

Assumption 1 *[to be completed] The probability model $f(x, \boldsymbol{\theta})$ satisfies the standard conditions for the consistency and asymptotic normality of the MLE $\widehat{\boldsymbol{\theta}}$.*

Assumption 2 *$\mathbf{R}(\boldsymbol{\theta})$ is continuous in a neighborhood of $\boldsymbol{\theta}_0$, and $\operatorname{rank}(\mathbf{R}) = p$ where $\mathbf{R} = \mathbf{R}(\boldsymbol{\theta}_0)$. $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$ and $\tau_n \xrightarrow{p} \tau$ as $n \to \infty$.*

Let $\mathbf{V} = \mathbf{V}_{\boldsymbol{\theta}_0} = \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}$ be the asymptotic variance of the MLE under the sequences (9).

Theorem 1 *Under Assumptions 1 and 2, along the sequences (9)*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V}), \tag{10}$$

$$\sqrt{n}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathrm{Z} - \mathbf{V}\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'(\mathrm{Z} + \boldsymbol{h}), \tag{11}$$

$$D_n \xrightarrow[\boldsymbol{\theta}_n]{} \xi = (\mathrm{Z} + \boldsymbol{h})'\mathbf{B}(\mathrm{Z} + \boldsymbol{h}), \tag{12}$$

$$\hat{w} \xrightarrow[\boldsymbol{\theta}_n]{} w = \left(1 - \frac{\tau}{\xi}\right)_+, \tag{13}$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} w\mathrm{Z} + (1-w)\left(\mathrm{Z} - \mathbf{V}\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'(\mathrm{Z} + \boldsymbol{h})\right), \tag{14}$$

where

$$\mathbf{B} = \mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\left(\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}\mathbf{R}\right)\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'. \tag{15}$$

Theorem 1 gives expressions for the joint asymptotic distribution of the MLE, restricted MLE, and shrinkage estimators as a transformation of the normal random vector Z and the non-centrality parameter $\boldsymbol{h}$. The asymptotic distribution of $\widehat{\boldsymbol{\theta}}^*$ is written as a random weighted average of the asymptotic distributions of $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$. Since the distribution of $\widehat{\boldsymbol{\theta}}^*$ depends on $\boldsymbol{h}$, the estimator $\widehat{\boldsymbol{\theta}}^*$ is non-regular.

The asymptotic distribution is obtained for parameter sequences $\boldsymbol{\theta}_n$ tending towards the restricted parameter space $\boldsymbol{\Theta}_0$. The conventional case of fixed $\boldsymbol{\theta}$ can be obtained by letting $\boldsymbol{h}$ diverge towards infinity, in which case $\xi \to_p \infty$, $w \to_p 1$, and the distribution on the right-hand-side of (14) tends towards $\mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$.

It is important to understand that Theorem 1 does not require that the true value of $\boldsymbol{\theta}_n$ satisfy the restriction to $\boldsymbol{\Theta}_0$, only that it is in a $n^{-1/2}$-neighborhood of $\boldsymbol{\Theta}_0$. The distinction is important, as the size of this neighborhood is determined by $\boldsymbol{h}$ which we allow to be arbitrarily large.

Equation (12) also provides the asymptotic distribution ξ of the distance-type statistic D_n. The limit distribution ξ controls the weight w and thus the degree of shrinkage, so it is worth investigating further. Notice that its expected value is

$$\mathbb{E}\xi = \boldsymbol{h}'\mathbf{B}\boldsymbol{h} + \mathbb{E}\operatorname{tr}\left(\mathbf{B}\mathrm{Z}\mathrm{Z}'\right) = \boldsymbol{h}'\mathbf{B}\boldsymbol{h} + \operatorname{tr}(\mathbf{A}) \tag{16}$$

where $\mathbf{B}$ is from (15) and

$$\mathbf{A} = \left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}\mathbf{V}\mathbf{W}\mathbf{V}\mathbf{R}. \tag{17}$$

The matrices $\mathbf{A}$ and $\mathbf{B}$ play an important roles in our theory. Notice that in the full shrinkage case we have the simplifications $\mathbf{B} = \mathbf{W}$ and $\mathbf{A} = \mathbf{V}\mathbf{W}$. In the canonical case $\mathbf{W} = \mathbf{V}^{-1}$ we have the simplifications $\mathbf{A} = \mathbf{I}_p$ and $\operatorname{tr}(\mathbf{A}) = p$ and thus (16) is

$$\mathbb{E}\xi = \boldsymbol{h}'\mathbf{B}\boldsymbol{h} + p \tag{18}$$

In fact, in the canonical case, $\xi \sim \chi^2_p(\boldsymbol{h}'\mathbf{B}\boldsymbol{h})$, a non-central chi-square random variable with non-centrality parameter $\boldsymbol{h}'\mathbf{B}\boldsymbol{h}$ and degrees of freedom p.

In general, the scalar $\boldsymbol{h}'\mathbf{B}\boldsymbol{h}$ captures how the divergence of $\boldsymbol{\theta}_n$ from the restricted region $\boldsymbol{\Theta}_0$ affects the distribution of ξ.

4 Asymptotic Risk

The risk of an estimator T_n is its expected loss $\mathbb{E}_\theta \ell(T_n - \boldsymbol{\theta})$. In general this is difficult to evaluate, and may not even be finite unless T_n has a finite second moments. To obtain a useful approximation we calculate the asymptotic risk (as $n \to \infty$) and to ensure its existence we calculate a limiting trimmed risk. That is, we define the trimmed loss function $\ell_\zeta(\mathbf{u}) = \min\{\ell(\mathbf{u}), \zeta\}$ and define the asymptotic risk of an estimator T_n for $\boldsymbol{\theta}$ under the sequence (9) as

$$\lim_{\zeta\to\infty} \liminf_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell_{\zeta/n}(T_n - \boldsymbol{\theta}_n). \tag{19}$$

Recalling the matrix $\mathbf{A} = (\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}\mathbf{V}\mathbf{W}\mathbf{V}\mathbf{R}$ from (17) we define the scalar

$$\lambda_p = \frac{\operatorname{tr}(\mathbf{A})}{\lambda_{\max}(\mathbf{A})} \tag{20}$$

which satisfies $\lambda_p \le p$. In the canonical case $\mathbf{W} = \mathbf{V}^{-1}$ we find $\lambda_p = p$. In general, λ_p can be thought of as the effective shrinkage dimension.

Theorem 2 *Under Assumptions 1 and 2, $\lambda_p > 2$, and $0 < \tau < 2\left(\operatorname{tr}(\mathbf{A}) - 2\lambda_{\max}(\mathbf{A})\right)$, then*

$$\lim_{\zeta\to\infty} \liminf_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell_{\zeta/n}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_n) < \lim_{\zeta\to\infty} \liminf_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell_{\zeta/n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n) \tag{21}$$

for all $\boldsymbol{h}$. Furthermore, if we define the ball

$$\mathbf{H}(c) = \left\{\boldsymbol{h} : \boldsymbol{h}'\mathbf{B}\boldsymbol{h} \le \operatorname{tr}(\mathbf{A})\, c\right\} \tag{22}$$

then

$$\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \lim_{\zeta\to\infty} \liminf_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell_{\zeta/n}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_n) \le 1 - \frac{\tau}{\operatorname{tr}(\mathbf{W}\mathbf{V})} \frac{2\left(\operatorname{tr}(\mathbf{A}) - 2\lambda_{\max}(\mathbf{A})\right) - \tau}{\operatorname{tr}(\mathbf{A})\,(c+1)}. \tag{23}$$

Equation (21) shows that the asymptotic risk of the shrinkage estimator is strictly less than that of the MLE for all parameter values. As this holds for even extremely large values of $\boldsymbol{h}$, this shows that in a very real sense, the shrinkage estimator dominates the usual estimator.

The assumption $\lambda_p > 2$ is the critical condition needed to ensure that the shrinkage estimator has uniformly smaller asymptotic risk than the usual estimator. In the canonical case $\mathbf{W} = \mathbf{V}^{-1}$, Assumption 3 is equivalent to $p > 2$, which is Stein's (1956) classic conditions for shrinkage. As shown by Stein (1956) $p > 2$ is necessary in order for shrinkage to achieve global reductions in risk relative to unrestricted estimation. Assumption 3 generalizes $p > 2$ to allow for general weight matrices.

The condition $0 < \tau < 2\left(\mathrm{tr}(\mathbf{A}) - 2\lambda_{\max}(\mathbf{A})\right)$ simplifies to $0 < \tau < 2(p-2)$ in the canonical case, which is a standard restriction on the shrinkage parameter.

Equation (23) provides a uniform bound for the asymptotic risk in the ball $\boldsymbol{h} \in \mathbf{H}(c)$. The bound (23) is minimized by setting $\tau = \mathrm{tr}\left(\mathbf{A}\right) - 2\lambda_{\max}\left(\mathbf{A}\right)$ in which case (23) becomes

$$\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \lim_{\zeta\to\infty} \liminf_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell_{\zeta/n}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_n) \leq 1 - \frac{\left(\mathrm{tr}\left(\mathbf{A}\right) - 2\lambda_{\max}\left(\mathbf{A}\right)\right)^2}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)\mathrm{tr}\left(\mathbf{A}\right)(c+1)}. \tag{24}$$

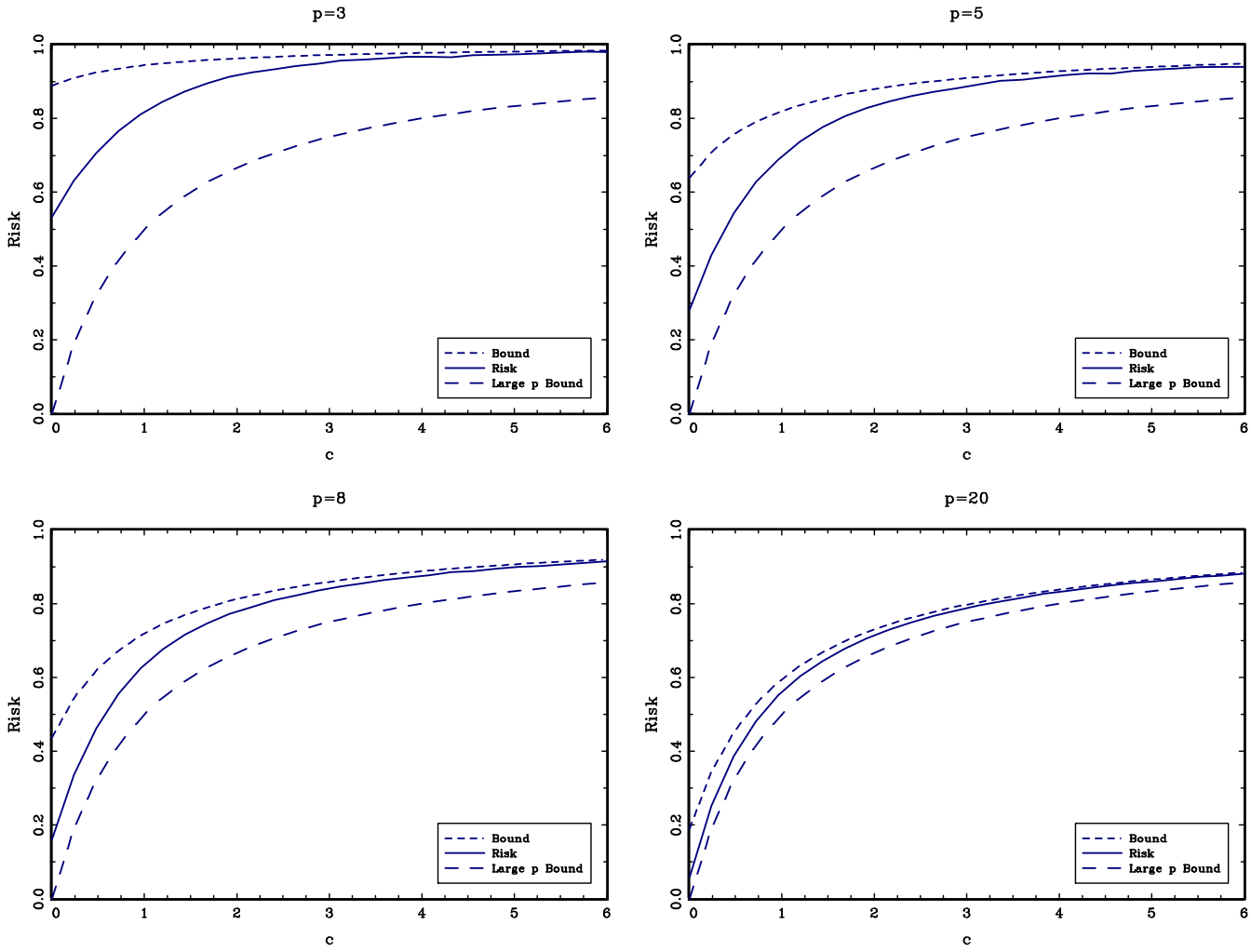
The fact that this is the tightest uniform bound motivates our recommendation (7) for the shrinkage parameter τ_n. With this choice the shrinkage estimator satisfies the tight uniform bound (24).

To illustrate these results numerically, we plot in Figure 1 the asymptotic risk of the shrinkage estimator $\widehat{\boldsymbol{\theta}}^*$ in the full shrinkage canonical case. The asymptotic risk is only a function of p and c, so we plot the risk as a function of c for $p = 3$, 5, 8 and 20. The asymptotic risk is plotted with the solid line. We also plot the upper bound (24) using the short dashes. Recall that the loss function has been normalized so that the asymptotic risk of the unrestricted MLE is 1, so values less than 1 indicate risk reduction relative to the unrestricted MLE.

From Figure 1 we can see that the asymptotic risk of the shrinkage estimator is monotonically decreasing as $c \to 0$, indicating (as expected) that the greatest risk reductions occur for parameter values near the restricted parameter space. We also can see that the asymptotic risk function decreases as p increases. Furthermore, we can observe that the upper bound (24) is not particularly tight for small p, but improves as p increases. This means that risk improvements implied by Theorem 3 are underestimates of the actual improvements in asymptotic risk due to shrinkage.

5 High Dimensional Models

In the previous section we showed numerically that accuracy of the risk bound (24) improves as the shrinkage dimension p increases. Indeed the bound (24) leads to a simple approximation for the asymptotic risk when the shrinkage dimension p is large.



Theorem 3 *Under Assumptions 1 and 2, if τ_n is set as in (7), and as $p \to \infty$, $\lambda_p \to \infty$ and*

$$\frac{\mathrm{tr}\,(\mathbf{A})}{\mathrm{tr}\,(\mathbf{WV})} \longrightarrow a, \tag{25}$$

then

$$\liminf_{p\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \lim_{\zeta\to\infty} \liminf_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell_\zeta(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_n) \leq 1 - \frac{a}{c+1}. \tag{26}$$

Essentially, equation (26) is a simplified version of (24). This is an asymptotic (large n) generalization of the results obtained by Casella and Hwang (1982). (See also Theorem 7.42 of Wasserman (2006).) These authors only considered the canonical, non-asymptotic, full shrinkage case. Theorem 3 generalizes these results to asymptotically distributions, arbitrary weight matrices, and partial shrinkage.

The asymptotic risk of the MLE is 1. The ideal risk of the restricted estimator (when $c = 0$) is $1 - a$. The risk in (26) varies between $1 - a$ and 1, depending on c. In fact, we can see that $1/(1 + c)$ is the percentage decrease in risk relative to the usual estimator obtained by shrinkage,

when shrunk towards the restricted estimator.

(26) quantifies the reduction in risk obtained by the shrinkage estimator as the ratio $a/(1+c)$. The gain from shrinkage is greatest when the ratio $a/(1+c)$ is large, meaning that there are many mild restrictions.

a is a measure of the effective number of restrictions relative to the total number of parameters. Note that $0 \leq a \leq 1$, with $a = 1$ in the full shrinkage case and $a = 0$ when there is no shrinkage. In the canonical case, $a = \lim_p \dfrac{p}{m}$, the ratio of the number of restrictions to the total number of parameters. In the full shrinkage case (26) simplifies to

$$\liminf_{p\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \lim_{\zeta\to\infty} \liminf_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell_\zeta(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_n) \leq \frac{c}{c+1}. \tag{27}$$

c is a measure of the strength of the restrictions. To gain insight, consider the canonical case $\mathbf{W} = \mathbf{V}^{-1}$, and write the distance statistic (6) as $D_n = pF_n$, where F_n is an F-type statistic for (1). Using (18), this has the approximate expectation

$$\mathbb{E}F_n \longrightarrow \frac{\mathbb{E}\xi}{p} = 1 + \frac{\boldsymbol{h}'\mathbf{B}\boldsymbol{h}}{p} \leq 1 + c$$

where the inequality is for $\boldsymbol{h} \in \mathbf{H}(c)$. This means that we can interpret c in terms of the expectation of the F-statistic for (1). We can view the empirically-observed $F_n = D_n/p$ as an estimate of $1 + c$ and thereby assess the expected reduction in risk relative to the usual estimator. For example, if $F_n \approx 2$ (a moderate value) then $c \approx 1$, suggesting that the percentage reduction in asymptotic risk due to shrinkage is 50%, a very large decrease. Even if the F statistic is very large, say $F_n \approx 10$, then $c \approx 9$, suggesting the percentage reduction in asymptotic risk due to shrinkage is 10%, which is quite substantial. Equation (26) indicates that substantial efficiency gains can be achieved by shrinkage for a large region of the parameter space.

We assess the high-dimensional bound numerically by including the bound (27) in the plots of Figure 1 (the long dashes). We can see that the large-p bound (27) lies beneath the finite-p bound (24) (the short dashes) and the actual asymptotic risk (the solid lines). The differences are quite substantial for small p, but diminish as p increases. For $p = 20$ the three lines are quite close, indicating that the large-p approximation (27) is reasonally accurate for $p = 20$. Thus the technical approximation $p \to \infty$ seems to be a useful approximation even for moderate shrinkage dimensions.

Nevertheless, we have found that gains are most substantial in high dimensional models which are reasonably close to a low dimensional model. This is quite appropriate for econometric applications. It is common to see applications where the unconstrained model is quite high dimensional yet the unconstrained model is not substantially different from a low dimensional specification. This is precisely the context where shrinkage will be most beneficial. The shrinkage estimator will efficiently combine both model estimates, shrinking the high dimensional model towards the low dimensional model.

The asymptotic approximation in Theorem 3 might seem a bit odd, inthat we first take the sample size n to infinity and then take the dimension p to infinity. As p gets large it may seem more appropriate to view the underlying model as nonparametric, in which case the assumption that the usual estimator is $n^{-1/2}$ normal should be replaced by a nonparametric rate of convergence to a noncentral normal distribution. Instead, we view (26) as reflecting an approximation appropriate for finite high-dimensional models. We leave to future work the important extension to the case where $\widehat{\boldsymbol{\theta}}$ is nonparametric.

6 Minimax Risk

We have shown that the generalized shrinkage estimator has substantially lower asymptotic risk than the MLE. Does our shrinkage estimator have the lowest possible risk, or can an alternative shrinkage estimator attain even lower asymptotic risk? In this section we explore this question by proposing a local minimax efficiency bound.

The efficiency theory of Hájek (1970, 1972) defines the asymptotic maximum risk of a sequence of estimators T_n for $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h}$ with arbitrary $\boldsymbol{h}$ as

$$\sup_{I \subset \mathbb{R}^m} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(T_n - \boldsymbol{\theta}_n\right) \tag{28}$$

where the first supremum is taken over all finite subsets I of $\mathbb{R}^m$. The minimax theorem (e.g. Theorem 8.11 of van der Vaart (1998)) demonstrates that under quite mild regularity conditions (28) is bounded below by 1. This demonstrates that no estimator has smaller minimax risk than the MLE over unbounded $\boldsymbol{h}$.

A limitation with this theorem is that taking the maximum risk over all intervals is a very stringent requirement. It does not allow for local improvements such as those demonstrated in Theorems 2 and 3. To remove this limitation we can define the local asymptotic maximum risk of a sequence of estimators T_n as

$$\sup_{I \subset \mathbf{H}(c)} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(T_n - \boldsymbol{\theta}_n\right) \tag{29}$$

which replaces the supremum over all subsets of $\mathbb{R}^m$ with the supremum over all finite subsets of $\mathbf{H}(c)$. In the case of full shrinkage ($p = m$) then (29) is equivalent to

$$\liminf_{n\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(T_n - \boldsymbol{\theta}_n\right).$$

The standard method to establish the efficiency bound (28) is to first establish the bound in the non-asymptotic normal sampling model, and then extend to the asymptotic context via the limit of experiments theory. Thus to establish (29) we need to start with a similar bound for the normal sampling model. Unfortunately, we do not have a sharp bound for this case. An important breakthrough was obtained in a result known as Pinsker's Theorem (see Nussbam (1999)) which

provides a sharp bound for the normal sampling model by taking $p \to \infty$. The existing theory has established the bound for the full shrinkage canonical model (e.g., $p = m$ and $\mathbf{W} = \mathbf{V}^{-1}$). Therefore our first goal is to extend Pinsker's Theorem to the partial shrinkage non-canonical model.

The following is a generalization of Theorem 7.28 of Wasserman (2006).

Theorem 4 *Suppose* $\mathrm{Z} \sim \mathrm{N}_m(\boldsymbol{h}, \mathbf{V})$ *and* $\lambda_p > 8$. *For any estimator* $T = T(\mathrm{Z})$,

$$\sup_{\boldsymbol{h} \in \mathbf{H}(c)} \mathbb{E}_{\boldsymbol{h}} \ell(T - \boldsymbol{h}) \geq 1 - \left[\frac{1}{1+c} + \left(\frac{2}{1+c} + 4c \right) \lambda_p^{-1/3} \right] \frac{\operatorname{tr}(\mathbf{A})}{\operatorname{tr}(\mathbf{WV})}. \tag{30}$$

Combined with the limits of experiments technique, Theorem 4 allows us to establish an asymptotic (large n) local efficiency bound for the estimation of $\boldsymbol{\theta}$ in parametric models.

Theorem 5 *Suppose that* $X_1, ..., X_n$ *is a random sample from a density* $f(x, \boldsymbol{\theta})$ *with respect to a measure* μ *indexed by a parameter* $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$, *and the density is differentiable in quadratic mean, that is*

$$\int \left[f(x, \boldsymbol{\theta} + \boldsymbol{h})^{1/2} - f(x, \boldsymbol{\theta})^{1/2} - \frac{1}{2} \boldsymbol{h}' \boldsymbol{g}(x, \boldsymbol{\theta}) f(x, \boldsymbol{\theta})^{1/2} \right]^2 d\mu = o\left(\|\boldsymbol{h}\|^2 \right), \qquad \boldsymbol{h} \to 0$$

where $\boldsymbol{g}(x, \boldsymbol{\theta}) = \dfrac{\partial}{\partial \boldsymbol{\theta}} \log f(x, \boldsymbol{\theta})$, *and* $\mathbf{I}_{\boldsymbol{\theta}} = \mathbb{E} \boldsymbol{g}(X_i, \boldsymbol{\theta}) \boldsymbol{g}(X_i, \boldsymbol{\theta})' > 0$. *If* $\lambda_p > 8$, *then for any sequence of estimators* T_n,

$$\sup_{I \subset \mathbf{H}(c)} \liminf_{n \to \infty} \sup_{\boldsymbol{h} \in I} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(T_n - \boldsymbol{\theta}_n \right) \geq 1 - \left[\frac{1}{1+c} + \left(\frac{2}{1+c} + 4c \right) \lambda_p^{-1/3} \right] \frac{\operatorname{tr}(\mathbf{A})}{\operatorname{tr}(\mathbf{WV})} \tag{31}$$

where $\mathbf{V} = \mathbf{I}_{\boldsymbol{\theta}}^{-1}$. *Thus, if as* $p \to \infty$, $\lambda_p \to \infty$ *and (25),*

$$\liminf_{p \to \infty} \sup_{I \subset \mathbf{H}(c)} \liminf_{n \to \infty} \inf_{T} \sup_{\boldsymbol{h} \in I} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(T - \boldsymbol{\theta}_n \right) \geq 1 - \frac{a}{c+1}. \tag{32}$$

Theorem 5 provides a lower bound on the asymptotic local minimax risk for $\boldsymbol{h}$ in the ball $\mathbf{H}(c)$. (31) is the case of finite p, and (32) shows that the bound takes a simple form when p is large. Since this lower bound is equal to the upper bound (26) attained by out shrinkage estimator, (32) is sharp. This proves that the shrinkage estimator is asymptotically minimax efficient over the local sets $\mathbf{H}(c)$. To our knowledge, Theorem 5 is new. It is the first large sample local efficiency bound for shrinkage estimation.

Note that the equality of (26) and (32) holds for all values of c. This is a very strong efficiency property.

7 Simulation

We illustrate the numerical magnitude of the shrinkage improvements in a simple numerical simulation. The model is a Gaussian AR(p)

$$\begin{aligned} y_t &= \mu + \alpha_1 y_{t-1} + \cdots + \alpha_{p+1} y_{t-p-1} + e_t \\ e_t &\sim \mathrm{N}\left(0, \sigma^2\right) \end{aligned}$$

We set $\mu = 0$, $\sigma^2 = 1$ and

$$\begin{aligned} a &= \alpha_1 + \cdots + \alpha_{p+1} \\ &= 0.6 \end{aligned}$$

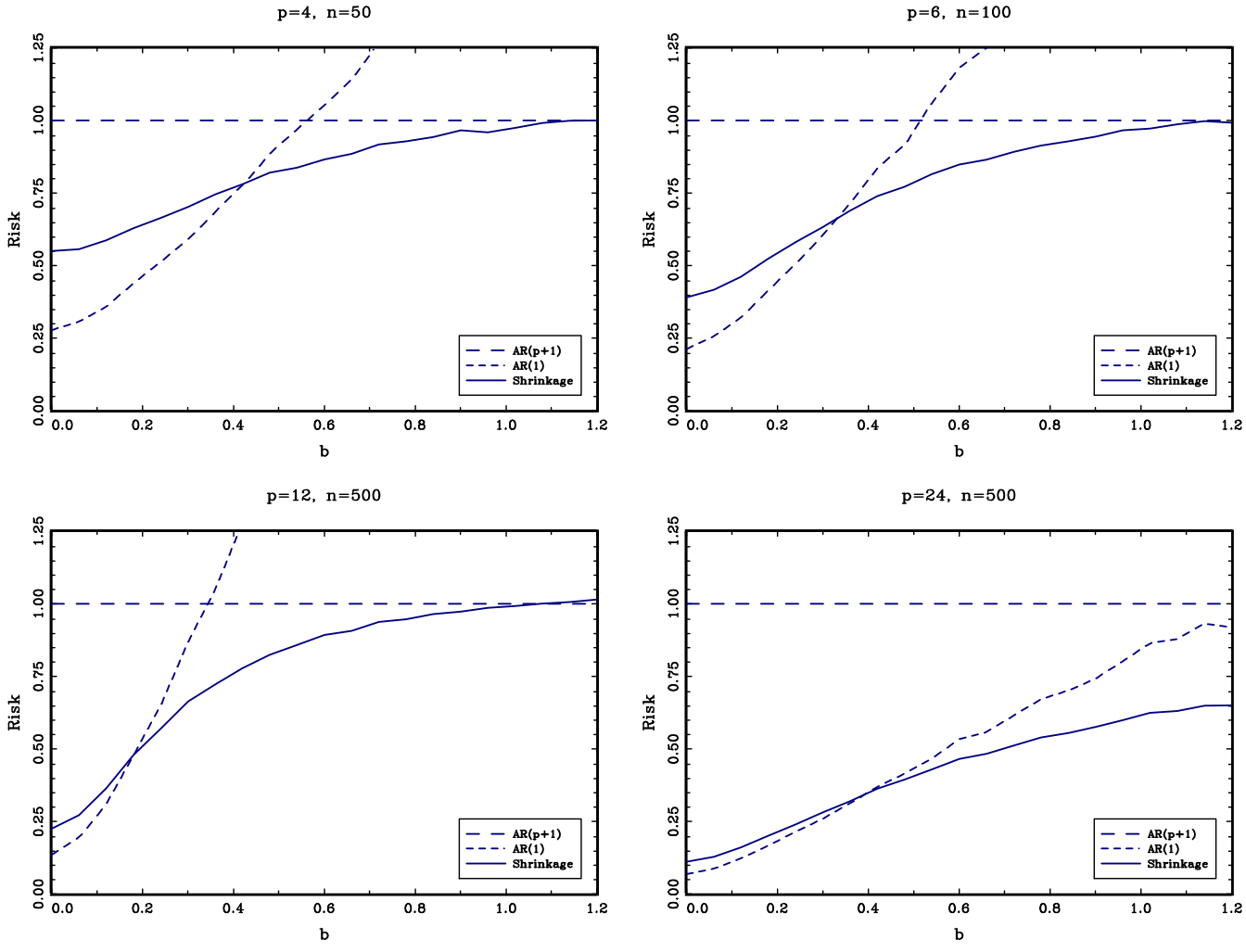
The sum of the autoregressive coefficients controls the low frequency serial dependence in y_t. For $j \geq 2$ we set

$$\alpha_j = \frac{b}{p}$$

so that $\sum_{j=2}^{p+1} \alpha_j = b$ and b is a free parameter whose value indexes the deviation of the process from an AR(1).

Our unrestricted model is the AR(p+1) and our restricted model is an AR(1) (both with intercepts) and for the weight matrix we use the canonical case. Thus our shrinkage estimator shrinks the AR(p+1) towards the AR(1). This is a reasonable context for illustration as it is common in time series to find that low dimensional models are reasonably good approximations (e.g. produce reliable forecasts) yet there is evidence that high dimensional models may be better actual descriptions of the data.

We calculate the finite sample (canonical) risk of the three estimators (unrestricted, restricted, and shrinkage) and plot the risk as a function of b (the deviation from the AR(1)) for different values of p and n in Figure 2. The risk of the unrestricted AR(p+1) is normalized to 1 and is shown with the long dashes. The risk of the restricted AR(1) is shown with the short dashes, and as might be expected, is quite low for small b (where the restriction is true or close to true) but increasing and unbounded with b. The risk of the shrinkage estimator typically lies beneath that of the unrestricted estimator, and is nearly as good as the restricted estimator for large p and n. The reductions in risk relative to the unrestricted estimator are large for wide values of the local deviation parameter b.


p=4, n=50
p=6, n=100
p=12, n=500
p=24, n=500
Risk
b
0.00
0.25
0.50
0.75
1.00
1.25
0.0
0.2
0.4
0.6
0.8
1.0
1.2
AR(p+1)
AR(1)
Shrinkage

8 Appendix

Proof of Theorem 1: (10) and (11) are standard. [Formal proof and regularity conditions to be provided.] (12), (5), and (14) follow by the continuous mapping theorem. ■

The following is a version of Stein's Lemma (Stein, 1981), and will be used in the proof of Theorem 2.

Lemma 1 *If* $\mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$ *is* $m \times 1$, $\mathbf{K}$ *is* $m \times m$, *and* $\eta(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}^m$ *is absolutely continuous, then*

$$\mathbb{E}\left(\eta\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{K}\mathrm{Z}\right) = \mathbb{E}\,\mathrm{tr}\left(\frac{\partial}{\partial \mathbf{x}}\eta\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{K}\mathbf{V}\right).$$

Proof: Let $\phi_{\mathbf{V}}(\mathbf{x})$ denote the $\mathrm{N}(\mathbf{0}, \mathbf{V})$ density function. By multivariate integration by parts

$$\begin{aligned}
\mathbb{E}\left(\eta\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{K}\mathrm{Z}\right) &= \int \eta\left(\mathbf{x}+\boldsymbol{h}\right)'\mathbf{K}\mathbf{V}\mathbf{V}^{-1}\mathbf{x}\,\phi_{\mathbf{V}}(\mathbf{x})\,(\mathbf{dx}) \\
&= \int \mathrm{tr}\left(\frac{\partial}{\partial \mathbf{x}}\eta\left(\mathbf{x}+\boldsymbol{h}\right)'\mathbf{K}\mathbf{V}\right)\phi_{\mathbf{V}}(\mathbf{x})\,(\mathbf{dx}) \\
&= \mathbb{E}\,\mathrm{tr}\left(\frac{\partial}{\partial \mathbf{x}}\eta\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{K}\mathbf{V}\right).
\end{aligned}$$

■

Proof of Theorem 2: First, take any estimator T_n which satisfies $\sqrt{n}\left(T_n - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \xi$ with $\mathbb{E}_{\boldsymbol{h}}\ell(\xi) < \infty$. Note that

$$n\ell_{\zeta/n}(T_n - \boldsymbol{\theta}_n) = \ell_{\zeta}(\sqrt{n}\left(T_n - \boldsymbol{\theta}_n\right)) \leq \zeta.$$

Then by the portmanteau lemma,

$$\begin{aligned}
\lim_{\zeta\to\infty}\liminf_{n\to\infty}\mathbb{E}_{\boldsymbol{\theta}_n} n\ell_{\zeta}(T_n - \boldsymbol{\theta}_n) &= \lim_{\zeta\to\infty}\liminf_{n\to\infty}\mathbb{E}_{\boldsymbol{\theta}_n}\ell_{\zeta}(\sqrt{n}\left(T_n - \boldsymbol{\theta}_n\right)) \qquad (33) \\
&= \lim_{\zeta\to\infty}\mathbb{E}_{\boldsymbol{h}}\ell_{\zeta}(\xi) \\
&= \mathbb{E}_{\boldsymbol{h}}\ell(\xi) \\
&= \frac{\mathbb{E}_{\boldsymbol{h}}\left(\xi'\mathbf{W}\xi\right)}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)}. \qquad (34)
\end{aligned}$$

Thus to evaluate the left-hand-side of (33) it is sufficient to calculate (34).

Next, since $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$ by (10), then

$$\lim_{\zeta\to\infty}\liminf_{n\to\infty}\mathbb{E}_{\boldsymbol{\theta}_n}\ell_{\zeta}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n) = \frac{\mathbb{E}_{\boldsymbol{h}}\left(\mathrm{Z}'\mathbf{W}\mathrm{Z}\right)}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)} = \frac{\mathrm{tr}\left(\mathbf{W}\mathbb{E}_{\boldsymbol{h}}\left(\mathrm{Z}\mathrm{Z}'\right)\right)}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)} = \frac{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)} = 1. \qquad (35)$$

Now consider the estimator

$$\widehat{\boldsymbol{\theta}}^{JS} = \widehat{\boldsymbol{\theta}} - \left(\frac{\tau_n}{D_n}\right)\left(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\right)$$

which is the same as $\widehat{\boldsymbol{\theta}}^*$ but without positive-part trimming. The pointwise risk of $\widehat{\boldsymbol{\theta}}^*$ is strictly smaller than that of $\widehat{\boldsymbol{\theta}}^{JS}$ (as shown in Theorem 5.5.4 of Lehman and Casella (1998)). Thus for all $\boldsymbol{h}$,

$$\lim_{\zeta\to\infty}\liminf_{n\to\infty}\mathbb{E}_{\boldsymbol{\theta}_n} n\ell_{\zeta/n}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_n) < \lim_{\zeta\to\infty}\liminf_{n\to\infty}\mathbb{E}_{\boldsymbol{\theta}_n} n\ell_{\zeta/n}(\widehat{\boldsymbol{\theta}}^{JS} - \boldsymbol{\theta}_n). \tag{36}$$

The estimator $\widehat{\boldsymbol{\theta}}^{JS}$ has the asymptotic distribution

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{JS} - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathrm{Z}^* = \mathrm{Z} - \left(\frac{\tau}{(\mathrm{Z}+\boldsymbol{h})'\,\mathbf{B}\,(\mathrm{Z}+\boldsymbol{h})}\right)\mathbf{V}\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\left(\mathrm{Z}+\boldsymbol{h}\right).$$

We calculate that

$$\begin{aligned}
\lim_{\zeta\to\infty}\liminf_{n\to\infty}\mathbb{E}_{\boldsymbol{\theta}_n} n\ell_{\zeta/n}(\widehat{\boldsymbol{\theta}}^{JS} - \boldsymbol{\theta}_n) &= \frac{\mathbb{E}_{\boldsymbol{h}}(\mathrm{Z}^{*\prime}\mathbf{W}\mathrm{Z}^*)}{\mathrm{tr}\,(\mathbf{W}\mathbf{V})} \\
&= \frac{\mathbb{E}_{\boldsymbol{h}}(\mathrm{Z}'\mathbf{W}\mathrm{Z})}{\mathrm{tr}\,(\mathbf{W}\mathbf{V})} \\
&\quad + \frac{\tau^2}{\mathrm{tr}\,(\mathbf{W}\mathbf{V})}\mathbb{E}_{\boldsymbol{h}}\left(\frac{(\mathrm{Z}+\boldsymbol{h})'\,\mathbf{R}\,(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\,\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}\mathbf{R}\,(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\,\mathbf{R}'\,(\mathrm{Z}+\boldsymbol{h})}{\left((\mathrm{Z}+\boldsymbol{h})'\,\mathbf{B}\,(\mathrm{Z}+\boldsymbol{h})\right)^2}\right) \\
&\quad - 2\frac{\tau}{\mathrm{tr}\,(\mathbf{W}\mathbf{V})}\mathbb{E}_{\boldsymbol{h}}\left(\frac{(\mathrm{Z}+\boldsymbol{h})'\,\mathbf{R}\,(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\,\mathbf{R}'\mathbf{V}\mathbf{W}\mathrm{Z}}{(\mathrm{Z}+\boldsymbol{h})'\,\mathbf{B}\,(\mathrm{Z}+\boldsymbol{h})}\right) \\
&= 1 + \frac{\tau^2}{\mathrm{tr}\,(\mathbf{W}\mathbf{V})}\mathbb{E}_{\boldsymbol{h}}\left(\frac{1}{(\mathrm{Z}+\boldsymbol{h})'\,\mathbf{B}\,(\mathrm{Z}+\boldsymbol{h})}\right) \\
&\quad - 2\frac{\tau}{\mathrm{tr}\,(\mathbf{W}\mathbf{V})}\mathbb{E}_{\boldsymbol{h}}\left(\eta(\mathrm{Z}+\boldsymbol{h})'\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{W}\mathrm{Z}\right)
\end{aligned} \tag{37}$$

where

$$\eta(\mathbf{x}) = \left(\frac{1}{\mathbf{x}'\mathbf{B}\mathbf{x}}\right)\mathbf{x}.$$

Since

$$\frac{\partial}{\partial\mathbf{x}}\eta(\mathbf{x})' = \left(\frac{1}{\mathbf{x}'\mathbf{B}\mathbf{x}}\right)\mathbf{I} - \frac{2}{(\mathbf{x}'\mathbf{B}\mathbf{x})^2}\mathbf{B}\mathbf{x}\mathbf{x}',$$

then by Lemma 1 (Stein's Lemma)

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left(\eta(\mathrm{Z}+\boldsymbol{h})'\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{W}\mathrm{Z}\right) &= \mathbb{E}_{\boldsymbol{h}}\,\mathrm{tr}\left(\frac{\partial}{\partial\mathbf{x}}\eta\,(\mathrm{Z}+\boldsymbol{h})'\,\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}\right) \\
&= \mathbb{E}_{\boldsymbol{h}}\,\mathrm{tr}\left(\frac{\mathbf{R}\,(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\,\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}}{(\mathrm{Z}+\boldsymbol{h})'\,\mathbf{B}\,(\mathrm{Z}+\boldsymbol{h})}\right) \\
&\quad - 2\mathbb{E}_{\boldsymbol{h}}\,\mathrm{tr}\left(\frac{\mathbf{B}\,(\mathrm{Z}+\boldsymbol{h})\,(\mathrm{Z}+\boldsymbol{h})'\,\mathbf{R}\,(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\,\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}}{\left((\mathrm{Z}+\boldsymbol{h})'\,\mathbf{B}\,(\mathrm{Z}+\boldsymbol{h})\right)^2}\right).
\end{aligned}$$

Using the inequality $\operatorname{tr}(\mathbf{CD}) \leq \lambda_{\max}(\mathbf{C}) \operatorname{tr}(\mathbf{D})$, this is larger than

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{h}} \operatorname{tr}\left(\frac{\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right) \\ & -2\mathbb{E}_{\boldsymbol{h}} \operatorname{tr}\left(\frac{\mathbf{B}(\mathrm{Z}+\boldsymbol{h})(\mathrm{Z}+\boldsymbol{h})'}{\left((\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})\right)^2}\right) \lambda_{\max}\left(\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}\right) \\ = \quad & \mathbb{E}_{\boldsymbol{h}}\left(\frac{\operatorname{tr}(\mathbf{A})-2\lambda_{\max}(\mathbf{A})}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right). \end{aligned}$$

Thus (37) is smaller than

$$\begin{aligned} & 1+\frac{\tau^2}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}_{\boldsymbol{h}}\left(\frac{1}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right)-2\frac{\tau}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}_{\boldsymbol{h}}\left(\frac{\operatorname{tr}(\mathbf{A})-2\lambda_{\max}(\mathbf{A})}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right) \\ = \quad & 1-\frac{\tau}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}_{\boldsymbol{h}}\left(\frac{2\left(\operatorname{tr}(\mathbf{A})-2\lambda_{\max}(\mathbf{A})\right)-\tau}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right) \\ \leq \quad & 1-\frac{\tau}{\operatorname{tr}(\mathbf{WV})}\frac{2\left(\operatorname{tr}(\mathbf{A})-2\lambda_{\max}(\mathbf{A})\right)-\tau}{\mathbb{E}_{\boldsymbol{h}}\left((\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})\right)} \end{aligned} \tag{38}$$

where the second inequality is Jensen's and uses the assumption that $\tau < 2\left(\operatorname{tr}(\mathbf{A})-2\lambda_{\max}(\mathbf{A})\right)$.

We calculate that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{h}}\left((\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})\right) &= \boldsymbol{h}'\mathbf{B}\boldsymbol{h}+\mathbb{E}_{\boldsymbol{h}}\operatorname{tr}\left(\mathbf{B}\mathrm{Z}\mathrm{Z}'\right) \\ &= \boldsymbol{h}'\mathbf{B}\boldsymbol{h}+\operatorname{tr}(\mathbf{A}) \\ &\leq (c+1)\operatorname{tr}(\mathbf{A}) \end{aligned}$$

where the inequality is for $\boldsymbol{h} \in \mathbf{H}(c)$. Substituting into (38) and using (36) we find

$$\lim_{\zeta\to\infty}\liminf_{n\to\infty}\mathbb{E}_{\boldsymbol{\theta}_n}\ell_\zeta(\widehat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}_n) < 1-\frac{\tau}{\operatorname{tr}(\mathbf{WV})}\frac{2\left(\operatorname{tr}(\mathbf{A})-2\lambda_{\max}(\mathbf{A})\right)-\tau}{(c+1)\operatorname{tr}(\mathbf{A})}$$

which is (23). As this bound is strictly less than 1, combined with (35) we have established (21). ■

Proof of Theorem 4. Without loss of generality we can set $\mathbf{V} = \mathbf{I}_m$ and $\mathbf{R} = \begin{pmatrix} \mathbf{0}_{m-p} \\ \mathbf{I}_p \end{pmatrix}$. To see this, start by making the transformations $\boldsymbol{h} \mapsto \mathbf{V}^{-1/2}\boldsymbol{h}$, $\mathbf{R} \mapsto \mathbf{V}^{1/2}\mathbf{R}$, and $\mathbf{W} \mapsto \mathbf{V}^{1/2}\mathbf{W}\mathbf{V}^{1/2}$ so that $\mathbf{V} = \mathbf{I}_m$. Then write $\mathbf{R} = \mathbf{Q}\begin{pmatrix} \mathbf{0}_{m-p} \\ \mathbf{I}_p \end{pmatrix}\mathbf{G}$ where $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_p$ and $\mathbf{G}$ is full rank. Make the transformations $\boldsymbol{h} \mapsto \mathbf{Q}'\boldsymbol{h}$, $\mathbf{R} \mapsto \mathbf{Q}'\mathbf{R}\mathbf{G}^{-1}$ and $\mathbf{W} \mapsto \mathbf{Q}\mathbf{W}\mathbf{Q}'$. Hence $\mathbf{V} = \mathbf{I}_m$ and $\mathbf{R} = \begin{pmatrix} \mathbf{0}_{m-p} \\ \mathbf{I}_p \end{pmatrix}$ as claimed.

Partition $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2)$, $T = (T_1, T_2)$, $\mathrm{Z} = (\mathrm{Z}_1, \mathrm{Z}_2)$ and $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}$ comformably with

R. Note that after these transformations $\mathbf{A} = \mathbf{W}_{22}$ and $\mathbf{H}(c) = \left\{\boldsymbol{h} : \boldsymbol{h}_2'\mathbf{W}_{22}\boldsymbol{h}_2 \leq \operatorname{tr}(\mathbf{W}_{22})\, c\right\}$.

Set $\eta = 1 - 2\lambda_p^{-1/3}$ and note that $0 < \eta < 1$ since $\lambda_p > 8$. Fix $\omega > 0$. Let $\Pi_1(\boldsymbol{h}_1)$ and $\Pi_2(\boldsymbol{h}_2)$ be the independent priors $\boldsymbol{h}_1 \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{m-p}\omega)$ and $\boldsymbol{h}_2 \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_p c\eta)$. Let $\widetilde{T}_1 = \mathbb{E}(\boldsymbol{h}_1 \mid \mathrm{Z})$ and $\widetilde{T}_2 = \mathbb{E}(\boldsymbol{h}_2 \mid \mathrm{Z})$ be the Bayes estimators of $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ under these priors. By standard calculations, $\widetilde{T}_1 = \dfrac{\omega}{1+\omega}\mathrm{Z}_1$ and $\widetilde{T}_2 = \dfrac{c\eta}{1+c\eta}\mathrm{Z}_2$. Also, let $\Pi_2^*(\boldsymbol{h}_2)$ be the prior $\Pi_2(\boldsymbol{h}_2)$ truncated to the region $\mathbf{H}_2(c) = \left\{\boldsymbol{h}_2 : \boldsymbol{h}_2'\mathbf{W}_{22}\boldsymbol{h}_2 \leq \operatorname{tr}(\mathbf{W}_{22})\, c\right\}$, and let $\widetilde{T}_2^* = \mathbb{E}(\boldsymbol{h}_2 \mid \mathrm{Z})$ be the Bayes estimator of $\boldsymbol{h}_2$ under this truncated prior. Since a Bayes estimator must lie in the prior support, it follows that $\widetilde{T}_2^* \in \mathbf{H}_2(c)$ or

$$\widetilde{T}_2^{*\prime}\mathbf{W}_{22}\widetilde{T}_2^* \leq \operatorname{tr}(\mathbf{W}_{22})\, c. \tag{39}$$

Also, since Z_1 and Z_2 are independent, and Π_1 and Π_2^* are independent, it follows that $\widetilde{T}_2^*$ is a function of Z_2 only, and $\widetilde{T}_1 - \boldsymbol{h}_1$ and $\widetilde{T}_2^* - \boldsymbol{h}_2$ are independent.

Set $\widetilde{T} = \left(\widetilde{T}_1, \widetilde{T}_2^*\right)$. For any estimator T, since a supremum is larger than an average and the support of $\Pi_1 \times \Pi_2^*$ is $\mathbf{H}(c)$,

$$\begin{aligned}
\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \mathbb{E}_{\boldsymbol{h}}\ell(T-\boldsymbol{h}) &\geq \int\int \mathbb{E}_{\boldsymbol{h}}\ell(T-\boldsymbol{h})d\Pi_1(\boldsymbol{h}_1)d\Pi_2^*(\boldsymbol{h}_2) && (40)\\
&\geq \int\int \mathbb{E}_{\boldsymbol{h}}\ell(\widetilde{T}-\boldsymbol{h})d\Pi_1(\boldsymbol{h}_1)d\Pi_2^*(\boldsymbol{h}_2)\\
&= \frac{1}{\operatorname{tr}(\mathbf{W})}\int\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_1-\boldsymbol{h}_1\right)'\mathbf{W}_{11}\left(\widetilde{T}_1-\boldsymbol{h}_1\right)\right]d\Pi_1(\boldsymbol{h}_1)d\Pi_2^*(\boldsymbol{h}_2)\\
&\quad + \frac{2}{\operatorname{tr}(\mathbf{W})}\int\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_1-\boldsymbol{h}_1\right)'\mathbf{W}_{12}\left(\widetilde{T}_2^*-\boldsymbol{h}_2\right)\right]d\Pi_1(\boldsymbol{h}_1)d\Pi_2^*(\boldsymbol{h}_2)\\
&\quad + \frac{1}{\operatorname{tr}(\mathbf{W})}\int\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_2^*-\boldsymbol{h}_2\right)'\mathbf{W}_{22}\left(\widetilde{T}_2^*-\boldsymbol{h}_2\right)\right]d\Pi_1(\boldsymbol{h}_1)d\Pi_2^*(\boldsymbol{h}_2) && (41)\\
&= \frac{1}{\operatorname{tr}(\mathbf{W})}\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_1-\boldsymbol{h}_1\right)'\mathbf{W}_{11}\left(\widetilde{T}_1-\boldsymbol{h}_1\right)\right]d\Pi_1(\boldsymbol{h}_1) && (42)\\
&\quad + \frac{2}{\operatorname{tr}(\mathbf{W})}\left(\int \mathbb{E}_{\boldsymbol{h}}\left(\widetilde{T}_1-\boldsymbol{h}_1\right)d\Pi_1(\boldsymbol{h}_1)\right)'\mathbf{W}_{12}\left(\int\left(\widetilde{T}_2^*-\boldsymbol{h}_2\right)d\Pi_2^*(\boldsymbol{h}_2)\right) && (43)\\
&\quad + \frac{1}{\operatorname{tr}(\mathbf{W})}\frac{\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_2^*-\boldsymbol{h}_2\right)'\mathbf{W}_{22}\left(\widetilde{T}_2^*-\boldsymbol{h}_2\right)\right]d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} && (44)\\
&\quad - \frac{\int_{\mathbf{H}_2(c)^c} \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_2^*-\boldsymbol{h}_2\right)'\mathbf{W}_{22}\left(\widetilde{T}_2^*-\boldsymbol{h}_2\right)\right]d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} && (45)
\end{aligned}$$

where the second inequality is because the Bayes estimator $\widetilde{T}$ minimizes the right-hand-side of (40). The final equality uses the fact that $\widetilde{T}_1 - \boldsymbol{h}_1$ and $\widetilde{T}_2^* - \boldsymbol{h}_2$ are independent, and breaks the integral (41) over the truncated prior (which has support on $\mathbf{H}_2(c)$) into the difference of the integrals over the non-truncated prior over the $\mathbb{R}^m$ and $\mathbf{H}_2(c)^c$, respectively. We now treat the four components (42)-(45) separately.

First, since $\widetilde{T}_1 = \dfrac{\omega}{1+\omega}\mathrm{Z}_1$ and $\Pi_1(\boldsymbol{h}_1) = \mathrm{N}\left(\mathbf{0}, \mathbf{I}_{m-p}\omega\right)$, we calculate that

$$\begin{aligned}
& \frac{1}{\operatorname{tr}(\mathbf{W})}\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_1 - \boldsymbol{h}_1\right)'\mathbf{W}_{11}\left(\widetilde{T}_1 - \boldsymbol{h}_1\right)\right] d\Pi_1(\boldsymbol{h}_1) \\
= & \frac{1}{\operatorname{tr}(\mathbf{W})}\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\frac{\omega}{1+\omega}\mathrm{Z}_1 - \boldsymbol{h}_1\right)'\mathbf{W}_{11}\left(\frac{\omega}{1+\omega}\mathrm{Z}_1 - \boldsymbol{h}_1\right)\right] d\Pi_1(\boldsymbol{h}_1) \\
= & \frac{1}{\operatorname{tr}(\mathbf{W})}\int \left[\frac{1}{(1+\omega)^2}\boldsymbol{h}_1'\mathbf{W}_{11}\boldsymbol{h}_1 + \frac{\omega^2}{(1+\omega)^2}\operatorname{tr}(\mathbf{W}_{11})\right] d\Pi_1(\boldsymbol{h}_1) \\
= & \frac{\operatorname{tr}(\mathbf{W}_{11})}{\operatorname{tr}(\mathbf{W})}\frac{\omega}{1+\omega}. \qquad (46)
\end{aligned}$$

Second, since

$$\int \mathbb{E}_{\boldsymbol{h}}\left(\widetilde{T}_1 - \boldsymbol{h}_1\right) d\Pi_1(\boldsymbol{h}_1) = -\frac{1}{1+\omega}\int \boldsymbol{h}_1 d\Pi_1(\boldsymbol{h}_1) = 0$$

it follows that (43) equals zero.

Third, take (44). Because $\widetilde{T}_2$ is the Bayes estimator under the prior Π_2,

$$\begin{aligned}
& \frac{1}{\operatorname{tr}(\mathbf{W})}\frac{\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)'\mathbf{W}_{22}\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} \\
\geq & \frac{1}{\operatorname{tr}(\mathbf{W})}\frac{\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_2 - \boldsymbol{h}_2\right)'\mathbf{W}_{22}\left(\widetilde{T}_2 - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} \\
\geq & \frac{1}{\operatorname{tr}(\mathbf{W})}\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_2 - \boldsymbol{h}_2\right)'\mathbf{W}_{22}\left(\widetilde{T}_2 - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2) \\
= & \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})}\frac{c\eta}{1+c\eta} \qquad (47) \\
= & \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})}\left(\frac{c}{1+c} - \frac{2\lambda_p^{-1/3}}{1+c}\right) \qquad (48)
\end{aligned}$$

where (47) is a calculation similar to (46) using $\widetilde{T}_2 = \dfrac{c\eta}{1+c\eta}\mathrm{Z}_2$ and $\boldsymbol{h}_2 \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}_p c\eta\right)$. (48) makes a simple expansion using $\eta = 1 - 2\lambda_p^{-1/3}$.

Fourth, take (45). Our goal is to show that this term is negligible for large p, and our argument is based on the proof Theorem 7.28 of Wasserman (2006). Set

$$q = \frac{\boldsymbol{h}_2'\mathbf{W}_{22}\boldsymbol{h}_2}{c\operatorname{tr}(\mathbf{W}_{22})}.$$

Since $\boldsymbol{h}_2 \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_p c\eta)$ we see that $\mathbb{E}q = \eta$. Use $(\boldsymbol{a}+\boldsymbol{b})'(\boldsymbol{a}+\boldsymbol{b}) \leq 2\boldsymbol{a}'\boldsymbol{a} + 2\boldsymbol{b}'\boldsymbol{b}$ and (39) to find that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)' \mathbf{W}_{22}\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)\right] &\leq 2\mathbb{E}_{\boldsymbol{h}}\left(\widetilde{T}_2^{*\prime}\mathbf{W}_{22}\widetilde{T}_2^*\right) + 2\boldsymbol{h}_2'\mathbf{W}_{22}\boldsymbol{h}_2 \\
&\leq 2\,\mathrm{tr}\left(\mathbf{W}_{22}\right)c + 2\boldsymbol{h}_2'\mathbf{W}_{22}\boldsymbol{h}_2 \\
&= 2\,\mathrm{tr}\left(\mathbf{W}_{22}\right)c\,(1+q) \\
&\leq 2\,\mathrm{tr}\left(\mathbf{W}_{22}\right)c\,(2+q-\eta)\,. \qquad (49)
\end{aligned}$$

Note that $\boldsymbol{h}_2 \in \mathbf{H}_2(c)^c$ is equivalent to $q > 1$. Using (49) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\int_{\mathbf{H}_2(c)^c} \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)' \mathbf{W}_{22}\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2) \\
\leq\ & 2\,\mathrm{tr}\left(\mathbf{W}_{22}\right)c\left[2\int_{\mathbf{H}_2(c)^c} d\Pi_2(\boldsymbol{h}_2) + \int_{\mathbf{H}_2(c)^c}(q-\eta)\,d\Pi_2(\boldsymbol{h}_2)\right] \\
\leq\ & 2\,\mathrm{tr}\left(\mathbf{W}_{22}\right)c\left[2\mathbb{P}\left(q>1\right) + \mathrm{var}\left(q\right)^{1/2}\mathbb{P}\left(q>1\right)^{1/2}\right]. \qquad (50)
\end{aligned}$$

Letting w_j denote the eigenvalues of $\mathbf{W}_{22}$ then we can write

$$q - \mathbb{E}q = \frac{\eta}{\sum_{j=1}^p w_j}\sum_{j=1}^p w_j\left(y_j^2 - 1\right)$$

where y_j are iid $N(0,1)$. Thus

$$\mathrm{var}\left(q\right) = \frac{\eta^2}{\left(\sum_{j=1}^p w_j\right)^2}\sum_{j=1}^p w_j^2\,\mathrm{var}\left(y_j^2\right) \leq 2\lambda_p^{-1} \qquad (51)$$

since $\lambda_p = \dfrac{\sum_{j=1}^p w_j}{\max_j w_j} = \mathrm{tr}\left(\mathbf{W}_{22}\right)/\lambda_{\max}\left(\mathbf{W}_{22}\right)$. By Markov's inequality, (51), and $1-\eta = 2\lambda_p^{-1/3}$,

$$\mathbb{P}\left(q>1\right) = \mathbb{P}\left(q-\eta > 1-\eta\right) \leq \frac{\mathrm{var}\left(q\right)}{\left(1-\eta\right)^2} \leq \frac{\lambda_p^{-1/3}}{2}. \qquad (52)$$

Furthermore, (52) and $\lambda_p^{-1/3} \leq 2^{-1}$ imply that

$$\begin{aligned}
\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2) &= 1 - \mathbb{P}\left(q>1\right) \\
&\geq 1 - \frac{\lambda_p^{-1/3}}{2} \\
&\geq \frac{3}{4}. \qquad (53)
\end{aligned}$$

It follows from (50), (51), (52), (53) and $\lambda_p^{-1/3} \leq 2^{-1}$ that

$$\begin{aligned} & \frac{1}{\operatorname{tr}(\mathbf{W})} \frac{\int_{\mathbf{H}_2(c)^c} \mathbb{E}_{\boldsymbol{h}} \left[\left(\widetilde{T}_2^* - \boldsymbol{h}_2 \right)' \mathbf{W}_{22} \left(\widetilde{T}_2^* - \boldsymbol{h}_2 \right) \right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} \\ \leq & \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})} \frac{2c\left(\lambda_p^{-1/3} + \lambda_p^{-2/3}\right)}{3/4} \\ \leq & \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})} 4c\lambda_p^{-1/3} \end{aligned} \tag{54}$$

Together, (46) and (54) applied to (42)-(44) show that

$$\sup_{\boldsymbol{h} \in \mathbf{H}(c)} \mathbb{E}_{\boldsymbol{h}} \ell(T - \boldsymbol{h}) \geq \frac{\omega}{1+\omega} \frac{\operatorname{tr}(\mathbf{W}_{11})}{\operatorname{tr}(\mathbf{W})} + \left(\frac{c}{1+c} - \left(\frac{2}{1+c} + 4c \right) \lambda_p^{-1/3} \right) \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})}.$$

Since ω is arbitrary we conclude that

$$\begin{aligned} \sup_{\boldsymbol{h} \in \mathbf{H}(c)} \mathbb{E}_{\boldsymbol{h}} \ell(T - \boldsymbol{h}) & \geq \frac{\operatorname{tr}(\mathbf{W}_{11})}{\operatorname{tr}(\mathbf{W})} + \left(\frac{c}{1+c} - \left(\frac{2}{1+c} + 4c \right) \lambda_p^{-1/3} \right) \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})} \\ & = 1 - \left(\frac{1}{1+c} + \left(\frac{2}{1+c} + 4c \right) \lambda_p^{-1/3} \right) \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})} \end{aligned}$$

which is (30) since $\dfrac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})} = \dfrac{\operatorname{tr}(\mathbf{A})}{\operatorname{tr}(\mathbf{WV})}$ under the transformations made at the beginning of the proof.

The innovation in the proof technique (relative, for example, to the arguments of van der Vaart (1998) and Wasserman (2006)) is the use of the Bayes estimator $\widetilde{T}_2^*$ based on the truncated prior Π_2^*. ■

Proof of Theorem 5. The proof technique closely follows the argument in Theorem 8.11 of van der Vaart (1998). The only difference is the last step, where we bound the risk of the limiting experiment using Theorem 4 rather than van der Vaart's Proposition 8.6.

Let $\mathbb{Q}$ denote the rational vectors in $\mathbf{H}(c)$ placed in arbitrary order, and let I_k denote the first k vectors in this sequence. Then the left side of (31) is larger than

$$\lim_{k \to \infty} \liminf_{n \to \infty} \sup_{\boldsymbol{h} \in I_k} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(T_n - \boldsymbol{\theta}_n\right). \tag{55}$$

There exists a subsequence $\{n_k\}$ of $\{n\}$ such that this expression is equal to

$$\lim_{k \to \infty} \sup_{\boldsymbol{h} \in I_k} \mathbb{E}_{\boldsymbol{\theta}_{n_k}} n_k \ell\left(T_{n_k} - \boldsymbol{\theta}_{n_k}\right) \geq \lim_{k \to \infty} \sup_{\boldsymbol{h} \in I_K} \mathbb{E}_{\boldsymbol{\theta}_{n_k}} n_k \ell\left(T_{n_k} - \boldsymbol{\theta}_{n_k}\right) \tag{56}$$

for any $K < \infty$.

For simplicity, assume that the sequence $\sqrt{n_k}\left(T_{n_k}-\boldsymbol{\theta}\right)$ is tight for fixed $\boldsymbol{\theta}$. (Van der Vaart and Wellner (1996) show that a compactification device can be used to induce tightness.) Given tightness, Prohorov's theorem shows that there is a subsequence $\{n_j\}$ of $\{n_k\}$ along which $\sqrt{n_j}\left(T_{n_j}-\boldsymbol{\theta}\right)$ converges in distribution. By the asymptotic normality of the local limit experiment, and Le Cam's third lemma, the same is true for the sequence $\sqrt{n_j}\left(T_{n_j}-\boldsymbol{\theta}_{n_j}\right)$. Thus (56) equals

$$\lim_{j\to\infty}\sup_{\boldsymbol{h}\in I_K}\mathbb{E}_{\boldsymbol{\theta}_{n_j}} n_j\ell\left(T_{n_j}-\boldsymbol{\theta}_{n_j}\right). \tag{57}$$

By Van der Vaart, Theorem 8.3, the limiting distribution of $\sqrt{n_j}\left(T_{n_j}-\boldsymbol{\theta}_{n_j}\right)$ takes the form

$$\sqrt{n_j}\left(T_{n_j}-\boldsymbol{\theta}_{n_j}\right)\xrightarrow[\boldsymbol{\theta}_{n_j}]{} T-\boldsymbol{h}$$

where $T=T\left(\mathrm{Z}\right)$ is a (possibly randomized) estimator of $\boldsymbol{h}$ based on $\mathrm{Z}\sim\mathrm{N}_m\left(\boldsymbol{h},\mathbf{V}\right)$. By the portmanteau lemma, for every $\boldsymbol{h}$,

$$\liminf_{j\to\infty}\mathbb{E}_{\boldsymbol{\theta}_{n_j}} n_j\ell\left(T_{n_j}-\boldsymbol{\theta}_{n_j}\right)=\liminf_{j\to\infty}\mathbb{E}_{\boldsymbol{\theta}_{n_j}}\ell\left(\sqrt{n_j}\left(T_{n_j}-\boldsymbol{\theta}_{n_j}\right)\right)\geq\mathbb{E}_{\boldsymbol{h}}\ell(T-\boldsymbol{h})$$

and since I_K is a finite set it follows that (57) is larger than $\sup_{\boldsymbol{h}\in I_K}\mathbb{E}_{\boldsymbol{h}}\ell(T-\boldsymbol{h})$. Since K is arbitrary, we deduce that (55) is larger than

$$\begin{aligned}\sup_{\boldsymbol{h}\in\mathbb{Q}}\mathbb{E}_{\boldsymbol{h}}\ell(T-\boldsymbol{h}) &= \sup_{\boldsymbol{h}\in\mathbf{H}(c)}\mathbb{E}_{\boldsymbol{h}}\ell(T-\boldsymbol{h})\\ &\geq 1-\left[\frac{1}{1+c}+\left(\frac{2}{1+c}+4c\right)\lambda_p^{-1/3}\right]\frac{\mathrm{tr}\left(\mathbf{A}\right)}{\mathrm{tr}\left(\mathbf{WV}\right)}\end{aligned}$$

where the equality holds since the loss function is continuous in $\boldsymbol{h}$, and the inequality is Theorem 4. We have shown (31).

Equivalently

$$\sup_{I\subset\mathbf{H}(c)}\liminf_{n\to\infty}\inf_{T}\sup_{\boldsymbol{h}\in I}\mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(T-\boldsymbol{\theta}_n\right)\geq 1-\left[\frac{1}{1+c}+\left(\frac{2}{1+c}+4c\right)\lambda_p^{-1/3}\right]\frac{\mathrm{tr}\left(\mathbf{A}\right)}{\mathrm{tr}\left(\mathbf{WV}\right)}.$$

For each c, taking the limit as $p\to\infty$, we obtain

$$\liminf_{p\to\infty}\sup_{I\subset\mathbf{H}(c)}\liminf_{n\to\infty}\inf_{T}\sup_{\boldsymbol{h}\in I}\mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(T-\boldsymbol{\theta}_n\right)\geq 1-\frac{a}{1+c}$$

which is (32). ■

References

[1] Baranchick, A. (1964): "Multiple regression and estimation of the mean of a multivariate normal distribution," Technical Report No. 51, Department of Statistics, Stanford University.

[2] Casella, George and J.T.G. Hwang (1982): "Limit expressions for the risk of James-Stein estimators," *Canadian Journal of Statistics*, 10, 305-309.

[3] Hájek, J. (1970): A characterization of limiting distributions of regular estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14, 323-330.,

[4] Hájek, J. (1972): "Local asymptotic minimax and admissibility in estimation," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 175-194.

[5] Hansen, Bruce E. (2007): "Least squares model averaging," *Econometrica*, 75, 1175-1189.

[6] Hjort, Nils Lid and Gerda Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879-899.

[7] James W. and Charles M. Stein (1961): "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361-380.

[8] Judge, George and M. E. Bock (1978): *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*, North-Holland.

[9] Lehmann, E.L. and George Casella (1998): *Theory of Point Estimation*, 2nd Edition, New York: Springer.

[10] Liu, Chu-An (2011): "A Plug-In Averaging Estimator for Regressions with Heteroskedastic Errors," Department of Economics, University of Wisconsin.

[11] Magnus, Jan R. and Heinz Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: Wiley.

[12] Newey, Whitney K. and Kenneth D. West (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.

[13] Nussbam, Michael (1999): "Minimax risk: Pinsker bound," *Encyclopedia of Statistical Sciences*, Update Volume 3, 451-460 (S. Kotz, Ed.), John Wiley, New York

[14] Oman, Samuel D. (1982a): "Contracting towards subspaces when estimating the mean of a multivariate normal distribution," *Journal of Multivariate Analysis*, 12, 270-290.

[15] Oman, Samuel D. (1982b): "Shrinking towards subspaces in multiple linear regression," *Technometrics*, 24, 307-311.

[16] Saleh, A. K. Md. Ehsanses (2006): *Theory of Preliminary Test and Stein-Type Estimation with Applications*, Hoboken, Wiley.

[17] Stein, Charles M. (1956): "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proc. Third Berkeley Symp. Math. Statist. Probab.*, 1, 197-206.

[18] Stein, Charles M. (1981): "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, 9, 1135-1151.

[19] van der Vaart, Aad W. (1998): *Asymptotic Statistics*, New York: Cambridge University Press.

[20] van der Vaart, Aad W. and Jon A. Wellner (1996): *Weak Convergence and Empirical Processes*, New York: Springer.

[21] Wasserman, Larry (2006): *All of Nonparametric Statistics*, New York: Springer.