Efficient Shrinkage in Parametric Models

Bruce E. Hansen*
University of Wisconsin†

September 2012
Revised: February 2013

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 a 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 risk 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 (Hájek, 1972) and local minimax efficiency theory for Gaussian models (Pinsker, 1980).

*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 routinely asserted that the conventional MLE is 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 subsequent 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 so we allow for this possibility. Weighted MSE is a standard criteria in the shrinkage literature, including Bhattacharya (1966), Sclove (1968), and Berger (1976a, 1976b, 1982). What is different about our approach relative to these papers is that our estimator does not require the weight matrix to be positive definite. This may be particularly important in econometric applications where nuisance parameters are commonplace.

We benefit from the recent theory of efficient high-dimensional Gaussian shrinkage, specifically Pinkser's Theorem (Pinsker, 1980), 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 (Hájek, 1972) 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 estimator equals this lower bound when the shrinkage dimension diverges towards infinity. This shows that the proposed shrinkage estimator is minimax efficient in high-dimensional models.

There are limitations to the theory presented in this paper. First, our efficiency theory is confined to parametric models, while most econometric applications are semi-parametric. Second, our efficiency theory for high-dimensional models employs a sequential asymptotic argument. A deeper theory would employ a joint asymptotic limit. Third, our analysis is confined to weighted quadratic loss functions. Fourth, we do not provide methods for confidence interval construction or inference after shrinkage. These four limitations are important, yet pose difficult technical challenges, and raise issues which hopefully can be addressed in future research.

The literature on shrinkage estimation is 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).

The theory of efficient high-dimensional Gaussian shrinkage is credited to Pinsker (1980), though Beran (2010) points out that the idea has antecedents in Stein (1956). Reviews are provided by Nussbaum (1999) and Wasserman (2006, chapter 7). Extensions to asymptotically Gaussian regression have been made by Golubev (1991), Golubev and Nussbaum (1990), and Efromovich (1996).

There also has been a recent explosion of interest in the Lasso (Tibshirani, 1996) and its variants, which simultaneously selects variables and shrinks coefficients in linear regression. Lasso methods are complementary to shrinkage but have important conceptual differences. The Lasso is known to work well in sparse models (high-dimensional models with a true small-dimensional structure). In contrast, shrinkage methods do not exploit sparsity, and can work well when there are many non-zero but small parameters. Furthermore, Lasso has been previously developed exclusively for regression models, while this paper focuses on likelihood models.

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 high-dimensional approximation, showing that the gains are substantial and broad in the parameters space. Section 6 presents a new local minimax efficiency bound. Section 7 illustrates the performance in a simulation experience using a probit model. Mathematical proofs are left to 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 $m \times p$ 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 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

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

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

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

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

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

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

$$\ell(\mathbf{u}) = \frac{\mathbf{u}'\mathbf{W}\mathbf{u}}{\mathrm{tr}(\mathbf{V}\mathbf{W})} \tag{4}$$

and define the loss of an estimator T_n for the parameter $\boldsymbol{\theta}$ as

$$\ell\left(T_n - \boldsymbol{\theta}\right) = \frac{\left(T_n - \boldsymbol{\theta}\right)' \mathbf{W} \left(T_n - \boldsymbol{\theta}\right)}{\mathrm{tr}(\mathbf{V}\mathbf{W})}.$$

We have scaled the quadratic function by $\mathrm{tr}(\mathbf{V}\mathbf{W})$ as a normalization so that the asymptotic risk of the MLE is unity. The weight matrix $\mathbf{W}$ need not be positive definite, and indeed this is appropriate when a subset of $\boldsymbol{\theta}$ are nuisance parameters.

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 (4) 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 proposed shrinkage estimator of $\boldsymbol{\theta}$ is the weighted average of the MLE and restricted MLE

$$\widehat{\boldsymbol{\theta}}_n^* = \hat{w}\widehat{\boldsymbol{\theta}}_n + (1 - \hat{w})\widetilde{\boldsymbol{\theta}}_n \tag{5}$$

where

$$\hat{w} = \left(1 - \frac{\tau}{D_n}\right)_+ \tag{6}$$

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

$$D_n = n\left(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n\right)' \mathbf{W} \left(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n\right), \tag{7}$$

a distance-type statistic for the restriction (1) in $\boldsymbol{\Theta}$. The scalar $\tau \geq 0$ controls the degree of shrinkage. In practice, if $\mathbf{W}$ and/or τ are replaced with consistent estimates $\widehat{\mathbf{W}}$ and $\hat{\tau}$ our asymptotic theory is unaffected.

In the canonical case ($\mathbf{W} = \mathbf{V}^{-1}$) we recommend $\tau = p$ or $\tau = p - 2$. The choice $\tau = p - 2$ originates with James and Stein (1961), but the distinction is small when p is large.

In the targeted shrinkage case we recommend

$$\tau = \mathrm{tr}(\mathbf{A}) \tag{8}$$

where

$$\mathbf{A} = \left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}\mathbf{R}. \tag{9}$$

A consistent estimate is $\hat{\tau} = \text{tr}(\widehat{\mathbf{A}})$ where $\widehat{\mathbf{A}} = \left(\widehat{\mathbf{R}}'\widehat{\mathbf{V}}\widehat{\mathbf{R}}\right)^{-1}\widehat{\mathbf{R}}'\widehat{\mathbf{V}}\mathbf{W}\widehat{\mathbf{V}}\widehat{\mathbf{R}}$ and $\widehat{\mathbf{R}} = \mathbf{R}(\widehat{\boldsymbol{\theta}}_n)$. Notice that in the canonical case that $\mathbf{A} = \mathbf{I}_p$ and $\tau = p$, and thus (9) includes the canonical case recommendation as a special case. These recommendations will be justified in Section 4.

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. The latter is also a special case of Hansen's (2007) Mallows Model Averaging (MMA) estimator with two models.

In general, the degree of shrinkage depends on the ratio τ/D_n. When $D_n < \tau$ then $\hat{w} = 0$ and $\widehat{\boldsymbol{\theta}}_n^* = \widehat{\boldsymbol{\theta}}_n$ equals the restricted estimator. When $D_n > \tau$ then $\widehat{\boldsymbol{\theta}}_n^*$ is a weighted average of the unrestricted and restricted estimators, with more weight on the unrestricted estimator when D_n/τ is large.

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

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h} \tag{10}$$

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 between the parameter and 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 (10).

Assumption 1

1. *X_i are iid.*
2. *The model $f(x, \boldsymbol{\theta})$ satisfies the conditions of Theorem 3.3 of Newey and McFadden (1994).*
3. *$\mathbf{R}(\boldsymbol{\theta})$ is continuous in a neighborhood of $\boldsymbol{\theta}_0$.*
4. *$\text{rank}(\mathbf{R}) = p$ where $\mathbf{R} = \mathbf{R}(\boldsymbol{\theta}_0)$.*

Let $\mathbf{V} = \mathbf{V}_{\boldsymbol{\theta}_0} = \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}$ be the asymptotic variance of the MLE under the sequences (10).

Theorem 1 *Under Assumption 1, along the sequences (10)*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathrm{Z} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{V}\right), \tag{11}$$

$$\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathrm{Z} - \mathbf{V}\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\left(\mathrm{Z} + \boldsymbol{h}\right), \tag{12}$$

$$D_n \xrightarrow[\boldsymbol{\theta}_n]{} \xi = \left(\mathrm{Z} + \boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z} + \boldsymbol{h}\right), \tag{13}$$

$$\hat{w} \xrightarrow[\boldsymbol{\theta}_n]{} w = \left(1 - \frac{\tau}{\xi}\right)_+, \tag{14}$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} w\mathrm{Z} + (1-w)\left(\mathrm{Z} - \mathbf{V}\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\left(\mathrm{Z} + \boldsymbol{h}\right)\right), \tag{15}$$

where

$$\mathbf{B} = \mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\left(\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}\mathbf{R}\right)\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'. \tag{16}$$

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 $\widehat{\boldsymbol{\theta}}_n^*$ is written as a random weighted average of the asymptotic distributions of $\widehat{\boldsymbol{\theta}}_n$ and $\widetilde{\boldsymbol{\theta}}_n$. Since the distribution of $\widehat{\boldsymbol{\theta}}_n^*$ depends on $\boldsymbol{h}$, the estimator $\widehat{\boldsymbol{\theta}}_n^*$ 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 (15) tends towards $\mathrm{Z} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{V}\right)$.

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 (13) 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

$$\mathbb{E}\xi = \boldsymbol{h}'\mathbf{B}\boldsymbol{h} + \mathbb{E}\,\mathrm{tr}\left(\mathbf{B}\mathrm{Z}\mathrm{Z}'\right) = \boldsymbol{h}'\mathbf{B}\boldsymbol{h} + \mathrm{tr}\left(\mathbf{A}\right) \tag{17}$$

where $\mathbf{B}$ is from (16) and $\mathbf{A}$ was defined in (9) The matrices $\mathbf{A}$ and $\mathbf{B}$ play important roles in our theory. Notice that in the full shrinkage case we have the simplifications $\mathbf{B} = \mathbf{W}$ and $\mathbf{A} = \mathbf{W}\mathbf{V}$. In the canonical case $\mathbf{W} = \mathbf{V}^{-1}$ we find that (17) simplifies to

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

Furthremore, 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 $\mathbb{E}\ell(T_n - \boldsymbol{\theta})$. In general this expectation is difficult to evaluate, and may not even be finite unless T_n has a finite second moments. To obtain a useful approximation and ensure existence we use a trimmed loss and take limits as the sample size $n \to \infty$.

Define

$$\ell_\zeta(\mathbf{u}) = \min\{\ell(\mathbf{u}), \zeta\}, \tag{19}$$

which is the quadratic function (4) trimmed at ζ. Let $T = \{T_n : n = 1, 2, ...\}$ denote a sequence of estimators, and let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_n : n = 1, 2, ...\}$ denote the sequence of parameter values (10). We define the asymptotic (trimmed) risk of the estimator sequence T for the parameter sequence $\boldsymbol{\theta}$ as

$$\rho(T, \boldsymbol{\theta}) = \lim_{\zeta\to\infty} \liminf_{n\to\infty} \mathbb{E}\, n\ell_{\zeta/n}(T_n - \boldsymbol{\theta}_n). \tag{20}$$

This is the expected loss, using the trimmed loss function, but in large samples ($n \to \infty$) and with arbitrarily negligible trimming ($\zeta \to \infty$).

The definition (20) is convenient when T_n has an asymptotic distribution. Specifically, suppose that

$$\sqrt{n}\,(T_n - \boldsymbol{\theta}_n) \xrightarrow[\boldsymbol{\theta}_n]{} \xi \tag{21}$$

for some distribution ξ. Observe that since $\ell(\mathbf{u})$ is quadratic,

$$n\ell_{\zeta/n}(T_n - \boldsymbol{\theta}_n) = \ell_\zeta(\sqrt{n}\,(T_n - \boldsymbol{\theta}_n)). \tag{22}$$

As $\ell_\zeta(\mathbf{u})$ is bounded, (21) and the portmanteau lemma imply that

$$\lim_{n\to\infty} \mathbb{E}\,\ell_\zeta(\sqrt{n}\,(T_n - \boldsymbol{\theta}_n)) = \mathbb{E}\,\ell_\zeta(\xi). \tag{23}$$

Equations (22) and (23) plus definition (4) combine to show that

$$\begin{aligned} \rho(T, \boldsymbol{\theta}) &= \lim_{\zeta\to\infty} \mathbb{E}\,\ell_\zeta(\xi) \\ &= \mathbb{E}\,\ell(\xi) \\ &= \frac{\mathbb{E}\,(\xi'\mathbf{W}\xi)}{\operatorname{tr}(\mathbf{W}\mathbf{V})}. \end{aligned} \tag{24}$$

Thus the asymptotic trimmed risk of any estimator T satisfying (21) can be calculated using (24).

Recalling the matrix $\mathbf{A} = (\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}\mathbf{V}\mathbf{W}\mathbf{V}\mathbf{R}$ from (9) we define the scalar

$$\lambda_p = \frac{\operatorname{tr}(\mathbf{A})}{\lambda_{\max}(\mathbf{A})} \tag{25}$$

which satisfies $\lambda_p \leq p$. 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 Assumption 1, $\lambda_p > 2$, and*

$$0 < \tau \leq 2\left(\operatorname{tr}(\mathbf{A}) - 2\lambda_{\max}(\mathbf{A})\right), \tag{26}$$

then

$$\rho(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\theta}) < \rho(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \tag{27}$$

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

$$\mathbf{H}(c) = \left\{\boldsymbol{h} : \boldsymbol{h}'\mathbf{B}\boldsymbol{h} \leq \operatorname{tr}(\mathbf{A})\, c\right\} \tag{28}$$

then

$$\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \rho(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\theta}) \leq 1 - \frac{\tau}{\operatorname{tr}(\mathbf{W}\mathbf{V})} \frac{2\left(\operatorname{tr}(\mathbf{A}) - 2\lambda_{\max}(\mathbf{A})\right) - \tau}{\operatorname{tr}(\mathbf{A})\,(c+1)}. \tag{29}$$

Equation (27) shows that the asymptotic risk of the shrinkage estimator is strictly less than that of the MLE for all parameter values, so long as the shrinkage parameter τ satisfies the condition (26). As (27) holds for even extremely large values of $\boldsymbol{h}$, this shows that in a very real sense the shrinkage estimator strictly dominates the usual estimator.

The assumption $\lambda_p > 2$ is the critical condition needed to ensure that the shrinkage estimator has globally smaller asymptotic risk than the usual estimator. In the canonical case $\mathbf{W} = \mathbf{V}^{-1}$, $\lambda_p > 2$ 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. $\lambda_p > 2$ generalizes $p > 2$ to allow for general weight matrices.

The condition (26) simplifies to $0 < \tau \leq 2(p-2)$ in the canonical case, which is a standard restriction on the shrinkage parameter. Notice that in the canonical case, $\tau = p$ satisfies (26) if $p \geq 4$, while $\tau = p - 2$ satisfies (26) for $p \geq 3$. In the general case, if we set $\tau = \operatorname{tr}(\mathbf{A})$ (as recommended in (8)) then (26) is satisfied if $\lambda_p \geq 4$, which generalizes the condition $p \geq 4$ from the canonical case. Equivalently, Theorem 2 shows that when $\tau = \operatorname{tr}(\mathbf{A})$, then $\lambda_p \geq 4$ is sufficient for the generalized shrinkage estimator to strictly dominate the MLE.

Equation (29) provides a uniform bound for the asymptotic risk in the ball $\boldsymbol{h} \in \mathbf{H}(c)$. If $\lambda_p \geq 4$ and we set $\tau = \operatorname{tr}(\mathbf{A})$ as recommended, then (29) becomes

$$\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \rho(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\theta}) \leq 1 - \frac{\operatorname{tr}(\mathbf{A})}{\operatorname{tr}(\mathbf{W}\mathbf{V})} \left(\frac{1 - 4/\lambda_p}{c+1}\right). \tag{30}$$

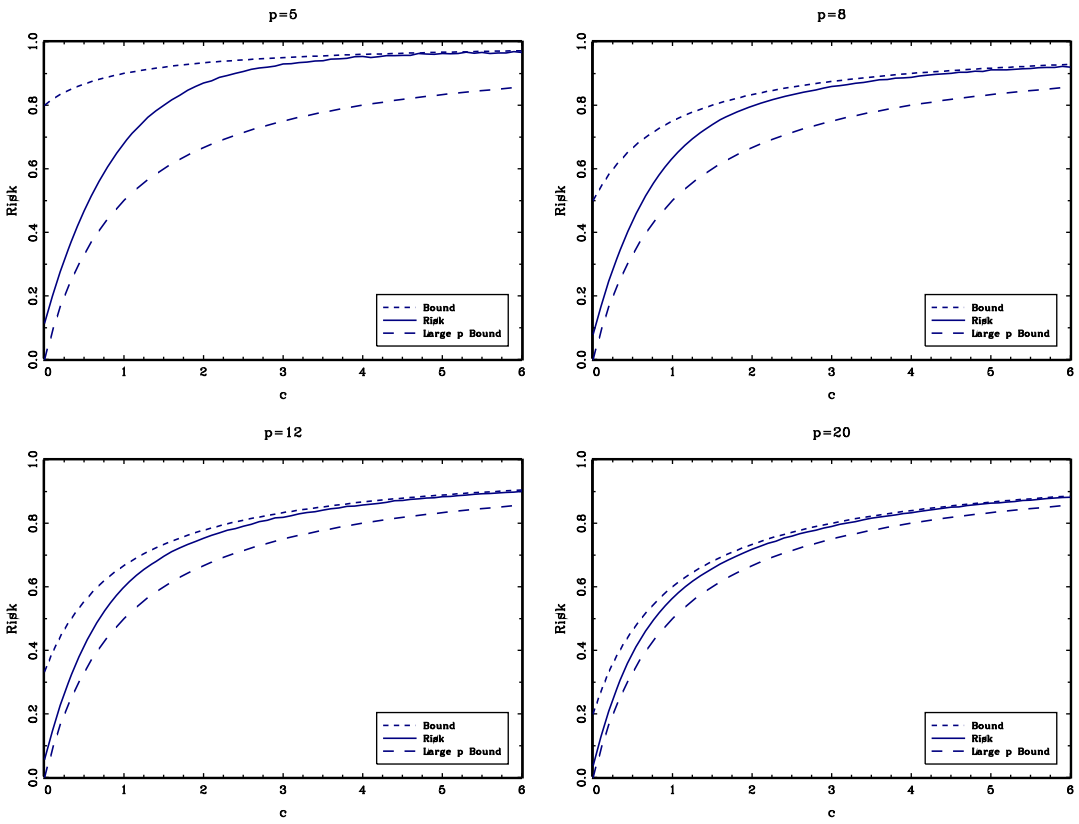

Figure 1: Asymptotic Risk of Shrinkage Estimators

In the canonical case this specializes to

$$\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \rho(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\theta}) \leq 1 - \frac{m}{p}\left(\frac{1-4/p}{c+1}\right). \tag{31}$$

To illustrate these results numerically, we plot in Figure 1 the asymptotic risk of the shrinkage estimator $\widehat{\boldsymbol{\theta}}_n^*$ in the full shrinkage ($m = p$) canonical case. The asymptotic risk is only a function of p and c, and we plot the risk as a function of c for $p = 5$, 8, 12, and 20. The asymptotic risk is plotted with the solid line. We also plot the upper bound (31) 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. (Figure 1 also plots a "Large p bound" which will be discussed in the following section.)

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 (31) is not particularly tight for small p, but improves as p increases. This means that risk improvements implied by Theorem 2 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 (30) improves as the shrinkage dimension p increases. Indeed the bound (30) leads to a simple approximation for the asymptotic risk when the shrinkage dimension p is large.

Theorem 3 *Under Assumption 1, if as* $p \to \infty$, $\lambda_p \to \infty$, $\tau / \operatorname{tr}(\mathbf{A}) \to 1$, *and*

$$\frac{\operatorname{tr}(\mathbf{A})}{\operatorname{tr}(\mathbf{WV})} \longrightarrow a, \tag{32}$$

then

$$\limsup_{p\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \rho(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\theta}) \leq 1 - \frac{a}{c+1}. \tag{33}$$

Equation (33) is a simplified version of (30). 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 asymptotic 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 (33) varies between $1 - a$ and 1, depending on c. Thus 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.

Equation (33) 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, (33) simplifies to

$$\limsup_{p\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \rho(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\theta}) \leq \frac{c}{c+1}. \tag{34}$$

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 (7) 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 (33) 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 (34) in the plots of Figure 1 (the long dashes). We can see that the large-p bound (34) lies beneath the finite-p bound (30) (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 (34) 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.

A limitation of Theorem 3 is that the sequential limits (first taking the sample size n to infinity and then taking the dimension p to infinity) is artificial. A deeper result would employ joint limits (taking n and p to infinity jointly). While desirable, this extension does not appear to be feasible given the present theoretical tools, and the sequential limit appears to be the best which can be attained. Because of the use of sequential limits, Theorem 3 should not be interpreted as nonparametric. Rather, it shows that in finite yet high-dimensional parametric models, the risk of the shrinkage estimator takes the simple form (33).

It is quite likely that nonparametric versions of Theorem 3 could be developed. For example, a nonparametric series regression estimator could be shrunk towards a simpler model, and we would expect improvements in asymptotic risk similar to (33). There are also conceptual similarities between shrinkage and penalization, for example Shen (1997). These connections are worth exploring.

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

$$\sup_{I \subset \mathbb{R}^m} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}\, n\ell\left(T_n - \boldsymbol{\theta}_n\right) \tag{35}$$

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 the asymptotic uniform risk (35) is bounded below by 1. This demonstrates that no estimator has smaller asymptotic uniform risk than the MLE over unbounded $\boldsymbol{h}$.

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

$$\sup_{I \subset \mathbf{H}(c)} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}\, n\ell\left(T_n - \boldsymbol{\theta}_n\right) \tag{36}$$

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 (36) is equivalent to

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

The standard method to establish the efficiency bound (35) 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 (36) 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 is Pinsker's Theorem (Pinsker, 1980) 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* $\mathrm{Z} \sim \mathrm{N}_m\left(\boldsymbol{h}, \mathbf{V}\right)$ *and* $\lambda_p > 8$, *where* λ_p *is defined in (25). For any estimator* $T = T(\mathrm{Z})$,

$$\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \mathbb{E}\, \ell(T - \boldsymbol{h}) \geq 1 - \left[\frac{1}{1+c} + \left(\frac{2}{1+c} + 4c\right)\lambda_p^{-1/3}\right]\frac{\operatorname{tr}\left(\mathbf{A}\right)}{\operatorname{tr}\left(\mathbf{WV}\right)}. \tag{37}$$

This is a finite sample lower bound on the quadratic risk for the normal sampling model. Typically in this literature this bound is expressed for the high-dimensional (large λ_p) case. Indeed, taking the limit as $p \to \infty$ as in Theorem 3, then the bound (37) simplifies to

$$\liminf_{p\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \mathbb{E}\, \ell(T - \boldsymbol{h}) \geq 1 - \frac{a}{c+1}. \tag{38}$$

We do not use (38) directly, but rather use (37) as an intermediate step towards establishing a

large n bound. It is worth noting that while Theorem 4 appears similar to existing results (e.g. Theorem 7.28 of Wasserman (2006)), its proof is a significant extension due to the need to break the parameter space into parts constrained by $\mathbf{H}(c)$ and those which are unconstrained.

Combined with the limits of experiments technique, Theorem 4 allows us to establish an asymptotic (large n) local minimax 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 $f(x, \boldsymbol{\theta})$ indexed by a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$, and the density is differentiable in quadratic mean, that is*

$$\int \left[f(x, \boldsymbol{\theta} + \boldsymbol{h})^{1/2} - f(x, \boldsymbol{\theta})^{1/2} - \frac{1}{2} \boldsymbol{h}' \boldsymbol{g}(x, \boldsymbol{\theta}) f(x, \boldsymbol{\theta})^{1/2} \right]^2 d\mu = o\left(\|\boldsymbol{h}\|^2 \right), \qquad \boldsymbol{h} \to 0 \tag{39}$$

where $\boldsymbol{g}(x, \boldsymbol{\theta}) = \dfrac{\partial}{\partial \boldsymbol{\theta}} \log f(x, \boldsymbol{\theta})$. Suppose that $\mathbf{I}_{\boldsymbol{\theta}} = \mathbb{E} \boldsymbol{g}(X_i, \boldsymbol{\theta}) \boldsymbol{g}(X_i, \boldsymbol{\theta})' > 0$ and set $\mathbf{V} = \mathbf{I}_{\boldsymbol{\theta}}^{-1}$. Finally, suppose that $\lambda_p > 8$, where λ_p is defined in (25). Then for any sequence of estimators T_n, on the sequence $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-1/2} \boldsymbol{h}$

$$\sup_{I \subset \mathbf{H}(c)} \liminf_{n \to \infty} \sup_{\boldsymbol{h} \in I} \mathbb{E}\, n \ell \left(T_n - \boldsymbol{\theta}_n \right) \geq 1 - \left[\frac{1}{1+c} + \left(\frac{2}{1+c} + 4c \right) \lambda_p^{-1/3} \right] \frac{\operatorname{tr}(\mathbf{A})}{\operatorname{tr}(\mathbf{WV})}. \tag{40}$$

Furthermore, suppose that as $p \to \infty$, $\lambda_p \to \infty$ and (32) holds. Then

$$\liminf_{p \to \infty} \sup_{I \subset \mathbf{H}(c)} \liminf_{n \to \infty} \sup_{\boldsymbol{h} \in I} \mathbb{E}\, n \ell \left(T_n - \boldsymbol{\theta}_n \right) \geq 1 - \frac{a}{c+1}. \tag{41}$$

Theorem 5 provides a lower bound on the asymptotic local minimax risk for $\boldsymbol{h}$ in the ball $\mathbf{H}(c)$. (40) is the case of finite p, and (41) shows that the bound takes a simple form when p is large. Since this lower bound is equal to the upper bound (33) attained by out shrinkage estimator, (41) 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.

Differentiability in quadratic mean (39) is weaker than the requirements for asymptotic normality of the MLE.

Note that the equality of (33) and (41) holds for all values of c. This is a very strong efficiency property.

In the case of full shrinkage ($p = m$) then (41) is equivalent to

$$\liminf_{p \to \infty} \liminf_{n \to \infty} \sup_{\boldsymbol{h} \in \mathbf{H}(c)} \mathbb{E}\, n \ell \left(T_n - \boldsymbol{\theta}_n \right) \geq \frac{c}{c+1}.$$

In the case of partial shrinkage the more complicated double supremum in (41) is needed since the set $\mathbf{H}(c)$ is unbounded.

Classic minimax theory (e.g. Theorem 8.11 of van der Vaart (1998)) applies to all bowl-shaped loss functions $\ell(\mathbf{u})$, not just quadratic loss, and thus it seems reasonable to conjecture that Theorem

1 will extend beyond quadratic loss. The challenge is that Pinsker's theorem specifically exploits the structure of the quadratic loss, and thus it is unclear how to extend Theorem 4 to allow for other loss functions. Allowing for more general loss functions would be a useful extension.

Similarly to Theorem 3, a limitation of the bound (41) is the use of the sequential limits, first taking n to infinity and then p to infinity. A deeper result would employ joint limits.

7 Simulation

We illustrate the numerical magnitude of the finite sample shrinkage improvements in a simple numerical simulation. The model is a binary probit. For $i = 1, ..., n$,

$$\begin{aligned} y_i &= 1\left(y_i^* \geq 0\right) \\ y_i^* &= X_{1i}'\boldsymbol{\beta}_1 + X_{2i}'\boldsymbol{\beta}_2 + e_i \\ e_i &\sim N(0,1). \end{aligned}$$

The regressors X_{i1} and X_{2i} are $k \times 1$ and $p \times 1$, respectively, with $p > k$. The first element of X_{1i} is an intercept, the remaining regressors are $N(0,1)$ with correlation ρ.

The regression coefficients are set as $\beta_0 = 0$, $\boldsymbol{\beta}_1 = (b, b, ..., b)'$ and $\boldsymbol{\beta}_2 = (c, c, ..., c)'$. Consequently, the control parameters of the model are c, n, p, k, b, and ρ. We found that the results were qualitatively insensitive to the choice of k , b, and ρ, so we fixed their values at $k = 4$, $b = 0$, and $\rho = 0$, and report results for different values of c, n, and p. We also experiment with the alternative specification $\boldsymbol{\beta}_2 = (c, 0, ..., 0)'$ (only one omitted regressor important) and the results were virturally identical so are not reported.

We are interested in comparing the finite sample risk of estimators of $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. The estimators will be functions of the following primary components:

1. $\widehat{\boldsymbol{\beta}}$ = unconstrained MLE. Probit of y_i on $(1, X_{1i}, X_{2i})$
2. $\widetilde{\boldsymbol{\beta}}$ = constrained MLE. Probit of y_i on $(1, X_{1i})$
3. $LR = 2\left(\log \mathcal{L}(\widehat{\boldsymbol{\beta}}) - \log \mathcal{L}(\widetilde{\boldsymbol{\beta}})\right)$, the likelihood ratio test for the restriction $\boldsymbol{\beta}_2 = 0$
4. $\widehat{\mathbf{V}}$ = estimate of the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$

We compare three estimators. The first is $\widehat{\boldsymbol{\beta}}$, the unconstrained MLE. The second is our canonical (partial) shrinkage estimator

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^* &= \hat{w}\widehat{\boldsymbol{\beta}} + (1 - \hat{w})\widetilde{\boldsymbol{\beta}} \\ \hat{w} &= \left(1 - \frac{p}{D_n}\right)_+ \\ D_n &= n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right)' \widehat{\mathbf{V}}^{-1}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right). \end{aligned}$$

The third is the pretest estimator

$$\overline{\boldsymbol{\beta}} = \begin{cases} \widehat{\boldsymbol{\beta}} & \text{if} \quad LR \geq q \\ \widetilde{\boldsymbol{\beta}} & \text{if} \quad LR < q \end{cases}$$

where q is the 95% quantile of the χ^2_p distribution. We include $\overline{\boldsymbol{\beta}}$ to provide a comparison with a conventional selection technique used routinely in applications.

The simulations were computed in R, and the MLE was calculated using the built-in glm program. One difficulty was that in some cases (when then sample size n was small and the number of parameters $k+p$ was large) the glm algorithm failed to converge for the unconstrained MLE and thus the reported estimate $\widehat{\boldsymbol{\beta}}$ was unreliable. For these cases we simply set all estimates equal to the restricted estimate $\widetilde{\boldsymbol{\beta}}$. This corresponds to empirical practice and does not bias our results as all estimators were treated symmetrically.

We compare the estimators by unweighted MSE. For any estimator $\widehat{\boldsymbol{\beta}}$,

$$MSE(\widehat{\boldsymbol{\beta}}) = \mathbb{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

This MSE is unweighted (e.g., is calculated using $\mathbf{W} = \mathbf{I}$) even though the generalized shrinkage estimator is optimized for $\mathbf{W} = \mathbf{V}^{-1}$, where $\mathbf{V}$ is the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$. We do this for simplicity, and not to avoid skewing our results in favor of the shrinkage estimator.

We normalize the MSE of all estimators by that of the unconstrained MLE.

We calculated the MSE by simulation using 30,000 replications. We display results for $n = \{200, 500\}$ and $p = \{4, 8\}$, and vary c on a 50-point grid. The results are displayed in Figure 2. The trimmed MSE are displayed as lines. The solid line is the relative trimmed MSE of the generalized shrinkage estimator, the dashed line is the relative trimmed MSE of the pretest estimator, and the dotted line is 1, the relative trimmed MSE of the unconstrained MLE.

Figure 2 shows convincingly that the generalized shrinkage estimator significantly dominates the other estimators. Its finite-sample MSE is less than that of the unconstrained estimator for all parameter values, and in some cases its MSE is a small fraction. It is also constructive to compare the shrinkage estimator with the pretest estimator. For nearly all parameter values the shrinkage estimator has lower trimmed MSE. The only exceptions are for very small values of c. Furthermore, the MSE of the pretest estimator is quite sensitive to the value of c, and for many values the pretest estimator has MSE much higher than the unconstrained estimator. This is a well-documented property of pretest estimators, but is worth emphasizing as pretests are still routinely used for selection in applied research. The numerical calculations shown in Figure 2 show that a much better estimator is the generalized shrinkage estimator.

Some readers may be surprised by the extremely strong performance of the shrinkage estimator relative to the MLE. However, this is precisely the lesson of Theorems 2 and 3. Shrinkage strictly improves asymptotic risk, and the improvements can be especially strong in high-dimensional cases.

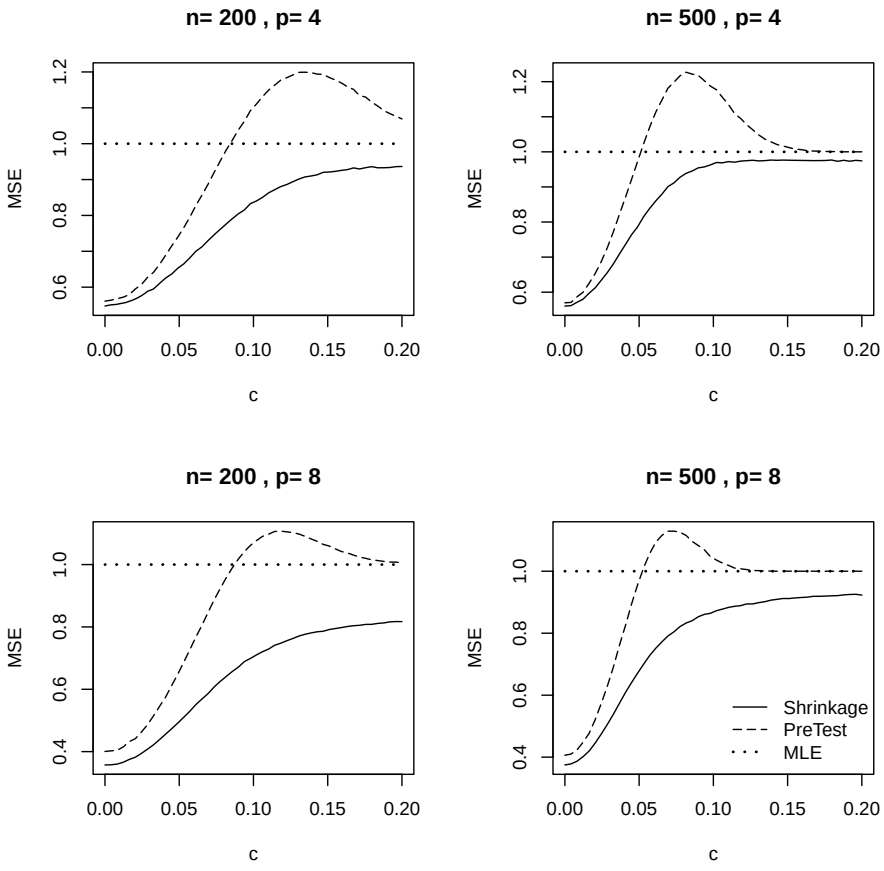

Figure 2: Mean Squared Error of Shrinkage, Pre-Test and Maximum Likelihood Estimators

8 Appendix

Proof of Theorem 1: (11) is Theorem 3.3 of Newey and McFadden (1994). (12) follows by standard arguments, see for example, the derivation in Section 9.1 of Newey and McFadden (1994). (13), (14), and (15) 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: Observe that $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$ under (11). Then (24) shows that

$$\rho(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{\mathbb{E}\left(\mathrm{Z}'\mathbf{W}\mathrm{Z}\right)}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)} = \frac{\mathrm{tr}\left(\mathbf{W}\mathbb{E}\left(\mathrm{Z}\mathrm{Z}'\right)\right)}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)} = \frac{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)} = 1. \tag{42}$$

Next, $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \xi$, where ξ is the random variable shown in (15). The variable ξ has a classic James-Stein distribution with positive-part trimming. Define the analogous random variable without positive part trimming

$$\xi^* = \mathrm{Z} - \left(\frac{\tau}{\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)}\right)\mathbf{V}\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\left(\mathrm{Z}+\boldsymbol{h}\right). \tag{43}$$

Then using (24) and the fact that the pointwise quadratic risk of ξ is strictly smaller than that of ξ^* (as shown, for example, by Theorem 5.5.4 of Lehman and Casella (1998)),

$$\rho(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\theta}) = \frac{\mathbb{E}\left(\xi'\mathbf{W}\xi\right)}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)} < \frac{\mathbb{E}\left(\xi^{*\prime}\mathbf{W}\xi^*\right)}{\mathrm{tr}\left(\mathbf{W}\mathbf{V}\right)}. \tag{44}$$

Using (43), we calculate that (44) equals

$$\begin{aligned}
&\frac{\mathbb{E}(\mathrm{Z}'\mathbf{W}\mathrm{Z})}{\operatorname{tr}(\mathbf{WV})} \\
&+\frac{\tau^2}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}\left(\frac{(\mathrm{Z}+\boldsymbol{h})'\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}\mathbf{W}\mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'(\mathrm{Z}+\boldsymbol{h})}{\left((\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})\right)^2}\right) \\
&-2\frac{\tau}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}\left(\frac{(\mathrm{Z}+\boldsymbol{h})'\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}\mathbf{W}\mathrm{Z}}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right) \\
&=1+\frac{\tau^2}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}\left(\frac{1}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right) \\
&-2\frac{\tau}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}\left(\eta(\mathrm{Z}+\boldsymbol{h})'\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{W}\mathrm{Z}\right) \qquad (45)
\end{aligned}$$

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)

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

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

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

Thus (45) is smaller than

$$\begin{aligned}
&1+\frac{\tau^2}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}\left(\frac{1}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right)-2\frac{\tau}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}\left(\frac{\operatorname{tr}(\mathbf{A})-2\lambda_{\max}(\mathbf{A})}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right)\\
&=1-\frac{\tau}{\operatorname{tr}(\mathbf{WV})}\mathbb{E}\left(\frac{2(\operatorname{tr}(\mathbf{A})-2\lambda_{\max}(\mathbf{A}))-\tau}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right)\\
&\leq 1-\frac{\tau}{\operatorname{tr}(\mathbf{WV})}\frac{2(\operatorname{tr}(\mathbf{A})-2\lambda_{\max}(\mathbf{A}))-\tau}{\mathbb{E}\left((\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})\right)} \qquad (46)
\end{aligned}$$

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

We calculate that

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

where the inequality is for $\boldsymbol{h}\in\mathbf{H}(c)$. Substituting into (46) we have established (29). As this bound is strictly less than 1, combined with (42) we have established (27). ■

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 $\mathbf{R}$. 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 $\boldsymbol{h}_2\sim\mathrm{N}(\mathbf{0},\mathbf{I}_pc\eta)$. 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

$$\widetilde{T}_2^{*\prime}\mathbf{W}_{22}\widetilde{T}_2^*\leq\operatorname{tr}(\mathbf{W}_{22})c. \qquad (47)$$

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 $T = T(\mathrm{Z})$, 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}\ell(T-\boldsymbol{h}) &\geq \int\int \mathbb{E}\ell(T-\boldsymbol{h}) d\Pi_1(\boldsymbol{h}_1) d\Pi_2^*(\boldsymbol{h}_2) && (48)\\
&\geq \int\int \mathbb{E}\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}\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)\\
&+ \frac{2}{\operatorname{tr}(\mathbf{W})} \int\int \mathbb{E}\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)\\
&+ \frac{1}{\operatorname{tr}(\mathbf{W})} \int\int \mathbb{E}\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) && (49)\\
&= \frac{1}{\operatorname{tr}(\mathbf{W})} \int \mathbb{E}\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) && (50)\\
&+ \frac{2}{\operatorname{tr}(\mathbf{W})} \left(\int \mathbb{E}\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) && (51)\\
&+ \frac{1}{\operatorname{tr}(\mathbf{W})} \frac{\int \mathbb{E}\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)} && (52)\\
&- \frac{1}{\operatorname{tr}(\mathbf{W})} \frac{\int_{\mathbf{H}_2(c)^c} \mathbb{E}\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)} && (53)
\end{aligned}$$

where the second inequality is because the Bayes estimator $\widetilde{T}$ minimizes the right-hand-side of (48). 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 (49) 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 (50)-(53) 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

$$\begin{aligned}
&\frac{1}{\operatorname{tr}(\mathbf{W})} \int \mathbb{E}\left[\left(\widetilde{T}_1-\boldsymbol{h}_1\right)' \mathbf{W}_{11} \left(\widetilde{T}_1-\boldsymbol{h}_1\right)\right] d\Pi_1(\boldsymbol{h}_1)\\
&= \frac{1}{\operatorname{tr}(\mathbf{W})} \int \mathbb{E}\left[\left(\frac{\omega}{1+\omega}\mathrm{Z}_1-\boldsymbol{h}_1\right)' \mathbf{W}_{11} \left(\frac{\omega}{1+\omega}\mathrm{Z}_1-\boldsymbol{h}_1\right)\right] d\Pi_1(\boldsymbol{h}_1)\\
&= \frac{1}{\operatorname{tr}(\mathbf{W})} \int \left[\frac{1}{(1+\omega)^2}\boldsymbol{h}_1'\mathbf{W}_{11}\boldsymbol{h}_1 + \frac{\omega^2}{(1+\omega)^2}\operatorname{tr}(\mathbf{W}_{11})\right] d\Pi_1(\boldsymbol{h}_1)\\
&= \frac{\operatorname{tr}(\mathbf{W}_{11})}{\operatorname{tr}(\mathbf{W})}\frac{\omega}{1+\omega}. && (54)
\end{aligned}$$

Second, since

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

it follows that (51) equals zero.

Third, take (52). Because $\widetilde{T}_2$ is the Bayes estimator under the prior Π_2,

$$\begin{aligned}
&\frac{1}{\operatorname{tr}(\mathbf{W})}\frac{\int \mathbb{E}\left[\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)' \mathbf{W}_{22}\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} \\
&\geq \frac{1}{\operatorname{tr}(\mathbf{W})}\frac{\int \mathbb{E}\left[\left(\widetilde{T}_2 - \boldsymbol{h}_2\right)' \mathbf{W}_{22}\left(\widetilde{T}_2 - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} \\
&\geq \frac{1}{\operatorname{tr}(\mathbf{W})}\int \mathbb{E}\left[\left(\widetilde{T}_2 - \boldsymbol{h}_2\right)' \mathbf{W}_{22}\left(\widetilde{T}_2 - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2) \\
&= \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})}\frac{c\eta}{1+c\eta} \qquad (55) \\
&= \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})}\left(\frac{c}{1+c} - \frac{2\lambda_p^{-1/3}}{1+c}\right) \qquad (56)
\end{aligned}$$

where (55) is a calculation similar to (54) using $\widetilde{T}_2 = \frac{c\eta}{1+c\eta}\mathrm{Z}_2$ and $\boldsymbol{h}_2 \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_p c\eta)$. (56) makes a simple expansion using $\eta = 1 - 2\lambda_p^{-1/3}$.

Fourth, take (53). Our goal is to show that this term is negligible for large p, and our argument is based on the proof of Theorem 7.28 from 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 (47) to find that

$$\begin{aligned}
\mathbb{E}\left[\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)' \mathbf{W}_{22}\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)\right] &\leq 2\mathbb{E}\left(\widetilde{T}_2^{*\prime} \mathbf{W}_{22} \widetilde{T}_2^*\right) + 2\boldsymbol{h}_2' \mathbf{W}_{22} \boldsymbol{h}_2 \\
&\leq 2\operatorname{tr}(\mathbf{W}_{22})\, c + 2\boldsymbol{h}_2' \mathbf{W}_{22} \boldsymbol{h}_2 \\
&= 2\operatorname{tr}(\mathbf{W}_{22})\, c\,(1+q) \\
&\leq 2\operatorname{tr}(\mathbf{W}_{22})\, c\,(2+q-\eta). \qquad (57)
\end{aligned}$$

Note that $\boldsymbol{h}_2 \in \mathbf{H}_2(c)^c$ is equivalent to $q > 1$. Using (57) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\int_{\mathbf{H}_2(c)^c} \mathbb{E}\left[\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)' \mathbf{W}_{22}\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2) \\
&\leq 2\operatorname{tr}(\mathbf{W}_{22})\, c\left[2\int_{\mathbf{H}_2(c)^c} d\Pi_2(\boldsymbol{h}_2) + \int_{\mathbf{H}_2(c)^c} (q-\eta)\, d\Pi_2(\boldsymbol{h}_2)\right] \\
&\leq 2\operatorname{tr}(\mathbf{W}_{22})\, c\left[2\mathbb{P}(q>1) + \operatorname{var}(q)^{1/2}\mathbb{P}(q>1)^{1/2}\right]. \qquad (58)
\end{aligned}$$

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

$$\operatorname{var}(q) = \frac{\eta^2}{\left(\sum_{j=1}^{p} w_j\right)^2} \sum_{j=1}^{p} w_j^2 \operatorname{var}\left(y_j^2\right) \leq 2\lambda_p^{-1} \tag{59}$$

since $\lambda_p = \dfrac{\sum_{j=1}^{p} w_j}{\max_j w_j} = \operatorname{tr}(\mathbf{W}_{22}) / \lambda_{\max}(\mathbf{W}_{22})$. By Markov's inequality, (59), and $1 - \eta = 2\lambda_p^{-1/3}$,

$$\mathbb{P}(q > 1) = \mathbb{P}(q - \eta > 1 - \eta) \leq \frac{\operatorname{var}(q)}{(1-\eta)^2} \leq \frac{\lambda_p^{-1/3}}{2}. \tag{60}$$

Furthermore, (60) and $\lambda_p^{-1/3} \leq 2^{-1}$ imply that

$$\begin{aligned} \int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2) &= 1 - \mathbb{P}(q > 1) \\ &\geq 1 - \frac{\lambda_p^{-1/3}}{2} \\ &\geq \frac{3}{4}. \end{aligned} \tag{61}$$

It follows from (58), (59), (60), (61) and $\lambda_p^{-1/3} \leq 2^{-1}$ that

$$\begin{aligned} &\frac{1}{\operatorname{tr}(\mathbf{W})} \frac{\int_{\mathbf{H}_2(c)^c} \mathbb{E}\left[\left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)' \mathbf{W}_{22} \left(\widetilde{T}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} \\ &\leq \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})} \frac{2c\left(\lambda_p^{-1/3} + \lambda_p^{-2/3}\right)}{3/4} \\ &\leq \frac{\operatorname{tr}(\mathbf{W}_{22})}{\operatorname{tr}(\mathbf{W})} 4c\lambda_p^{-1/3} \end{aligned} \tag{62}$$

Together, (54) and (62) applied to (50)-(52) show that

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

Since ω is arbitrary we conclude that

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

which is (37) 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 is based on the arguments in Theorem 8.11 of van der Vaart (1998), with two important differences. First, van der Vaart (1998) appeals to a compactification argument from Theorem 3.11.5 of Van der Vaart and Wellner (1996), while we use a different argument which allows for possibly singular $\mathbf{W}$. Second, we bound the risk of the limiting experiment using Theorem 4 rather than van der Vaart's Proposition 8.6.

Let $\mathbf{Q}(c)$ denote the rational vectors in $\mathbf{H}(c)$ placed in arbitrary order, and let Q_k denote the first k vectors in this sequence. Define $Z_n=\sqrt{n}\left(T_n-\boldsymbol{\theta}_n\right)$. There exists a subsequence $\{n_k\}$ of $\{n\}$ such that

$$\begin{aligned}\sup_{I\subset\mathbf{H}(c)}\liminf_{n\to\infty}\sup_{\boldsymbol{h}\in I}\mathbb{E}\,\ell\left(Z_n\right) &\geq \lim_{k\to\infty}\liminf_{n\to\infty}\sup_{\boldsymbol{h}\in Q_k}\mathbb{E}\,\ell\left(Z_n\right)\\ &= \lim_{k\to\infty}\sup_{\boldsymbol{h}\in Q_k}\mathbb{E}\,\ell\left(Z_{n_k}\right)\\ &\geq \lim_{k\to\infty}\sup_{\boldsymbol{h}\in Q_K}\mathbb{E}\,\ell\left(Z_{n_k}\right) \qquad (63)\end{aligned}$$

the final inequality for any $K<\infty$.

Since we allow $\mathbf{W}$ to have rank $r\leq m$, write $\mathbf{W}=\mathbf{G}_1\mathbf{G}_1'$ where $\mathbf{G}_1$ is $m\times r$ with rank r. Set $\mathbf{G}=[\mathbf{G}_1,\mathbf{G}_2]$ where $\mathbf{G}_2$ is $m\times(m-r)$ with rank $m-r$ and $\mathbf{G}_1'\mathbf{G}_2=0$. Define

$$Z_n^*=\mathbf{G}^{-1\prime}\begin{pmatrix}\mathbf{G}_1'Z_n\\ 0_{m-r}\end{pmatrix}$$

which replaces the linear combinations $\mathbf{G}_2'Z_n$ with zeros. Notice that since the loss function is a quadratic in $\mathbf{W}=\mathbf{G}_1\mathbf{G}_1'$, then $\ell\left(Z_n\right)=\ell\left(Z_n^*\right)$.

We next show that without loss of generality we can assume that Z_n^* is uniformly tight on a subsequence $\{n_k'\}$ of $\{n_k\}$. Suppose not. Then there exists some $\varepsilon>0$ such that for any $\zeta<\infty$,

$$\liminf_{k\to\infty}P\left(Z_{n_k}^{*\prime}\mathbf{G}\mathbf{G}'Z_{n_k}^*>\zeta\right)\geq\varepsilon. \qquad (64)$$

Set $\zeta = \operatorname{tr}(\mathbf{WV})/\varepsilon$. Since $\ell(Z_n^*) = Z_n^{*\prime}\mathbf{G}\mathbf{G}'Z_n^*/\operatorname{tr}(\mathbf{WV})$, (64) implies

$$\liminf_{k\to\infty} \mathbb{E}\,\ell\left(Z_{n_k}^*\right) = \liminf_{k\to\infty} \frac{\mathbb{E}\, Z_{n_k}^{*\prime}\mathbf{G}\mathbf{G}'Z_{n_k}^*}{\operatorname{tr}(\mathbf{WV})} \geq \frac{\zeta\varepsilon}{\operatorname{tr}(\mathbf{WV})} = 1$$

which is larger than (40). Thus for the remainder we assume that Z_n^* is uniformly tight on a subsequence $\{n_k'\}$ of $\{n_k\}$.

Tightness implies by Prohorov's theorem that there is a further subsequence $\{n_k''\}$ along which $Z_{n_k''}^*$ converges in distribution. For simplicity write $\{n_k''\} = \{n_k\}$. Theorem 8.3 of van der Vaart (1988) shows that differentiability in quadratic mean and $\mathbf{I}_{\boldsymbol{\theta}} > 0$ imply that the asymptotic distribution of $Z_{n_k}^*$ is $T(\mathrm{Z}) - \boldsymbol{h}$, where $T(\mathrm{Z})$ is a (possibly randomized) estimator of $\boldsymbol{h}$ based on $\mathrm{Z} \sim \mathrm{N}_m(\boldsymbol{h}, \mathbf{V})$. By the portmanteau lemma

$$\liminf_{k\to\infty} \mathbb{E}\,\ell\left(Z_{n_k}^*\right) \geq \mathbb{E}\,\ell(T(\mathrm{Z}) - \boldsymbol{h}).$$

Combined with (63), the fact that the set Q_K is finite, and $\ell(Z_n) = \ell(Z_n^*)$, we find that

$$\sup_{I\subset\mathbf{H}(c)} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}\,\ell(Z_n) \geq \sup_{\boldsymbol{h}\in Q_K} \mathbb{E}\,\ell(T(\mathrm{Z}) - \boldsymbol{h}).$$

Since K is arbitrary, and since $\ell(u)$ is continuous in $\boldsymbol{h}$, we deduce that

$$\begin{aligned}
\sup_{I\subset\mathbf{H}(c)} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}\,\ell(Z_n) &\geq \sup_{\boldsymbol{h}\in\mathbf{Q}(c)} \mathbb{E}\,\ell(T(\mathrm{Z}) - \boldsymbol{h}) \\
&= \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \mathbb{E}\,\ell(T(\mathrm{Z}) - \boldsymbol{h}) \\
&\geq 1 - \left[\frac{1}{1+c} + \left(\frac{2}{1+c} + 4c\right)\lambda_p^{-1/3}\right] \frac{\operatorname{tr}(\mathbf{A})}{\operatorname{tr}(\mathbf{WV})}
\end{aligned}$$

the final inequality by Theorem 4. We have shown (40).

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

$$\liminf_{p\to\infty} \sup_{I\subset\mathbf{H}(c)} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}\,\ell(Z_n) \geq 1 - \frac{a}{1+c}$$

which is (41). ■

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] Bhattacharya, P. K. (1966): "Estimating the mean of a multivariate normal population with general quadratic loss function," *The Annals of Mathematical Statistics*, 37, 1819-1824.

[3] Beran, Rudolf (2010): "The unbearable transparency of Stein estimation," *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jureckova*, 7, 25-34.

[4] Berger, James O. (1976a): "Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss," *The Annals of Statistics*, 4, 223-226.

[5] Berger, James O. (1976b): "Minimax estimation of a multivariate normal mean under arbitrary quadratic loss," *Journal of Multivariate Analysis*, 6, 256-264.

[6] Berger, James O. (1982): "Selecting a minimax estimator of a multivariate normal mean," *The Annals of Statistics*, 10, 81-92.

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

[8] Efromovich, Sam (1996): "On nonparametric regression for iid observations in a general setting," *Annals of Statistics*, 24, 1126-1144.

[9] Golubev, Grigory K. (1991): "LAN in problems of nonparametric estimation of functions and lower bounds for quadratic risks," *Theory of Probability and its Applications*, 36, 152-157.

[10] Golubev, Grigory K. and Michael Nussbaum (1990): "A risk bound in Sobolev class regression," *Annals of Statistics*, 18, 758–778.

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

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

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

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

[15] 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.

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

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

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

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

[20] Newey, Whitney K. and Daniel L. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics, Vol IV*, R.F. Engle and D.L McFadden, eds., 2113-2245. New York: Elsevier.

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

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

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

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

[25] Pinsker, M. S. (1980): "Optimal filtration of square-integrable signals in Gaussian white noise," *Problems of Information Transmission*, 16, 120-133.

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

[27] Sclove, Stanley L. (1968): "Improved estimators for coefficients in linear regression," *Journal of the American Statistical Association*, 63, 596-606.

[28] Shen, Xiaotong (1997): "On methods of sieves and penalization," *Annals of Statistics*, 25, 2555-2591.

[29] 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.

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

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

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

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