Efficient Shrinkage in Parametric Models

Bruce E. Hansen*
University of Wisconsin†

September 2012
Revised: June 2015

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 our 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 maximum likelihood estimator (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.

We also provide a new high-dimensional large sample local minimax efficiency bound. The bound is the lowest possible asymptotic risk, uniformly in a local region of the parameter space. Local minimax bounds are a stronger efficiency characterization than global minimax bounds. We show that our shrinkage estimator asymptotically achieves this local asymptotic minimax bound, while the MLE does not. Thus the shrinkage estimator, unlike the MLE, is locally minimax efficient.

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. I thank the Co-Editor, Associate Editor, and two referees for their insightful comments.

†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706.
behansen@wisc.edu

1 Introduction

In a conventional parametric setting, one where maximum likelihood estimation (MLE) applies, there is a considerable body of theory concerning the asymptotic efficiency properties of the MLE. First, the Cramér-Rao theorem establishes that no unbiased estimator can have a smaller variance than the inverse of the Fisher information, and the latter is the asymptotic variance of the MLE. Second, by the Hájek-Le Cam convolution theorem, this is the asymptotic distribution of the best regular estimator. Third, the Hájek asymptotic minimax theorem establishes that the minimax lower bound on the asymptotic risk among all estimators equals the asymptotic risk of the MLE. These results are typically summarized by the general assertion that "the MLE is asymptotically efficient".

In this paper we show that this understanding is incomplete. We show that the asymptotic risk of a feasible shrinkage estimator is strictly smaller than that of the MLE uniformly over all parameter sets local to the shrinkage parameter space. We also derive a local asymptotic minimax bound – which is a stronger efficiency bound than a global minimax bound – and show that our shrinkage estimator achieves this bound while the MLE does not. The local minimax bound is the lowest possible minimax risk, uniformly in local regions of the parameter space.

Our shrinkage estimator depends on three choices available to the researcher. First, the researcher needs to select a parameter of interest, which could be the entire parameter vector, a subset of the parameter vector, or a nonlinear transformation. Second, the researcher needs to select a loss function to evaluate the estimates of the parameter of interest. We require that the loss function is locally quadratic but otherwise do not restrict its choice. Third, the researcher needs to select a shrinkage direction, or equivalently a set of (potentially nonlinear) parametric restrictions. The purpose of the shrinkage direction is to provide a plausible set of potential simplifications to define the local regions to evaluate the minimax risk.

The shrinkage estimator takes a very simple form, as a weighted average of the unrestricted and restricted MLE, with the weight inversely proportional to the loss function evaluated at the two estimates. The estimator is fully data-dependent, with no need for selection of a tuning parameter. The estimator is a generalization of the classical James-Stein estimator in the normal sampling model.

We evaluate the performance of the estimator using the local asymptotic normality approach of Le Cam (1972) and 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 (2015), 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). We find that if the shrinkage dimension exceeds two, the asymptotic risk of the shrinkage estimator is strictly smaller than that of the unrestricted MLE, globally in the parameter space. 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 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.

Estimation risk (and the shrinkage estimator itself) depend on the choice of loss function. Our asymptotic theory shows that this choice only affects asymptotic performance via its local quadratic approximation. Thus there may be little reason in practice to use anything more complicated than quadratic loss. The choice of weight matrix, however, could be quite important in practice. As a default choice the weight matrix can be set to equal either the identity matrix or the inverse of the asymptotic covariance matrix, but in other cases a user may wish to select a specific weight matrix. Weighted MSE is a standard criterion in the shrinkage literature, including Bhattacharya (1966), Sclove (1968), and Berger (1976a, 1976b, 1982).

We also develop an high-dimensional local asymptotic minimax bound, and show that our shrinkage estimator achieves this bound and is thus locally asymptotically minimax efficient. Our local minimax bound differs from the classic Hájek (1972) bound by calculating the lowest possible risk in local regions of the parameter space, not just globally. This produces a minimax bound which is a function of the size of the local region. The minimax bound is high-dimensional because we use Pinkser's Theorem (Pinsker, 1980), which gives a lower minimax bound for estimation of high dimensional normal means. Our asymptotic minimax theorem combines Pinker's Theorem with classic large-sample minimax efficiency theory (Hájek, 1972) to provide a new asymptotic local minimax efficiency bound.

We show that the asymptotic risk of our shrinkage estimator equals the asymptotic local minimax efficiency bound on all local parameter regions, and is thus locally asymptotically minimax efficient. This is in contrast to the unrestricted MLE, which has larger local minimax risk and is thus locally minimax inefficient.

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 shrinkage employs a sequential asymptotic argument, where we first take limits as the sample size diverges and second take limits as the shrinkage dimension increases. A deeper theory would employ a joint asymptotic limit. Third, our analysis is confined to locally quadratic loss functions and thus excludes non-differentiable loss functions such as absolute error loss. Fourth, we do not provide methods for confidence interval construction or inference after shrinkage. These four limitations are important, 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 and showed that this estimator has smaller MSE than the MLE. 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).

Stein-type shrinkage is related to model averaging. In fact, in linear regression with two nested models, the Mallows model averaging (MMA) estimator of Hansen (2007) is precisely a Stein-type shrinakge estimator. This paper, however, is concerned with nonlinear likelihood-type contexts. While Bayesian model averaging techniques have been developed, and frequentist model averaging models have been introduced by Burnham and Anderson (2002) and Hjort and Claeskens (2003), there is no optimality theory concerning how to select the frequentist model averaging weights in the nonlinear likelihood context. One of the motivations behind this paper was to uncover the optimal method for model averaging. The risk bounds established in this paper demonstrate that – at least in the context of two models – there is a well-defined efficiency bound, and the Stein-type shrinkage estimator achieves this bound. Consequently, there is no need to consider alternative model averaging techniques.

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 primarily developed for regression models (see also Liao (2013) for application in GMM), while this paper focuses on likelihood models.

There is also some relationship between the methods proposed in this paper and methods designed to minimize an estimated loss function. These methods include the focused information criterion of Claeskens and Hjort (2003), the plug-in estimator of Liu (2015), and the focused moment selection criterion of DiTraglia (2014). These methods, however, do not have the minimax efficiency properties of the shrinkage-type estimators proposed here.

This paper is concerned exclusively with point estimation and explicitly ignores inference. Inference with shrinkage estimators poses particular challenges. A useful literature review of shrinkage confidence sets can be found on page 423 of Lehmann and Casella (1998). One related paper is Casella and Hwang (1987) who discuss shrinkage towards subspaces in the context of confidence interval construction. Particularly promising approaches include Tseng and Brown (1997), Beran (2010), and McCloskey(2015). Developing methods for the shrinkage estimators described in this

paper is an important topic for future research.

The organization of the paper is as follows. Section 2 presents the general framework, and describes the choice of loss function and shrinkage direction. Section 3 presents the shrinkage estimator. Section 4 presents the asymptotic distribution of the estimator. Section 5 develops a bound for its asymptotic risk. Section 6 uses a high-dimensional approximation, showing that the gains are substantial and broad in the parameter space. Section 7 contrasts our results with those for superefficient estimators. Section 8 presents the local minimax efficiency bound. Section 9 illustrates the performance in two simulation experiments, the first using a probit model, and the second using a model of exponential means. Mathematical proofs are left to the appendix.

2 Model

Suppose that we observe a random array $\widetilde{X}_n = \{X_{1n}, ..., X_{nn}\}$ of independent and identically distributed realizations from a density $f(x, \boldsymbol{\theta}_n)$ indexed by a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$. Let $\mathcal{I}_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}} \left(-\dfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(X_i, \boldsymbol{\theta}) \right)$ denote the information matrix, where $\mathbb{E}_{\boldsymbol{\theta}}$ denotes expectation with respect to the density $f(x, \boldsymbol{\theta})$.

The goal is estimation of a parameter of interest $\boldsymbol{\beta}_n = \mathbf{g}(\boldsymbol{\theta}_n)$ for some differentiable function $\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^q$. Let $\mathbf{G}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{g}(\boldsymbol{\theta})'$.

The accuracy of an estimator $\boldsymbol{T}_n = \boldsymbol{T}_n(\widetilde{X}_n)$ of $\boldsymbol{\beta}_n$ is measured by a known loss function $\ell(\boldsymbol{\beta}_n, \boldsymbol{T}_n)$, so that the risk is the expected loss

$$R(\boldsymbol{\beta}_n, \boldsymbol{T}_n) = \mathbb{E}_{\boldsymbol{\theta}_n} \ell(\boldsymbol{\beta}_n, \boldsymbol{T}_n). \tag{1}$$

The estimation setting is augmented by the belief that the true value of $\boldsymbol{\theta}_n$ may be close (in a sense to be made precise later) to a restricted parameter space $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ defined by a differentiable parametric restriction

$$\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}\} \tag{2}$$

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

The goal is local minimax estimation efficiency: parameter estimation with minimum expected loss, the latter maximized in local regions of the parameter space. The role of the restriction (2) will be to define the local regions to evaluate the minimax risk.

This setting is largely described by the choice of parameter of interest $\boldsymbol{\beta}$, the loss function $\ell(\boldsymbol{\beta}, \boldsymbol{T})$, and the parametric restriction $\mathbf{r}(\boldsymbol{\theta})$. We now describe these choices.

2.1 Parameter of Interest

In a general estimation context the parameter of interest $\boldsymbol{\beta} = \boldsymbol{\theta}$ may be the entire parameter vector, in which case $\mathbf{G}(\boldsymbol{\theta}) = \mathbf{I}_m$.

In other contexts only a subset $\boldsymbol{\theta}_1 \in \mathbb{R}^q$ may of interest where we have partitioned $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

In this case we can set $\boldsymbol{\beta} = \boldsymbol{\theta}_1$ and $\mathbf{G}(\boldsymbol{\theta}) = (\mathbf{I}_q, \mathbf{0})'$, and typically call $\boldsymbol{\theta}_2$ the "nuisance parameters". This situation is typical in many econometric contexts when there are a large number of control variables (e.g. fixed effects) whose coefficients are not reported and are thus not the focus of the study. However, it should be noted that the partitioning into "parameters of interest" and "nuisance parameters" depends on the context. For example, if interest is in the distribution of the fixed effects then these parameters would be treated as parameters of interest.

The parameter of interest can also be a (nonlinear) transformation of the parameter vector. For example in probit regression $\mathbb{P}(y_i = 1 \mid X_i = x) = \Phi(x'\boldsymbol{\theta})$ the parameter of interest could be the parameters $\boldsymbol{\theta}$, the marginal effects $\boldsymbol{\beta} = \boldsymbol{\theta}\phi(x'\boldsymbol{\theta})$, or the probability $\boldsymbol{\beta} = \Phi(x'\boldsymbol{\theta})$ at a specific value of x or at a set of values of x.

2.2 Loss and Risk Function

We will require that the loss function $\ell(\boldsymbol{\beta}, \boldsymbol{T})$ has a second derivative with respect to the second argument. (See Assumption 2 in Section 4). This allows for smooth loss functions such as quadratic and linex, but excludes non-smooth loss functions such as absolute value and check function loss. For our minimax efficiency theory we will also impose that the loss function $\ell(\boldsymbol{\beta}, \boldsymbol{T})$ is uniquely minimized at $\boldsymbol{T} = \boldsymbol{\beta}$. (See Assumption 4 in Section 8.)

A leading loss function of interest is weighted squared error, which takes the form

$$\ell(\boldsymbol{\beta}, \boldsymbol{T}) = (\boldsymbol{T} - \boldsymbol{\beta})' \mathbf{W} (\boldsymbol{T} - \boldsymbol{\beta}) \tag{3}$$

for some weight matrix $\mathbf{W} > 0$. Risk under weighted squared error loss is weighted mean-squared error, since in this case

$$\begin{aligned} R(\boldsymbol{\beta}, \boldsymbol{T}_n) &= \operatorname{tr}\left[\mathbf{W}\mathbb{E}_{\boldsymbol{\theta}}(\boldsymbol{T}_n - \boldsymbol{\beta})(\boldsymbol{T}_n - \boldsymbol{\beta})'\right] \\ &= \sum_{j=1}^{q}\sum_{k=1}^{q} W_{jk}\mathbb{E}_{\boldsymbol{\theta}}\left[(T_{nj} - \beta_j)(T_{nk} - \beta_k)\right]. \end{aligned}$$

Allowing for flexible choice of the weight matrix incorporates many important potential applications.

One generic choice for the weight matrix is $\mathbf{W} = \mathbf{I}_q$ so that (3) is unweighted quadratic loss. This can be appropriate when the coefficients in $\boldsymbol{\beta}$ are scaled to be of roughly equal magnitude and are of equal importance. Another generic choice is $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{-1} = \left(\mathbf{G}'\mathcal{I}_{\boldsymbol{\theta}}^{-1}\mathbf{G}\right)^{-1}$, the inverse of the asymptotic variance of the MLE for $\boldsymbol{\beta}$ (the MLE for $\boldsymbol{\beta}$ is discussed in Section 3.1). Setting $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{-1}$ is an excellent choice as it renders the loss function invariant to rotations of the coefficient vector $\boldsymbol{\beta}$. The choice $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{-1}$ also has the advantage that some formulae will simplify, as we shall see later. Because of these advantages, we call $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{-1}$ the **default choice** for the weight matrix.

We now list some examples of loss functions derived directly from econometric problems.

Example 1 *In the (possibly nonlinear) regression model* $y_i = g(x_i, \boldsymbol{\beta}) + e_i$ *with* $e_i \sim f(e, \eta)$ *for some parametric density* $f(e, \eta)$*, an estimate of the conditional mean function takes the form* $g(x, \boldsymbol{T})$*. A common measure of accuracy is the integrated distance, which is*

$$\ell(\boldsymbol{\beta}, \boldsymbol{T}) = \int d\left(g(x, \boldsymbol{T}) - g(x, \boldsymbol{\beta})\right) w(x) dx$$

for some smooth distance measure $d(u) \geq 0$*. In general,* $\ell(\boldsymbol{\beta}, \boldsymbol{T})$ *is a nonquadratic yet smooth function of* $\boldsymbol{T}$*. If the regression function is linear in the parameters,* $g(x, \boldsymbol{\beta}) = x'\boldsymbol{\beta}$ *, and* $d(u) = u^2$ *is quadratic, then the integrated squared error equals*

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{T}) &= \int \left(x'\boldsymbol{T} - x'\boldsymbol{\beta}\right)^2 w(x) dx \\ &= (\boldsymbol{\beta} - \boldsymbol{T})' \mathbf{W} (\boldsymbol{\beta} - \boldsymbol{T}) \end{aligned}$$

with $\mathbf{W} = \int xx' w(x) dx = \mathbb{E}\left(x_i x_i' w(x_i)\right)$*, and thus takes the form (3) with a specific weight matrix.*

Example 2 *Suppose we are interested in a decision problem where we wish to take an action* $\mathbf{a}$ *to maximize a smooth utility (or profit) function* $u(\mathbf{a}, \boldsymbol{\theta})$*. The optimal action given* $\boldsymbol{\theta}$ *is* $\mathbf{a}(\boldsymbol{\theta}) = \operatorname{argmax}_{\mathbf{a}} u(\mathbf{a}, \boldsymbol{\theta})$ *and the estimated version is* $\mathbf{a}(\boldsymbol{T})$*. The loss in this context is the difference between optimality utility and realized utility, which is* $\ell(\boldsymbol{\theta}, \boldsymbol{T}) = u(\mathbf{a}(\boldsymbol{\theta}), \boldsymbol{\theta}) - u(\mathbf{a}(\boldsymbol{T}), \boldsymbol{\theta})$*. In general, this loss is a nonquadratic yet smooth function of* $\boldsymbol{T}$*.*

Example 3 *The estimate of the density of* x_i *is* $f(x, \boldsymbol{T})$*. Suppose the goal is to estimate this density accurately. Measures of the distance between densities include the KLIC and Hellinger distance, which give rise to the loss functions*

$$\ell_1(\boldsymbol{\theta}, \boldsymbol{T}) = \int \log\left(\frac{f(x, \boldsymbol{T})}{f(x, \boldsymbol{\theta})}\right) f(x, \boldsymbol{T}) dx,$$

$$\ell_2(\boldsymbol{\theta}, \boldsymbol{T}) = \int \log\left(\frac{f(x, \boldsymbol{\theta})}{f(x, \boldsymbol{T})}\right) f(x, \boldsymbol{\theta}) dx,$$

and

$$\ell_3(\boldsymbol{\theta}, \boldsymbol{T}) = \frac{1}{2} \int \left(f(x, \boldsymbol{\theta})^{1/2} - f(x, \boldsymbol{T})^{1/2}\right)^2 dx.$$

These are (generally) nonquadratic yet smooth functions of $\boldsymbol{T}$*, yet satisfy the properties of a loss function.*

These examples show that in some contexts the loss function can be motivated by the econometric problem. In general, the choice of loss function is important as the shrinkage estimator will depend on its choice.

2.3 Shrinkage Direction

The restriction (2) defines the shrinkage direction, and its choice is critical for the construction of our shrinkage estimator. The restriction is not believed to be true, but instead represents a belief about where $\boldsymbol{\theta}$ is likely to be close. The restricted space $\boldsymbol{\Theta}_0$ should be a plausible simplification, centering, or "prior" about the likely value of $\boldsymbol{\theta}$. The restriction can be a tight specification, a structural model, a set of exclusion restrictions, parameter symmetry (such as coefficient equality) or any other restriction.

Restricted parameter spaces are common in applied economics, as they are routinely specified for the purpose of hypothesis testing. Restrictions are often tested because they are believed to be plausible simplifications of the unrestricted model specification. Rather than using such restrictions for testing, these restrictions can be used to construct a shrinkage estimator, and thereby improve estimation efficiency. Consequently, since restrictions are routine in applied economics, shrinkage estimators can be made similarly routine.

An important (and classical) 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**, and is probably the more relevant empirical situation. 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.

Quite commonly, $\boldsymbol{\Theta}_0$ 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 $\mathbf{r}(\boldsymbol{\theta}) = \boldsymbol{\theta}_2$ and $\mathbf{R} = (\mathbf{0}, \mathbf{I}_p)'$. This is appropriate when the model with $\boldsymbol{\theta}_2 = \mathbf{0}$ is viewed as a useful yet parsimonous approximation. For example, $\boldsymbol{\theta}_2$ may be the coefficients for a large number of control variables which are included in the model for robustness but whose magnitude is a priori unclear. This is a common situation in applied economics.

Alternatively, $\boldsymbol{\Theta}_0$ may also be a linear subspace, in which case we can write

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{R}'\boldsymbol{\theta} - \mathbf{a} \tag{4}$$

where $\mathbf{R}$ is $m \times p$ and $\mathbf{a}$ is $p \times 1$. One common example is to shrink the elements of $\boldsymbol{\theta}$ to a common

mean, in which case we would set

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 0 \\ & & \vdots \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix} \tag{5}$$

and $\mathbf{a} = \mathbf{0}$. This is appropriate when $\boldsymbol{\theta}$ are disaggregate coefficients (such as slope coefficients for heterogenous groups) and a useful approximation is to shrink these coefficients to a common value.

In other cases, $\boldsymbol{\Theta}_0$ may be a nonlinear subspace. For example one of the restrictions could be $r_j(\boldsymbol{\theta}) = \theta_1\theta_2 - 1$ which would shrink the coefficients towards $\theta_1\theta_2 = 1$. In general, nonlinear restrictions may be useful when an economic model or hypothesis implies a set of nonlinear restrictions on the coefficients.

The shrinkage direction $\mathbf{r}(\boldsymbol{\theta})$ may, but does not necessarily need to, relate to the parameter of interest $\mathbf{g}(\boldsymbol{\theta})$. Thus the matrices $\mathbf{R}$ and $\mathbf{G}$ can be the same, different, or one can be a subset of the other. These cases are all allowed, and should be dictated by the underlying economic problem and focus of the study.

2.4 Canonical Case

The primary focus of the theoretical shrinkage literature has been on the simplified setting where $\boldsymbol{\beta} = \boldsymbol{\theta}$ (the full parameter vector is of interest and thus $\mathbf{G} = \mathbf{I}_m$) and the loss function is (3) with $\mathbf{W} = \mathbf{V}^{-1}$ (our recommended default choice when $\boldsymbol{\beta} = \boldsymbol{\theta}$). Under these assumptions many of the formulae simplify. We will consequently call the setting $\mathbf{G} = \mathbf{I}_m$ and $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{-1}$ the **canonical case**, and use these conditions to sharpen our understanding of the assumptions.

For applications, however, the canonical case is overly restrictive so our methods will not require these choices.

3 Estimation

3.1 Unrestricted Estimator

The standard estimator of $\boldsymbol{\theta}$ is the unrestricted maximum likelihood estimator (MLE). The log likelihood function is

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}). \tag{6}$$

The MLE maximizes (6) over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

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

We assume that the maximum is unique so that $\widehat{\boldsymbol{\theta}}_n$ is well defined (and similarly with the other extremum estimators defined below). Let $\widehat{\mathbf{V}}_n$ denote any consistent estimate of the asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$, such as

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

The MLE for $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_n = \mathbf{g}(\widehat{\boldsymbol{\theta}}_n)$. Let $\widehat{\mathbf{V}}_{\boldsymbol{\beta}} = \widehat{\mathbf{G}}_n' \widehat{\mathbf{V}}_n \widehat{\mathbf{G}}_n$ denote the standard estimator of the asymptotic covariance matrix for $\widehat{\boldsymbol{\beta}}_n$, where $\widehat{\mathbf{G}}_n = \mathbf{G}(\widehat{\boldsymbol{\theta}}_n)$.

3.2 Restricted Estimator

We wish to contract (or shrink) $\widehat{\boldsymbol{\theta}}_n$ towards the restricted space $\boldsymbol{\Theta}_0$. To do we first define a restricted estimator $\widetilde{\boldsymbol{\theta}}_n$ which (at least approximately) satisfies $\widetilde{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}_0$. We consider three possible restricted estimators.

1. Restricted maximum likelihood (RML)

$$\widetilde{\boldsymbol{\theta}}_n^R = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0}{\operatorname{argmax}} \mathcal{L}_n(\boldsymbol{\theta}). \tag{7}$$

2. Efficient minimum distance (EMD)

$$\widetilde{\boldsymbol{\theta}}_n^E = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0}{\operatorname{argmin}} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \right)' \widehat{\mathbf{V}}_n^{-1} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \right). \tag{8}$$

3. Projection

$$\widetilde{\boldsymbol{\theta}}_n^P = \widehat{\boldsymbol{\theta}}_n - \widehat{\mathbf{V}}_n \widehat{\mathbf{R}}_n \left(\widehat{\mathbf{R}}_n' \widehat{\mathbf{V}}_n \widehat{\mathbf{R}}_n \right)^{-1} \mathbf{r}(\widehat{\boldsymbol{\theta}}_n) \tag{9}$$

for $\widehat{\mathbf{R}}_n = \mathbf{R}(\widehat{\boldsymbol{\theta}}_n)$.

The RML and EMD estimators satisfy $\widetilde{\boldsymbol{\theta}}_n^R, \widetilde{\boldsymbol{\theta}}_n^E \in \boldsymbol{\Theta}_0$, and the projection estimator satisfies $\widetilde{\boldsymbol{\theta}}_n^P \in \boldsymbol{\Theta}_0$ for linear restrictions (4). For non-linear restrictions, however, the projection estimator only asymptotically satisfies the restriction. The three estimators are asymptotically equivalent under our assumptions and thus we will generically write the restricted estimator as $\widetilde{\boldsymbol{\theta}}_n$ without drawing a distinction between (7), (8) and (9). It would also be possible to consider minimum distance or projection estimators as in (8) or (9) but with the weight matrix $\widehat{\mathbf{V}}_n^{-1}$ replaced with another choice. Other choices, however, would lead to asymptotically inefficient estimators, so we confine attention to the estimators (8)-(9).

Given the estimator $\widetilde{\boldsymbol{\theta}}_n$, our plug-in estimator for $\boldsymbol{\beta}$ is $\widetilde{\boldsymbol{\beta}}_n = \mathbf{g}(\widetilde{\boldsymbol{\theta}}_n)$.

3.3 Shrinkage Estimator

Our proposed shrinkage estimator of $\boldsymbol{\beta}$ is a weighted average of the MLE and the restricted estimator

$$\widehat{\boldsymbol{\beta}}_n^* = \widehat{w}_n \widehat{\boldsymbol{\beta}}_n + (1 - \widehat{w}_n)\widetilde{\boldsymbol{\beta}}_n \tag{10}$$

where the weight takes the form

$$\widehat{w}_n = \left(1 - \frac{\widehat{\tau}_n}{D_n}\right)_+. \tag{11}$$

In (11), $(x)_+ = x1(x \geq 0)$ is the "positive-part" function, the scalar "shrinkage parameter" $\widehat{\tau}_n \geq 0$ controls the degree of shrinkage, and D_n equals

$$D_n = n\ell\left(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n\right). \tag{12}$$

In the case of weighted quadratic loss (3) we have the simplication

$$D_n = n\left(\widehat{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_n\right)' \mathbf{W} \left(\widehat{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_n\right) \tag{13}$$

with $\mathbf{W}$ the weight matrix from (3). The statistic (12) is the scaled loss between the unrestricted and restricted estimators, and (13) is a distance-type statistic for the restriction (2) in $\boldsymbol{\Theta}$.

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

We will defer describing our recommended choice of shrinkage parameter $\widehat{\tau}_n$ until Section 5. At that point we will recommend a specific data-dependent choice (see equation (33)).

Simplifications occur in the canonical case ($\mathbf{G} = \mathbf{I}_m$ and $\mathbf{W} = \mathbf{V}^{-1}$). The full shrinkage estimator is the classic James-Stein estimator. The partial shrinkage estimator with linear $\mathbf{r}(\boldsymbol{\theta})$ is similar to Oman's (1982a) shrinkage estimator. The difference is that Oman (1982a) uses the estimator (9) with $\widehat{\mathbf{V}}_n$ replaced by $\mathbf{I}_m$, thus $\widetilde{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n - \widehat{\mathbf{R}}_n\left(\widehat{\mathbf{R}}_n'\widehat{\mathbf{R}}_n\right)^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\theta}}_n - \mathbf{a}\right)$, which is an inefficient choice unless $\mathbf{V} = \mathbf{I}_m$. The partial shrinkage estimator is also a special case of Hansen's (2007) Mallows Model Averaging (MMA) estimator with two models.

4 Asymptotic Distribution

Our analysis is asymptotic as sample size $n \to \infty$. We use the local asymptotic normality approach of Le Cam (1972) and van der Vaart (1998). Consider parameter sequences of the form

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h} \tag{14}$$

where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$ and $\boldsymbol{h} \in \mathbb{R}^m$. In this framework $\boldsymbol{\theta}_n$ is the true value of the parameter, $\boldsymbol{\theta}_0$ is a centering value and $\boldsymbol{h} \in \mathbb{R}^m$ is a localizing parameter. Given (14), we define $\boldsymbol{\beta}_n = \mathbf{g}(\boldsymbol{\theta}_n)$ as the

true value of the parameter of interest, and $\boldsymbol{\beta}_0 = \mathbf{g}(\boldsymbol{\theta}_0)$ as its centering value.

What is important about the parameter sequences (14) is that the centering value $\boldsymbol{\theta}_0$ is specified to lie in the restricted parameter space $\boldsymbol{\Theta}_0$. This means that the true parameter $\boldsymbol{\theta}_n$ is localized to $\boldsymbol{\Theta}_0$. This is what we mean when we say that the true parameter is close to the restricted parameter space.

The magnitude of the distance between the parameter $\boldsymbol{\theta}_n$ and the restricted set $\boldsymbol{\Theta}_0$ is determined by the localizing parameter $\boldsymbol{h}$ and the sample size n. For any fixed $\boldsymbol{h}$ the distance $n^{-1/2}\boldsymbol{h}$ shrinks as the sample size increases. However, we do not restrict the magnitude of $\boldsymbol{h}$ so 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 (14).

The following is a standard set of regularity conditions sufficient for asymptotic normality of the conventional estimators.

Assumption 1

1. *The observations X_{ni} are independent, identically distributed draws from the density $f(x, \boldsymbol{\theta}_n)$, where $\boldsymbol{\theta}_n$ satisfies (14), $\boldsymbol{\theta}_0$ is in the interior of $\boldsymbol{\Theta}_0$, and $\boldsymbol{\Theta}$ is compact.*
2. *If $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ then $f(x, \boldsymbol{\theta}) \neq f(x, \boldsymbol{\theta}')$.*
3. *$\log f(x, \boldsymbol{\theta})$ is continuous at each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with probability one.*
4. $\mathbb{E}_{\boldsymbol{\theta}_0} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\log f(X_{ni}, \boldsymbol{\theta})| < \infty$.
5. $\mathbb{E}_{\boldsymbol{\theta}_0} \dfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(X_{ni}, \boldsymbol{\theta}_0)$ *exists and is non-singular.*
6. *For some neighborhood $\aleph$ of $\boldsymbol{\theta}_0$,*
   (a) *$f(x, \boldsymbol{\theta})$ is twice continuously differentiable,*
   (b) $\int \sup_{\boldsymbol{\theta} \in \aleph} \left\| \dfrac{\partial}{\partial \boldsymbol{\theta}} f(x, \boldsymbol{\theta}) \right\| dx < \infty$,
   (c) $\int \sup_{\boldsymbol{\theta} \in \aleph} \left\| \dfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(x, \boldsymbol{\theta}) \right\| dx < \infty$,
   (d) $\mathbb{E}_{\boldsymbol{\theta}_0} \sup_{\boldsymbol{\theta} \in \aleph} \left\| \dfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(X_{ni}, \boldsymbol{\theta}) \right\| < \infty$.
7. *$\mathbf{R}(\boldsymbol{\theta})$ is continuous in some neighborhood of $\boldsymbol{\theta}_0$, and* $\operatorname{rank}(\mathbf{R}(\boldsymbol{\theta}_0)) = p$.
8. *$\mathbf{G}(\boldsymbol{\theta})$ is continuous in some neighborhood of $\boldsymbol{\theta}_0$.*

Assumptions 1.1-1.6 are the conditions listed in Theorem 3.3 of Newey and McFadden (1994) for the asymptotic normality of the MLE $\widehat{\boldsymbol{\theta}}_n$. Assumption 1.1 specifies that the observations are iid, that the parameter satisfies the local sequences (14) and that the centering value is in the interior of $\boldsymbol{\Theta}_0$ so that Taylor expansion methods can be employed. Assumption 1.2 establishes identification.

Assumptions 1.3 and 1.4 are used to establish consistency of $\widehat{\boldsymbol{\theta}}_n$. Assumption 1.5 requires the Fisher information to exist. Assumption 1.6 are regularity conditions to establish asymptotic normality.

Assumptions 1.7 and 1.8 allow asymptotic normality to extend to the estimators $\widehat{\boldsymbol{\beta}}_n$, $\widetilde{\boldsymbol{\theta}}_n$, and $\widetilde{\boldsymbol{\beta}}_n$.

We now specify regularity conditions for the loss function.

Assumption 2 *The loss function $\ell(\boldsymbol{\beta}, \boldsymbol{T})$ satisfies*

1. $\ell(\boldsymbol{\beta}, \boldsymbol{T}) \geq 0$
2. $\ell(\boldsymbol{\beta}, \boldsymbol{\beta}) = 0$
3. $\mathbf{W}(\boldsymbol{\beta}) = \frac{1}{2} \frac{\partial^2}{\partial \boldsymbol{T} \partial \boldsymbol{T}'} \ell(\boldsymbol{\beta}, \boldsymbol{T}) \Big|_{\boldsymbol{T}=\boldsymbol{\beta}}$ *is continuous in a neighborhood of $\boldsymbol{\beta}_0$.*

Assumptions 2.1 and 2.2 are conventional properties of loss function (as discussed, for example, in Chapter 1.1 of Lehmann and Casella (1998)). Assumption 2.3 is stronger, requiring the loss function to be smooth (locally quadratic). Under weighted quadratic loss (3), $\mathbf{W}(\boldsymbol{\beta}) = \mathbf{W}$. In general, we define $\mathbf{W} = \mathbf{W}(\boldsymbol{\beta}_0)$.

Finally, we allow the shrinkage parameter $\widehat{\tau}_n$ to be data-dependent and random but require that it converge in probability to a non-negative constant.

Assumption 3 $\widehat{\tau}_n \to_p \tau \geq 0$ *as* $n \to \infty$.

We can now present the asymptotic distributions of the maximum likelihood, restricted, and shrinkage estimators. Define the matrices $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta}_0)$, $\mathbf{R} = \mathbf{R}(\boldsymbol{\theta}_0)$,

$$\mathbf{V} = \left(\mathbb{E}_{\boldsymbol{\theta}_0} \left(-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(X_{ni}, \boldsymbol{\theta}_0) \right) \right)^{-1},$$

and

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

Theorem 1 *Under Assumptions 1-3, along the sequences (14),*

$$\sqrt{n} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V}), \tag{16}$$

$$\sqrt{n} \left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathrm{Z} - \mathbf{V}\mathbf{R} \left(\mathbf{R}'\mathbf{V}\mathbf{R} \right)^{-1} \mathbf{R}' (\mathrm{Z} + \boldsymbol{h}), \tag{17}$$

$$n\ell \left(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n \right) \xrightarrow[\boldsymbol{\theta}_n]{} \xi = (\mathrm{Z} + \boldsymbol{h})' \mathbf{B} (\mathrm{Z} + \boldsymbol{h}), \tag{18}$$

$$\widehat{w} \xrightarrow[\boldsymbol{\theta}_n]{} w(\mathrm{Z}) = \left(1 - \frac{\tau}{\xi} \right)_+. \tag{19}$$

The asymptotic distribution of the shrinkage estimator is

$$\sqrt{n} \left(\widehat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n \right) \xrightarrow[\boldsymbol{\theta}_n]{} w(\mathrm{Z}) \mathbf{G}'\mathrm{Z} + (1 - w(\mathrm{Z})) \mathbf{G}' \left(\mathrm{Z} - \mathbf{V}\mathbf{R} \left(\mathbf{R}'\mathbf{V}\mathbf{R} \right)^{-1} \mathbf{R}' (\mathrm{Z} + \boldsymbol{h}) \right). \tag{20}$$

Furthermore, equations (16)-(20) hold jointly.

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

The asymptotic distribution is obtained for parameter sequences $\boldsymbol{\theta}_n$ tending towards a point in the restricted parameter space $\boldsymbol{\Theta}_0$. The case of fixed $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$ can be obtained by letting $\boldsymbol{h}$ diverge towards infinity, in which case $\xi \to_p \infty$, $w(\mathrm{Z}) \to_p 1$, and the distribution on the right-hand-side of (20) tends towards $\mathbf{G}'\mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$ where $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{G}'\mathbf{V}\mathbf{G}$.

It is important to understand that Theorem 1 does not require that the true parameter value $\boldsymbol{\theta}_n$ 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 $\boldsymbol{h}$ can be arbitrarily large.

Since the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n^*$ in (20) depends on $\boldsymbol{h}$, the estimator $\widehat{\boldsymbol{\beta}}_n^*$ is non-regular. (An estimator $\boldsymbol{T}_n$ is called regular if $\sqrt{n}\left(\boldsymbol{T}_n - \boldsymbol{\beta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \psi$ for some random variable ψ which does not depend on $\boldsymbol{h}$. See van der Vaart (1998, p. 115).)

The matrix $\mathbf{B}$ defined in (15) will play an important role in our theory. Notice that if $\mathbf{G}$ is in the range space of $\mathbf{R}$ then we have the simplification $\mathbf{B} = \mathbf{G}\mathbf{W}\mathbf{G}'$. This occurs when all parameters of interest $\boldsymbol{\beta}$ are included in the restrictions (4), and includes in particular the special cases $\mathbf{R} = \mathbf{I}_m$ and $\mathbf{R} = \mathbf{G}$. Also, note that in the full shrinkage canonical case ($\mathbf{R} = \mathbf{G} = \mathbf{I}_m$ and $\mathbf{W} = \mathbf{V}^{-1}$) then $\mathbf{B} = \mathbf{V}^{-1}$.

Equation (18) also provides the asymptotic distribution of the distance-type statistic D_n. The limit random variable ξ controls the weight $w(\mathrm{Z})$ and thus the degree of shrinkage, so it is worth investigating further. Notice that its expected value is

$$\mathbb{E}\xi = \boldsymbol{h}'\mathbf{B}\boldsymbol{h} + \mathbb{E}\operatorname{tr}\left(\mathbf{B}\mathrm{Z}\mathrm{Z}'\right) = \boldsymbol{h}'\mathbf{B}\boldsymbol{h} + \operatorname{tr}\left(\mathbf{B}\mathbf{V}\right). \tag{21}$$

In the canonical case $\mathbf{G} = \mathbf{I}_m$ and $\mathbf{W} = \mathbf{V}^{-1}$ we find that (21) simplifies to

$$\mathbb{E}\xi = \boldsymbol{h}'\mathbf{B}\boldsymbol{h} + p. \tag{22}$$

Furthermore, in this 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.

5 Asymptotic Risk

The risk $R(\boldsymbol{\beta}, \boldsymbol{T}_n)$ of an estimator $\boldsymbol{T}_n$ in general is difficult to evaluate, and may not even be finite unless $\boldsymbol{T}_n$ has sufficient finite moments. To obtain a useful approximation and ensure existence we use a trimmed loss and take limits as the sample size $n \to \infty$.

Specifically, let $\boldsymbol{T} = \{\boldsymbol{T}_n : n = 1, 2, ...\}$ denote a sequence of estimators. We define the **asymptotic risk** of the estimator sequence $\boldsymbol{T}$ for the parameter sequence $\boldsymbol{\theta}_n$ defined in (14) as

$$\rho(\boldsymbol{h}, \boldsymbol{T}) = \lim_{\zeta \to \infty} \liminf_{n \to \infty} \mathbb{E}_{\boldsymbol{\theta}_n} \min\left[n\ell\left(\boldsymbol{\beta}_n, \boldsymbol{T}_n\right), \zeta\right]. \tag{23}$$

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

This definition of asymptotic risk is convenient as it is well defined and easy to calculate whenever the estimator has an asymptotic distribution.

Lemma 1 *For any estimator $\boldsymbol{T}_n$ satisfying*

$$\sqrt{n}\left(\boldsymbol{T}_n - \boldsymbol{\beta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \psi, \tag{24}$$

for some random variable ψ, if $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$ as $n \to \infty$, and for any loss function satisfying Assumption 2, then

$$\rho(\boldsymbol{h}, \boldsymbol{T}) = \mathbb{E}\left(\psi' \mathbf{W} \psi\right). \tag{25}$$

Lemma 1 shows that the asymptotic risk of any estimator $\boldsymbol{T}_n$ satisfying (24) can be calculated using (25), which is expected weighted quadratic loss. This shows that for such estimators and smooth loss functions only the local properties of the loss function affect the asymptotic risk. Since Theorem 1 provides the asymptotic distribution of the shrinkage estimator, Lemma 1 provides a method to calculate its asymptotic risk.

Define the $q \times q$ matrix

$$\mathbf{A} = \mathbf{W}^{1/2\prime} \mathbf{G}' \mathbf{V} \mathbf{R} \left(\mathbf{R}' \mathbf{V} \mathbf{R}\right)^{-1} \mathbf{R}' \mathbf{V} \mathbf{G} \mathbf{W}^{1/2}, \tag{26}$$

let $\|\mathbf{A}\| = \lambda_{\max}(\mathbf{A})$ denote its largest eigenvalue, and define the ratio

$$d = \frac{\operatorname{tr}(\mathbf{A})}{\|\mathbf{A}\|}. \tag{27}$$

Notice that (27) satisfies $0 \leq d \leq \min[q, p]$. As discussed below, $d = \min[q, p]$ in the leading context when the weight matrix equals the default choice $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{-1}$ and in addition either $\mathbf{R}$ is in the range space of $\mathbf{G}$ (all restricted parameters are parameters of interest) or $\mathbf{G}$ is in the range space of $\mathbf{R}$ (all parameters of interest are restricted parameters). In general, d can be thought of as the effective shrinkage dimension: the number of restrictions on the parameters of interest.

Theorem 2 *Under Assumptions 1-3, if*

$$d > 2 \tag{28}$$

and

$$0 < \tau \leq 2\left(\operatorname{tr}(\mathbf{A}) - 2\|\mathbf{A}\|\right), \tag{29}$$

then for any $\boldsymbol{h}$

$$\rho(\boldsymbol{h}, \widehat{\boldsymbol{\beta}}^*) < \rho(\boldsymbol{h}, \widehat{\boldsymbol{\beta}}) = \operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}). \tag{30}$$

Furthermore, for any $0 < c < \infty$*, if we define the ball*

$$\mathbf{H}(c) = \left\{\boldsymbol{h} : \boldsymbol{h}'\mathbf{B}\boldsymbol{h} \leq \operatorname{tr}(\mathbf{A})\, c\right\} \tag{31}$$

with $\mathbf{A}$ *and* $\mathbf{B}$ *defined in (26) and (15), then*

$$\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \rho(\boldsymbol{h}, \widehat{\boldsymbol{\beta}}^*) \leq \operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}) - \frac{\tau\left(2\left(\operatorname{tr}(\mathbf{A}) - 2\left\|\mathbf{A}\right\|\right) - \tau\right)}{\operatorname{tr}(\mathbf{A})\,(c+1)}. \tag{32}$$

Equation (30) 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 (29). As (30) holds for even extremely large values of $\boldsymbol{h}$, this inequality shows that in a very real sense the shrinkage estimator strictly dominates the MLE.

The set $\mathbf{H}(c)$ defined in (31) is the set of localizing parameters $\boldsymbol{h}$ which are close to 0 with respect to the matrix $\mathbf{B}$ defined in (15). In the full shrinkage case $\mathbf{B}$ is full rank so $\mathbf{H}(c)$ is a ball. The magnitude of c controls the dimension of this ball. In the partial shrinkage case $\mathbf{B}$ has less than full rank so $\mathbf{H}(c)$ is tube-shaped. We can think of it as a ball with respect to the coefficients being shrunk and is the full parameter space in the other directions. The supremum in (32) is the largest risk across all localing parameters $\boldsymbol{h}$ in the set $\mathbf{H}(c)$, and thus contains all (unbounded) localizing parameters in the directions which are not being shrunk, and those in a ball with radius proportional to c in the directions which are being shrunk. The bound in (32) shows that the maximal risk in this set of localizing parameters is a function of this radius c; in particular the bound is tightest for small values of c.

The shrinkage parameter τ appears in the risk bound (32) as a quadratic expression, so there is a unique choice $\tau^* = \operatorname{tr}(\mathbf{A}) - 2\left\|\mathbf{A}\right\|$ which minimizes this bound. This is the optimal shrinkage parameter and our ideal choice.

The assumption (28) $d > 2$ is the critical condition needed to ensure that the shrinkage estimator can have globally smaller asymptotic risk than the MLE. The constant d defined in (27) is a measure of the sphericity of the matrix $\mathbf{A}$, and plays an important role in Theorem 2 and our theory which follows. The constant simplifies in certain special cases. The most important is when the weight matrix $\mathbf{W}$ has been set to equal what we have called the default choice $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{-1}$ and in addition either $\mathbf{R}$ is in the range space of $\mathbf{G}$ (all restricted parameters are parameters of interest) or $\mathbf{G}$ is in the range space of $\mathbf{R}$ (all parameters of interest are restricted parameters). Under these conditions we have the simplifications $\left\|\mathbf{A}\right\| = 1$, $\operatorname{tr}(\mathbf{A}) = \min[q, p]$, $d = \min[q, p]$, and $\tau^* = \min[q, p] - 2$. This includes the canonical case (where all parameters are of interest and $\mathbf{W} = \mathbf{V}^{-1}$) where $d = p$ and $\tau^* = p - 2$ which is the shrinkage parameter recommended by James and Stein (1961). In the canonical case (28) is equivalent to $p > 2$, which is Stein's (1956) classic condition 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. $d > 2$ is the generalization beyond the canonical case, to allow for general parameters of interest, general nonlinear restrictions, and general locally quadratic risk functions.

In the general case, the assumption $d > 2$ requires both $p > 2$ and $q > 2$, and excludes highly unbalanced matrices $\mathbf{A}$. The requirement $p > 2$ and $q > 2$ means that there are a minimum of three restrictions imposed in (2) and a minimum of three parameters of interest. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$ denote the ordered eigenvalues values of $\mathbf{A}$. $d > 2$ is equivalent to $\lambda_2 + \cdots + \lambda_q > \lambda_1$. This is violated only if λ_1 is much larger than the other eigenvalues. One sufficient condition for $d > 2$ is that for some $j > 1$, $\lambda_1/\lambda_j < j - 1$, in words, that the ratio of the largest eigenvalues to the j'th largest is not too large.

The condition $d > 2$ is necessary in order for the right-hand-side of (29) to be positive, which is necessary for the existence of a τ satisfying (29). The condition (29) simplifies to $0 < \tau \leq 2(p-2)$ in the canonical case, which is a standard restriction on the shrinkage parameter.

While the optimal shrinkage parameter $\tau^* = \operatorname{tr}(\mathbf{A}) - 2\|\mathbf{A}\|$ is generally unknown, it can be estimated. Set $\widehat{\mathbf{R}}_n = \mathbf{R}(\widehat{\boldsymbol{\theta}}_n)$, $\widehat{\mathbf{G}}_n = \mathbf{G}(\widehat{\boldsymbol{\theta}}_n)$, and the weight matrix estimate $\widehat{\mathbf{W}}_n$ of $\mathbf{W}$ can either be constructed from the specific context or as the second derivative of the loss function, e.g. $\widehat{\mathbf{W}}_n = \mathbf{W}\left(\widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\beta}}_n\right)$. We can then estimate τ^* by

$$\widehat{\tau}_n^* = \operatorname{tr}(\widehat{\mathbf{A}}_n) - 2\left\|\widehat{\mathbf{A}}_n\right\| \tag{33}$$

where

$$\widehat{\mathbf{A}}_n = \widehat{\mathbf{W}}_n^{1/2\prime}\widehat{\mathbf{G}}_n'\widehat{\mathbf{V}}_n\widehat{\mathbf{R}}_n\left(\widehat{\mathbf{R}}_n'\widehat{\mathbf{V}}_n\widehat{\mathbf{R}}_n\right)^{-1}\widehat{\mathbf{R}}_n'\widehat{\mathbf{V}}_n\widehat{\mathbf{G}}_n\widehat{\mathbf{W}}_n^{1/2}. \tag{34}$$

(33) is our practical recommendation for the shrinkage parameter so that the shrinkage estimator (10) is fully data-dependent. Notice that when $\mathbf{W} = \mathbf{I}_q$ then we can simplify (34) to

$$\widehat{\mathbf{A}}_n = \widehat{\mathbf{G}}_n'\widehat{\mathbf{V}}_n\widehat{\mathbf{R}}_n\left(\widehat{\mathbf{R}}_n'\widehat{\mathbf{V}}_n\widehat{\mathbf{R}}_n\right)^{-1}\widehat{\mathbf{R}}_n'\widehat{\mathbf{V}}_n\widehat{\mathbf{G}}_n.$$

We now reexpress Theorem 2 when the shrinkage parameter is set according to the recommendation (33).

Corollary 1 *If Assumptions 1 and 2 hold, $d > 2$, and $\widehat{\tau}_n^*$ is set by (33), then $\widehat{\tau}_n^* \to_p \tau^* = \operatorname{tr}(\mathbf{A}) - 2\|\mathbf{A}\|$ and hence the estimator (10) using the shrinkage parameter $\widehat{\tau}_n^*$ satisfies*

$$\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \rho(\boldsymbol{h}, \widehat{\boldsymbol{\beta}}^*) \leq \operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}) - \operatorname{tr}(\mathbf{A})\frac{(1-2/d)^2}{(c+1)}. \tag{35}$$

.

Corollary 1 shows that when $d > 2$ the shrinkage parameter is set according to our recommendation (33), then the shrinkage estimator will asymptotically dominate the MLE.

6 High Dimensional Shrinkage

From equation (35) it appears that the risk bound will simplify as the shrinkage dimension d increases. To investigate this further, we define the normalized asymptotic risk as the ratio of the asymptotic risk of an estimator divided by that of the MLE:

$$\overline{\rho}(\boldsymbol{h}, \boldsymbol{T}) = \frac{\rho(\boldsymbol{h}, \boldsymbol{T})}{\rho(\boldsymbol{h}, \widehat{\boldsymbol{\theta}})} = \frac{\rho(\boldsymbol{h}, \boldsymbol{T})}{\operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}})}. \tag{36}$$

Thus the normalized risk of the MLE is unity ($\overline{\rho}(\boldsymbol{h}, \widehat{\boldsymbol{\beta}}) = 1$) and Theorem 2 shows that the normalized risk of the shrinkage estimator is less than unity under the conditions (28) and (29).

We now investigate the behavior of the normalized risk as the shrinkage dimension d increases. Since $d \leq \min[q, p] \leq m$, then $d \to \infty$ implies $(q, p, m) \to \infty$ as well. As $d \to \infty$, we define

$$a = \lim_{d\to\infty} \frac{\operatorname{tr}(\mathbf{A})}{\operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}})} = \lim_{d\to\infty} \frac{\operatorname{tr}\left(\mathbf{W}\mathbf{G}'\mathbf{V}\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\right)}{\operatorname{tr}(\mathbf{W}\mathbf{G}'\mathbf{V}\mathbf{G})}. \tag{37}$$

The limit a is a measure of the effective number of restrictions relative to the total number of parameters of interest. Note that $0 \leq a \leq 1$, with $a = 1$ when $\mathbf{G}$ is in the range space of $\mathbf{R}$ (all parameters of interest are restricted) which includes full shrinkage. If the weight matrix is the default case $\mathbf{W} = \mathbf{V}_{\boldsymbol{\beta}}^{-1}$ and $\mathbf{R}$ is in the range space of $\mathbf{G}$ (which includes the canonical case) then $a = \lim_d \frac{p}{q}$, the ratio of the number of restrictions to the total number of parameters of interest.

We could restrict attention to shrinkage estimators using the optimal estimator (33), but we allow more broadly for any asymptotically equivalent shrinkage parameter choice. Specially we assume that as $d \to \infty$

$$\frac{\tau}{\tau^*} \longrightarrow 1. \tag{38}$$

Theorem 3 *Under Assumptions 1-3, if as $d \to \infty$, (37) and (38) hold, then for any $0 < c < \infty$,*

$$\limsup_{d\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \overline{\rho}(\boldsymbol{h}, \widehat{\boldsymbol{\beta}}^*) \leq 1 - \frac{a}{c+1}. \tag{39}$$

Equation (39) is a simplified version of (35). 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.

Recall that the asymptotic normalized risk of the MLE is 1. The ideal normalized risk of the restricted estimator (when $c = 0$) is $1 - a$. The risk in (39) 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 MLE obtained by shrinkage towards the restricted estimator.

Equation (39) 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.

In the full shrinkage case, (39) simplifies to

$$\limsup_{d\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \overline{\rho}(\boldsymbol{h}, \widehat{\boldsymbol{\beta}}^*) \leq \frac{c}{c+1}. \tag{40}$$

In general, 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 (13) as $D_n = pF_n$, where F_n is an F-type statistic for (2). Using (22), 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 (2). 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 (39) indicates that substantial efficiency gains can be achieved by shrinkage for a large region of the parameter space.

To illustrate these results numerically, we plot in Figure 1 the asymptotic normalized risk of the MLE $\widehat{\boldsymbol{\beta}}_n$ and our shrinkage estimator $\widehat{\boldsymbol{\beta}}_n^*$, along with the risk bounds, in the full shrinkage ($m = p = d = q$) canonical case. The asymptotic risk is only a function of p and c, and we plot the risk as a function of c for $p = 4$, 8, 12, and 20. The asymptotic normalized risk of the MLE is the upper bound of 1. The asymptotic normalized risk of the shrinkage estimator $\widehat{\boldsymbol{\beta}}_n^*$ is plotted with the solid line[1]. The upper bound (35) is plotted using the short dashes, and the "Large p" lower bound (40) plotted using the long dashes.

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 improvement in the asymptotic risk relative to the MLE decreases as p increases. Furthermore, we can observe that the upper bound (32) is not particularly tight for small p, but improves as p increases. This improvement is a consequence of Theorem 3, which shows that the bound simplifies as p increases. The numerical magnitudes, however, also imply that the risk improvements implied by Theorem 2 are underestimates of the actual improvements in asymptotic risk due to shrinkage. We can see that the large-p bound (40) lies beneath the finite-p bound (35) (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 (40) is reasonably accurate for $p = 20$. Thus the technical approximation $d \to \infty$ seems to be a useful approximation

[1] This is calculated by simulation from the asymptotic distribution using 1,000,000 simulation draws.

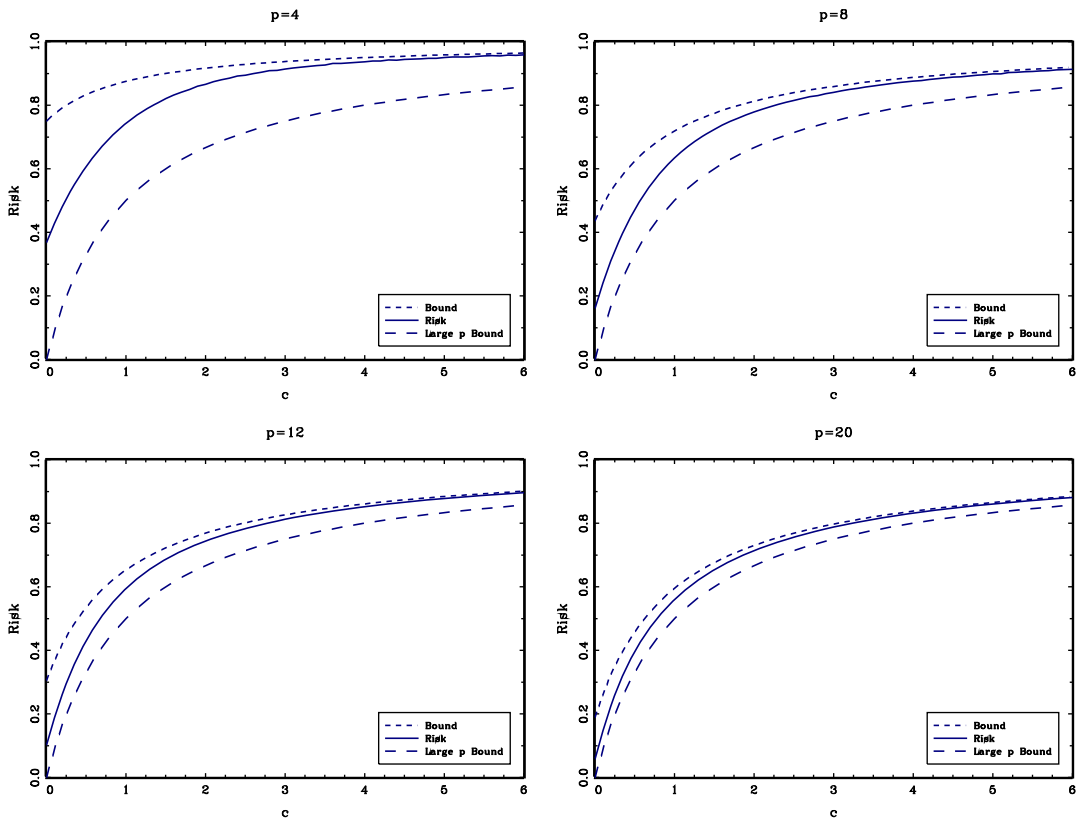

Figure 1: Asymptotic Risk of Shrinkage Estimators

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 an unconstrained model has a large number of parameters yet is not substantially different from a low dimensional model. 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 taking sequential limits (first taking the sample size n to infinity and then taking the dimension d to infinity) is artificial. A deeper result would employ joint limits (taking n and d to infinity jointly). This would be a desirable extension, but would require a different set of tools than those used in this paper. The primary difficulty is that it is unclear how to construct the appropriate limit experiment for a nonparametric estimation problem when simultaneously applying Stein's Lemma to calculate the asymptotic risk. Because of the use of sequential limits, Theorem 3 should not be interpreted as nonparametric.

We are hopeful 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 (39). There are also conceptual similarities

between shrinkage and some penalization methods used in nonparametrics, though in these settings penalization is typically required for regularization (see, e.g. Shen (1997)) rather than for risk reduction. Furtheremore, in nonparametric contexts convergence rates are slower than $n^{-1/2}$, so the asymptotic theory would need to be quite different. These connections are worth exploring.

7 Superefficiency

A seeming paradox for statistical efficiency is posed by the superefficient Hodges' estimator. Our theory provides some insight to dispell the apparant paradox. It is insightful to contrast shrinkage and superefficient estimators, as they have distinct properties. We consider the full shrinkage ($p = q = m$) case with $\boldsymbol{\theta}_0 = \mathbf{0}$ and $\mathbf{W} = \mathbf{V} = \mathbf{I}_p$.

Setting $D_n = n\widehat{\boldsymbol{\theta}}_n'\widehat{\boldsymbol{\theta}}_n$, the Hodges' and full shrinkage estimators can be written as

$$\widehat{\boldsymbol{\theta}}_n^H = \widehat{\boldsymbol{\theta}}_n 1\left(D_n \geq \sqrt{n}\right)$$

and

$$\widehat{\boldsymbol{\theta}}_n^* = \widehat{\boldsymbol{\theta}}_n \left(1 - \frac{p-2}{D_n}\right)_+.$$

There are some nominal similarities. Both $\widehat{\boldsymbol{\theta}}_n^H$ and $\widehat{\boldsymbol{\theta}}_n^*$ shrink $\widehat{\boldsymbol{\theta}}_n$ towards the zero vector, and both estimators equal the zero vector for sufficiently small D_n, but with different thresholds. One difference is that $\widehat{\boldsymbol{\theta}}_n^*$ is a smooth shrinkage function (soft thresholding). Another is that the Hodges' estimator implicitly uses a shrinkage parameter $\tau = \sqrt{n}$ while $\widehat{\boldsymbol{\theta}}_n^*$ sets $\theta = p - 2$.

We now study the asymptotic risk of the Hodges' estimator. Along the sequences (14) with $\boldsymbol{\theta}_0 = \mathbf{0}$, the estimator has the degenerate asymptotic distribution

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n^H - \boldsymbol{\theta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} -\boldsymbol{h}.$$

It follows that its uniform normalized local asymptotic risk for $\mathbf{H}(c) = \left\{\boldsymbol{h} : \boldsymbol{h}'\boldsymbol{h} \leq pc\right\}$ equals

$$\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \overline{\rho}(\boldsymbol{h}, \widehat{\boldsymbol{\theta}}^H) = \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \frac{\boldsymbol{h}'\boldsymbol{h}}{p} = c. \tag{41}$$

Thus the Hodges' estimator has uniformly lower asymptotic risk than the MLE on the sets $\boldsymbol{h} \in \mathbf{H}(c)$ for $c < 1$. The Hodges' estimator has lower risk than the MLE if $c < 1$ but it has higher risk than the MLE if $c > 1$. Thus its superefficiency is very local in nature, and it is globally inefficient. Since the bound (41) diverges to infinity as c increases, the theory shows that the Hodges' estimator is arbitrarily inefficient on arbitrarily large sets.

The difference with the shrinkage estimator is striking. Theorem 2 shows that $\widehat{\boldsymbol{\theta}}_n^*$ has lower asymptotic risk than the MLE for all values of $\boldsymbol{h}$, and the bound (32) holds for all c. In contrast, the asymptotic risk (41) of the Hodges' estimator becomes unbounded as $c \to \infty$, meaning that Hodges' estimator becomes arbitrarily inefficient on arbitrarily large sets.

Furthermore, when $p = d$ is large we can see that the full shrinkage estimator dominates the Hodges' estimator. As shown in (40)

$$\limsup_{d\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \overline{\rho}(\boldsymbol{h}, \widehat{\boldsymbol{\theta}}^*) \leq \frac{c}{c+1} < c = \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \overline{\rho}(\boldsymbol{h}, \widehat{\boldsymbol{\theta}}^H).$$

The shrinkage estimator escapes the Hodges' paradox and dominates the Hodges' estimator.

8 Minimax Risk

We have shown that the 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.

Classical asymptotic minimax theory defines the asymptotic maximum risk of a sequence of estimators $\boldsymbol{T}_n$ for $\boldsymbol{\beta}_n = \mathbf{g}(\boldsymbol{\theta}_n)$ where $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h}$ with arbitrary $\boldsymbol{h}$ as

$$\sup_{I\subset\mathbb{R}^m} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}_{\boldsymbol{\theta}_n}\, n\ell\left(\boldsymbol{\beta}_n, \boldsymbol{T}_n\right) \tag{42}$$

where the first supremum is taken over all finite subsets I of $\mathbb{R}^m$. The asymptotic minimax theorem (e.g. Theorem 8.11 of van der Vaart (1998)) demonstrates that under quite mild regularity conditions the asymptotic uniform risk (42) is bounded below by that of the MLE. This demonstrates that no estimator has smaller asymptotic uniform risk than the MLE over unbounded $\boldsymbol{h}$. This theorem is credited to Hájek (1970, 1972) who built on earlier work of Chernoff (1956) and others. An excellent exposition of this theory is chapter 8 of van der Vaart (1998).

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 $\boldsymbol{T}_n$ as

$$\sup_{I\subset\mathbf{H}(c)} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(\boldsymbol{\beta}_n, \boldsymbol{T}_n\right) \tag{43}$$

which replaces the supremum over all subsets of $\mathbb{R}^m$ with the supremum over all finite subsets of $\mathbf{H}(c)$. Since $\mathbf{H}(c) \subset \mathbb{R}^m$ it follows that (43) will be smaller than (42) and thus a tighter and more informative minimax bound.

On a side note, in the case of full shrinkage ($p = q = m$) then (43) is equivalent to

$$\liminf_{n\to\infty} \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(\boldsymbol{\beta}_n, \boldsymbol{T}_n\right). \tag{44}$$

The reason for the technical difference in the ordering of the supremum between (43) and (44) is that when $p < m$ or $q < m$ the set $\mathbf{H}(c)$ is unbounded is some directions (since the matrix $\mathbf{B}$ does

not have full rank), so the interior supremum must be taken only over a compact set. In the case of full shrinkage the set $\mathbf{H}(c)$ is compact so this complicated ordering of supremum is not necessary.

The standard method to establish the efficiency bound (42) 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 (43) 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., $p = q = m$ 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* $\mathrm{Z} \sim \mathrm{N}_m(\boldsymbol{h}, \mathbf{V})$. *Consider any estimator* $\boldsymbol{T} = \boldsymbol{T}(\mathrm{Z})$ *of* $\mathbf{G}'\boldsymbol{h}$. *For* $\mathbf{H}(c)$ *defined in (31) and any* $0 < c < \infty$, *for* d *sufficiently large, where* d *is defined in (27),*

$$\sup_{\boldsymbol{h} \in \mathbf{H}(c)} \mathbb{E}_{\boldsymbol{h}} \left(\boldsymbol{T} - \mathbf{G}'\boldsymbol{h} \right)' \mathbf{W} \left(\boldsymbol{T} - \mathbf{G}'\boldsymbol{h} \right) \geq \operatorname{tr}\left(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}} \right) - \left(\frac{1}{1+c} + \frac{4c}{(1+c)^{2/3}} d^{-1/3} \right) \operatorname{tr}\left(\mathbf{A} \right). \tag{45}$$

This is a finite sample lower bound on the weighted quadratic risk for the normal sampling model. Typically in this literature this bound is expressed for the high-dimensional (large d) case. We define the normalized loss as

$$\frac{\left(\boldsymbol{T} - \mathbf{G}'\boldsymbol{h} \right)' \mathbf{W} \left(\boldsymbol{T} - \mathbf{G}'\boldsymbol{h} \right)}{\operatorname{tr}\left(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}} \right)}$$

Then, taking the limit as $d \to \infty$ as in Theorem 3, the normalized version of the bound (45) simplifies to

$$\liminf_{d \to \infty} \sup_{\boldsymbol{h} \in \mathbf{H}(c)} \mathbb{E}_{\boldsymbol{h}} \frac{\left(\boldsymbol{T} - \mathbf{G}'\boldsymbol{h} \right)' \mathbf{W} \left(\boldsymbol{T} - \mathbf{G}'\boldsymbol{h} \right)}{\operatorname{tr}\left(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}} \right)} \geq 1 - \frac{a}{c+1}. \tag{46}$$

We do not use (46) directly, but rather use (45) 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.

Classic minimax theory (e.g. Theorem 8.11 of van der Vaart (1998)) applies to all bowl-shaped (subconvex) loss functions, not just quadratic loss, and thus it seems reasonable to conjecture that Theorem 4 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.

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. Unlike classical minimax theory, we do not restrict the loss function to be bowl-shaped. Instead we require the loss function to the locally quadratic (Assumption 2) and additionally impose the

technical requirement that it is uniquely minimized.

Assumption 4 *For all $\varepsilon > 0$, there is a $c_\varepsilon > 0$ such that $\inf_{|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2| \geq \varepsilon} \ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \geq c_\varepsilon$.*

Define the normalized loss function $\overline{\ell}(\boldsymbol{\beta}_n, \boldsymbol{T}_n) = \ell(\boldsymbol{\beta}_n, \boldsymbol{T}_n) / \operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}})$.

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*

$$\int \left[f(x, \boldsymbol{\theta} + \boldsymbol{h})^{1/2} - f(x, \boldsymbol{\theta})^{1/2} - \frac{1}{2} \boldsymbol{h}' f_{\boldsymbol{\theta}}(x) f(x)^{1/2} \right]^2 d\mu = o\left(\|\boldsymbol{h}\|^2 \right), \qquad \boldsymbol{h} \to 0 \tag{47}$$

where $f_{\boldsymbol{\theta}}(x) = \dfrac{\partial}{\partial \boldsymbol{\theta}} \log f(x, \boldsymbol{\theta})$. Suppose that $\mathcal{I}_{\boldsymbol{\theta}} = \mathbb{E} f_{\boldsymbol{\theta}}(X_i, \boldsymbol{\theta}) f_{\boldsymbol{\theta}}(X_i, \boldsymbol{\theta})' > 0$, set $\mathbf{V} = \mathcal{I}_{\boldsymbol{\theta}}^{-1}$, $\mathbf{A}$ as in (26), and $\mathbf{H}(c)$ as in (31). Suppose that the loss function $\ell(\boldsymbol{\beta}, \boldsymbol{T})$ satisfies Assumptions 2 and 4. Then for any sequence of estimators $\boldsymbol{T}_n$ for $\boldsymbol{\beta}_n = \mathbf{g}(\boldsymbol{\theta}_n)$ on the sequence $\boldsymbol{\theta}_n = \boldsymbol{\theta} + n^{-1/2}\boldsymbol{h}$, where $\boldsymbol{\theta}$ is in the interior of $\boldsymbol{\Theta}$, and any $0 < c < \infty$, and for d sufficiently large (defined in (27))

$$\sup_{I \subset \mathbf{H}(c)} \liminf_{n \to \infty} \sup_{\boldsymbol{h} \in I} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell(\boldsymbol{\beta}_n, \boldsymbol{T}_n) \geq \operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}) - \left[\frac{1}{1+c} + \frac{4c}{(1+c)^{2/3}} d^{-1/3} \right] \operatorname{tr}(\mathbf{A}). \tag{48}$$

Furthermore, suppose that as $d \to \infty$, (37) holds. Then

$$\liminf_{d \to \infty} \sup_{I \subset \mathbf{H}(c)} \liminf_{n \to \infty} \sup_{\boldsymbol{h} \in I} \mathbb{E}_{\boldsymbol{\theta}_n} n\overline{\ell}(\boldsymbol{\beta}_n, \boldsymbol{T}_n) \geq 1 - \frac{a}{c+1}. \tag{49}$$

Theorem 5 provides a lower bound on the asymptotic local minimax risk for $\boldsymbol{h}$ in the ball $\mathbf{H}(c)$. (48) is the case of finite d, and (49) shows that the bound takes a simple form when d is large. To our knowledge, Theorem 5 is new. It is the first large sample local efficiency bound for shrinkage estimation.

The bounds in (48) and (49) are the lowest possible asymptotic risks, uniformly for parameter values in the local regions $\mathbf{H}(c)$. The bounds are a function of c, which controls the size of the regions $\mathbf{H}(c)$. The bounds are increasing in c and asymptote as $c \to \infty$ to the classic minimax bounds. Equivalently, the bounds decrease as c tends to 0, meaning that the lower estimation efficiency can be improved on local regions of the parameter space.

The regularity conditions for Theorem 5 are quite weak. The main restriction on the probability framework is differentiability in quadratic mean (47), which is weaker than the requirements for asymptotic normality of the MLE. (See chapter 7 of van der Vaart (1998).) Assumption 2 restricts attention to locally quadratic loss functions (excluding, for example, absolute loss) but does not require the loss function to be subconvex (bowl-shaped) as is typical in existing minimax theory. The reason is that our theorem uses Pinsker's theorem rather than Anderson's Lemma (e.g. Lemma 8.5 of van der Vaart, 1998) and locally quadratic loss functions.

Similarly to Theorem 3, a limitation of the bound (49) is the use of the sequential limits, first taking n to infinity and then d to infinity. A deeper result would employ joint limits.

Equation (49) of Theorem 5 gives a lower bound for the local minimax risk of any estimator over the local parameter space $\mathbf{H}(c)$. The lower bound depends on the radius c of this local space. We call an estimator locally asymptotically minimax efficient if its local asymptotic risk is identical for all regions $\mathbf{H}(c)$.

Since the upper bound from Theorem 3 equals the lower bound from Theorem 5 for all values of c, this means that our shrinkage estimator $\widehat{\boldsymbol{\beta}}_n^*$ is locally asymptotically minimax efficient. In contrast, the MLE is not locally minimax efficient since its local asymptotic risk does not achieve the efficiency bound.

A global minimax bound is obtained by replacing $\mathbf{H}(c)$ with $\mathbb{R}^q$, or equivalently taking the limit as $c \to \infty$. From (49) we see that the global minimax bound is 1. This is the lowest possible minimax risk, globally in the parameter space. We call an estimator globally asymptotically minimax efficient if its global uniform risk if it achieves the global minimax bound. For example, both $\widehat{\boldsymbol{\beta}}_n^*$ and the MLE are globally minimax efficient.

9 Simulation

9.1 Binary Probit

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

$$\begin{aligned} y_i &= 1\left(y_i^* \geq 0\right) \\ y_i^* &= \theta_0 + X_{1i}'\boldsymbol{\theta}_1 + X_{2i}'\boldsymbol{\theta}_2 + e_i \\ e_i &\sim N(0,1). \end{aligned}$$

The regressors X_{i1} and X_{2i} are $k \times 1$ and $p \times 1$, respectively, with $p > k$. The regressor vector X_i is distributed $\mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$ where $\boldsymbol{\Sigma}_{jj} = 1$ and $\boldsymbol{\Sigma}_{jk} = \rho$ for $j \neq k$. The goal is estimation of marginal effects under quadratic loss and the belief that $\boldsymbol{\theta}_2$ may be close to the zero vector.

The regression coefficients are set as $\theta_0 = 0$, $\boldsymbol{\theta}_1 = (b, b, ..., b)'$ and $\boldsymbol{\theta}_2 = (c, c, ..., c)'$. 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 $\rho = 0.5$, and report results for different values of c, n, and p. We also experimented with the alternative specification $\boldsymbol{\theta}_2 = (c, 0, ..., 0)'$ (only one omitted regressor important) and the results were virtually identical so are not reported.

The estimators will be functions of the following primary components:

1. $\widehat{\boldsymbol{\theta}}$ = unrestricted MLE. Probit of y_i on $(1, X_{1i}, X_{2i})$
2. $\widetilde{\boldsymbol{\theta}}$ = restricted MLE. Probit of y_i on $(1, X_{1i})$
3. $LR_n = 2\left(\log \mathcal{L}(\widehat{\boldsymbol{\theta}}) - \log \mathcal{L}(\widetilde{\boldsymbol{\theta}})\right)$, the likelihood ratio test for the restriction $\boldsymbol{\theta}_2 = 0$

4. $\widehat{\mathbf{V}}_n$ = estimate of the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$

The restricted estimator is selected to reflect the belief that the coefficients on X_{2i} may be close to zero. Hence $\mathbf{R} = (\mathbf{0}_{k+1}, \mathbf{I}_p)'$.

Our parameter of interest is the vector of marginal effects of the regressors on the choice probabilities evaluated at the population means

$$\boldsymbol{\beta} = \mathbf{g}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\theta}_1 \phi(\theta_0) \\ \boldsymbol{\theta}_2 \phi(\theta_0) \end{pmatrix}.$$

The associated derivative matrix is

$$\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{g}(\boldsymbol{\theta})' = \begin{bmatrix} -\theta_0 \phi(\theta_0) \boldsymbol{\theta}_1' & -\theta_0 \phi(\theta_0) \boldsymbol{\theta}_2' \\ \mathbf{I}_k \phi(\theta_0) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \phi(\theta_0) \end{bmatrix}.$$

The unrestricted and restricted estimates of the marginal effects are $\widehat{\boldsymbol{\beta}} = \mathbf{g}(\widehat{\boldsymbol{\theta}})$ and $\widetilde{\boldsymbol{\beta}} = \mathbf{g}(\widetilde{\boldsymbol{\theta}})$.

We compare four conventional estimators and four shrinkage estimators of the marginal effects $\boldsymbol{\beta}$. The first is $\widehat{\boldsymbol{\beta}}$, the unrestricted MLE. The second is the pretest estimator

$$\widehat{\boldsymbol{\beta}}_{PT} = \begin{cases} \widehat{\boldsymbol{\beta}} & \text{if} \quad LR_n \geq q \\ \widetilde{\boldsymbol{\beta}} & \text{if} \quad LR_n < q \end{cases}$$

where q is the 95% quantile of the χ^2_p distribution.

The third estimator is the weighted AIC estimator $\widehat{\boldsymbol{\beta}}_{WAIC}$ of Burnham and Anderson (2002), which is

$$\widehat{\boldsymbol{\beta}}^* = \widehat{w}_n \widehat{\boldsymbol{\beta}} + (1 - \widehat{w}_n) \widetilde{\boldsymbol{\beta}} \tag{50}$$

with the weight

$$\widehat{w}_n = \left(1 + \exp\left(p - LR_n/2\right)\right)^{-1}.$$

The motivation is that the weight for each estimator is proportional to the exponential of negative one-half of each model's AIC.

The fourth estimator is an approximate Bayesian model averaging (BMA) estimator $\widehat{\boldsymbol{\beta}}_{BMA}$, which takes the same form but with the weight

$$\widehat{w}_n = \left(1 + \exp\left(p \log(n)/2 - LR_n/2\right)\right)^{-1}.$$

This is equivalent to making the weight for each estimator proportional to the exponential of negative one-half of each model's BIC. The motivation is that this is approximately the Bayes estimator when each model has equal prior probability and the probit parameters have diffuse priors.

We consider four shrinkage estimators contructed using distinct loss functions. These all take

the form (50) but with different weights $\widehat{w}_n$. Our first estimator is the default shrinkage estimator, which sets

$$\widehat{w}_n = \left(1 - \frac{p-2}{n\left(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\right)' \widehat{\mathbf{V}}_n^{-1}\left(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\right)}\right)_+ .$$

This estimator is constructed assuming the parameter of interest is the entire coefficient vector $\boldsymbol{\theta}$ and is evaluated using quadratic loss with weight matrix $\mathbf{W} = \widehat{\mathbf{V}}_n^{-1}$.

The second shrinkage estimator sets

$$\begin{aligned} \widehat{w}_n &= \left(1 - \frac{\operatorname{tr}(\widehat{\mathbf{A}}) - 2\left\|\widehat{\mathbf{A}}\right\|}{n\left(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\right)'\left(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\right)}\right)_+ \\ \widehat{\mathbf{A}} &= \widehat{\mathbf{V}}_n \mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{V}}_n\mathbf{R}\right)^{-1}\mathbf{R}'\widehat{\mathbf{V}}_n. \end{aligned} \tag{51}$$

This estimator is constructed assuming the parameter of interest is the entire coefficient vector $\boldsymbol{\theta}$ and is evaluated using unweighted quadratic loss.

The third shrinkage estimator sets

$$\begin{aligned} \widehat{w}_n &= \left(1 - \frac{\operatorname{tr}(\widehat{\mathbf{A}}) - 2\left\|\widehat{\mathbf{A}}\right\|}{n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right)'\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right)}\right)_+ \\ \widehat{\mathbf{A}} &= \widehat{\mathbf{G}}'\widehat{\mathbf{V}}_n \mathbf{R}\left(\mathbf{R}'\widehat{\mathbf{V}}_n\mathbf{R}\right)^{-1}\mathbf{R}'\widehat{\mathbf{V}}_n\widehat{\mathbf{G}}' \end{aligned}$$

with

$$\widehat{\mathbf{G}} = \begin{bmatrix} -\widehat{\theta}_0\phi(\widehat{\theta}_0)\widehat{\theta}_1' & -\widehat{\theta}_0\phi(\widehat{\theta}_0)\widehat{\boldsymbol{\theta}}_2' \\ \mathbf{I}_k\phi(\widehat{\theta}_0) & \mathbf{0}_p \\ \mathbf{0}_k & \mathbf{I}_p\phi(\widehat{\theta}_0) \end{bmatrix}.$$

This estimator is constructed assuming the parameter of interest is the vector of partial effects and is evaluated using unweighted quadratic loss.

The fourth shrinkage estimator is similar to the third, but is constructed assuming the parameter of interest is the vector of changes in the choice probability by increasing a single regressor from its mean by two standard deviations

$$\boldsymbol{\pi} = \begin{pmatrix} \Phi(2\boldsymbol{\theta}_1 + \theta_0) - \Phi(\theta_0) \\ \Phi(2\boldsymbol{\theta}_2 + \theta_0) - \Phi(\theta_0) \end{pmatrix},$$

and is evaluated using unweighted quadratic loss.

Our primary purpose is to compare the MLE with the default shrinkage estimator. The other estimators are included for comparison purposes.

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 $\widehat{\boldsymbol{\theta}}$ was unreliable. For these cases we simply set all estimates equal to the restricted estimates. This would seem to correspond to empirical practice, so should not bias our results as all estimators were treated symmetrically.

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

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

Thus our evaluation is based on the assumption that the parameter of interest is $\boldsymbol{\beta}$ and is evaluated using unweighted quadratic loss.

We calculated the MSE by simulation using 10,000 replications. To simplify the presentation, we normalize the MSE of the estimators by that of the unconstrained MLE. Thus an MSE less than one has lower risk than the MLE, and an MSE exceeding one has higher risk.

Our calculations revealed that in these experiments, the four shrinkage estimators had very similar MSE. This shows that in at least this application, their performance is robust to the specification of the loss function. Since their results are quite similar, we only report the MSE for the default estimator which uses the weights (51).

In Figure 2, we display the results for $n = \{200, 500\}$ and $p = \{4, 8\}$, and vary c on a 50-point grid from 0 to 0.20. The normalized MSE are displayed as lines. The solid line is the normalized MSE of the (default) shrinkage estimator, the short dashed line is the normalized MSE of the pretest estimator, the longer dashed line is the normalized MSE of the weighted AIC estimator, the dashed-dot line is the normalized MSE of the BMA estimator, and the dotted line is 1, the normalized MSE of the unrestricted MLE. Once again, the normalized MSE of the other three shrinkage estimators are omitted as they are nearly identical with that of the default shrinkage estimator

Figure 2 shows convincingly that the shrinkage estimator significantly dominates the MLE. Its finite-sample MSE is less than that of the MLE for all parameter values, and in some cases its MSE is a small fraction. (For robustness, similar calculations were made for values of c up to 0.5, and this result is uniform over this region.)

It is also constructive to compare the shrinkage estimator with the other estimators. No other estimator uniformly dominates the MLE. The other estimators have lower MSE for values of c near zero, but their MSE exceeds that of the MLE for moderate values of c. The most sensitive estimator is BMA, which has quite low MSE for very small c, but very large MSE for moderate c. The poor MSE performance of the pretest and BMA estimators is a well-documented property of such estimators, but is worth repeating here as both pretests and BMA are routinely used in applied research. The numerical calculations shown in Figure 2 show that a much lower MSE estimator is obtained by shrinkage.

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

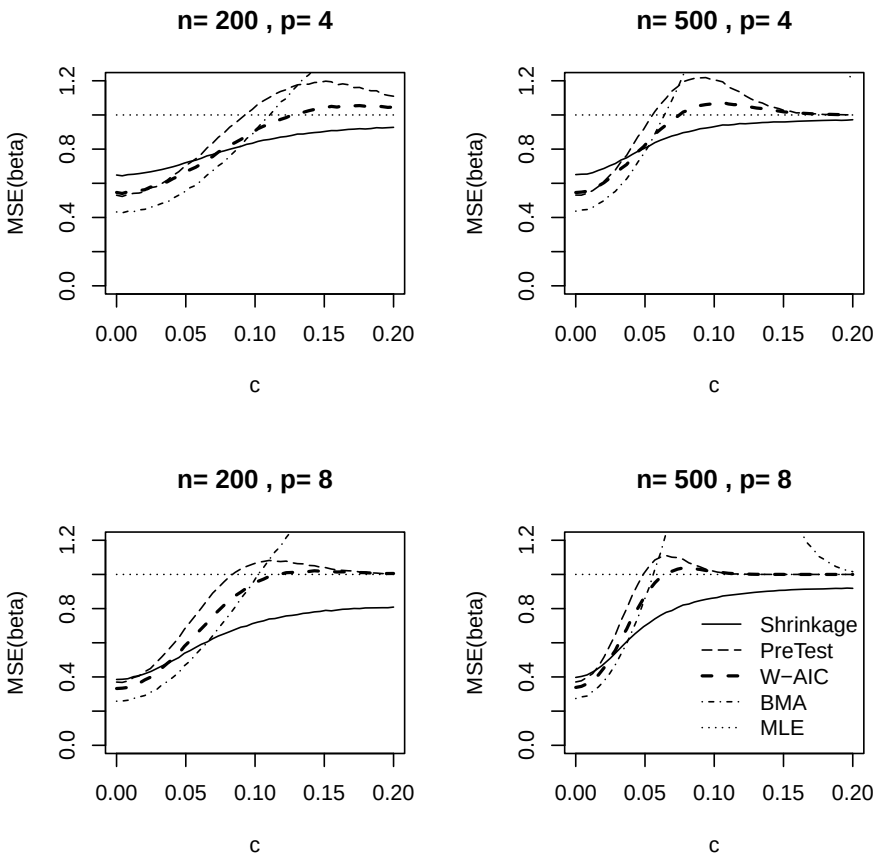

Figure 2: Finite Sample Mean Squared Error, Probit Design

improves asymptotic risk, and the improvements can be especially strong in high-dimensional cases.

9.2 Exponential Means

Our second simulation experiment is a simple model of exponential means. This example was selected for its simplicity, yet is a context where the log-likelihood is meaningfully different from quadratic in small samples.

Let y_{ji}, $i = 1, ..., n$, be independent draws from the exponential distribution with mean θ_j, for $j = 0, ..., p$. We consider estimation of $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_p)'$ under quadratic loss with the belief that the coefficients may be close to one another, that is, the shrinkage direction is $\theta_0 = \theta_1 = \cdots = \theta_p$. Thus the matrix $\mathbf{R}$ takes the form (5).

We are interested in the small sample case, so consider $n = 10$ and 40. We are also interested in the "large p" case so consider $p = 10$ and 40. We set the coefficients according to the law $\theta_j = 1 + cj/p$, so the coefficients are uniformly distributed in the interval $[1, 1 + c]$. Thus for small values of c, the coefficients are close to one another, and for large values of c the coefficients are less close. We then vary c on a 50-point grid in the interval $[0, 2]$.

In this setting, the unrestricted MLE is the subgroup means, that is, $\widehat{\theta}_j = \dfrac{1}{n}\sum_{i=1}^{n} y_{ji}$. The associated log-likelihood is $\log \mathcal{L}(\widehat{\boldsymbol{\theta}}) = -n\sum_{j=0}^{p} \log \widehat{\theta}_j - n(p+1)$, and the estimate of the asymptotic covariance matrix is $\widehat{\mathbf{V}} = \text{diag}\left\{\widehat{\theta}_0, ..., \widehat{\theta}_p^2\right\}$.

The restricted MLE is $\widetilde{\theta} = (p+1)^{-1}\sum_{j=0}^{p} \widehat{\theta}_j$ with associated log-likelihood $\log \mathcal{L}(\widetilde{\theta}) = -n(p+1)\left(\log \widetilde{\theta} + 1\right)$.

We compare the same conventional estimators as the previous simulation experiment: unrestricted MLE, pretest, weighted AIC, and BMA. For the shrinkage estimator we treat the parameter of interest as the entire coefficient vector θ and use unweighted squared error loss. Thus the weights take the form (51) with $\widehat{\mathbf{V}}_n = diag\{\widehat{\theta}_0^2, \widehat{\theta}_1^2, ...\widehat{\theta}_p^2\}$.

Again, we calculate the MSE of the estimators by simulation using 10,000 simulation replications, and normalize the MSE of each estimator by the MSE of the unrestricted MLE. We display the results in Figure 3 for $n = \{10, 40\}$ and $p = \{10, 40\}$, graphed as a function of c.

The results are similar to those from the previous example. The shrinkage estimator uniformly dominates the MLE. In some cases (especially the smaller sample size) the reduction in risk is quite substantial. The shrinkage estimator is the only estimator which uniformly dominates the MLE, and the shrinkage estimator has lower MSE than the other estimators for more values of c. The BMA estimator has the lowest MSE for the smallest values of c, but again has very high MSE for intermediate values.

10 Conclusion

This paper has shown how to improve upon nonlinear MLE in quite general contexts by shrinkage. Implementation requires a choice of weight matrix and shrinkage direction. Asymptotically, the shrinkage estimator uniformly has lower risk than the MLE and achieves a local minimax efficiency bound.

Technically, an important contribution is the extension of Pinsker-type high-dimensional local minimax theory to an asymptotic setting. An important caveat, however, is that the theory relies on sequential asymptotic limits (first sample size n and then shrinkage dimension p diverge) which limits the "high-dimensional" interpretation. Hopefully this limitation can be removed in future work.

The results in this paper are also confined to parametric (likelihood) models. The shrinkage estimators and asymptotic MSE results are straightforward to extend to semiparametric estimators, but the extension of the efficiency theory appears to be quite challenging. This would be an interesting topic for future research.

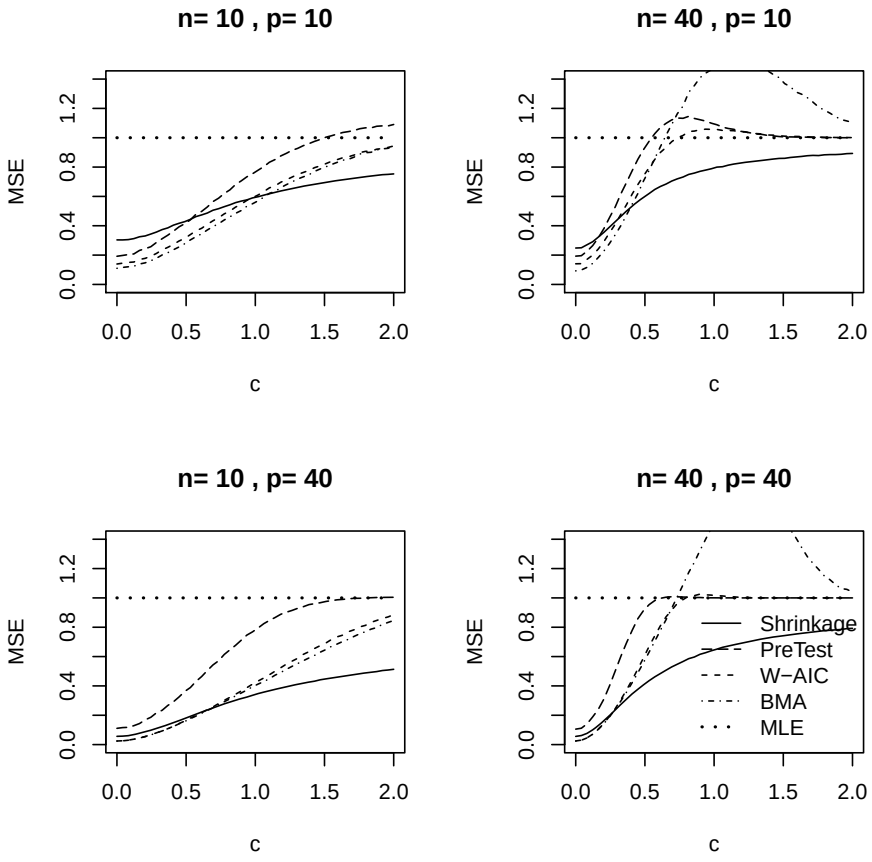

Figure 3: Finite Sample Mean Squared Error, Exponential Means Design

11 Appendix

Proof of Theorem 1: (16) is Theorem 3.3 of Newey and McFadden (1994). (17) follows by standard arguments, see for example, the derivation in Section 9.1 of Newey and McFadden (1994). Under (16), (17), and Assumption 1.8, we can apply the delta method to find that for some $\boldsymbol{\theta}_n^* \to_p \boldsymbol{\theta}_0$,

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right) &= \sqrt{n}\left(\mathbf{g}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{g}\left(\boldsymbol{\theta}_n\right)\right) \\ &= \boldsymbol{G}\left(\boldsymbol{\theta}_n^*\right)' \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) \\ &\xrightarrow[\boldsymbol{\theta}_n]{} \mathbf{G}'\mathrm{Z}\end{aligned}$$

and similarly

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right) &= \boldsymbol{G}\left(\boldsymbol{\theta}_n^*\right)' \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) \\ &\xrightarrow[\boldsymbol{\theta}_n]{} \mathbf{G}'\left(\left(\mathrm{Z} + \boldsymbol{h}\right) - \mathbf{VR}\left(\mathbf{R}'\mathbf{VR}\right)^{-1}\mathbf{R}'\left(\mathrm{Z} + \boldsymbol{h}\right)\right),\end{aligned}$$

implying

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} -\mathbf{G}'\mathbf{VR}\left(\mathbf{R}'\mathbf{VR}\right)^{-1}\mathbf{R}'\left(\mathrm{Z} + \boldsymbol{h}\right). \tag{52}$$

Taking a second order Taylor expansion around $\widehat{\boldsymbol{\beta}}_n$

$$\begin{aligned}n\ell\left(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n\right) = n\ell\left(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\beta}}_n\right) &+ n\,\frac{\partial}{\partial \boldsymbol{\beta}}\ell\left(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}\right)\Big|'_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_n}\left(\widetilde{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\beta}}_n\right) \\ &+ n\left(\widetilde{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\beta}}_n\right)'\mathbf{W}\left(\boldsymbol{\beta}_n^*\right)\left(\widetilde{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\beta}}_n\right)\end{aligned}$$

where $\boldsymbol{\beta}_n^*$ lies on a line segment joining $\widetilde{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\beta}}_n$. By Assumption 2.1, $\ell\left(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\beta}}_n\right) = 0$. By Assumption 2.3, the fact that $\ell\left(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}\right)$ is minimized at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_n$, and the differentiability implied by Assumption 2.3, then $\frac{\partial}{\partial \boldsymbol{\beta}}\ell\left(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}\right)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_n} = 0$. Under Assumption 2.3 and the consistency of $\widetilde{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\beta}}_n$, it follows that $\mathbf{W}\left(\boldsymbol{\beta}_n^*\right) \to_p \mathbf{W}$. Combined with (52) we find

$$\begin{aligned}n\ell\left(\widehat{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\beta}}_n\right) &= n\left(\widetilde{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\beta}}_n\right)'\mathbf{W}\left(\boldsymbol{\beta}_n^*\right)\left(\widetilde{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\beta}}_n\right) \\ &\xrightarrow[\boldsymbol{\theta}_n]{} \left(\mathrm{Z} + \boldsymbol{h}\right)'\mathbf{R}\left(\mathbf{R}'\mathbf{VR}\right)^{-1}\mathbf{R}'\mathbf{VGWG}'\mathbf{VR}\left(\mathbf{R}'\mathbf{VR}\right)^{-1}\mathbf{R}'\left(\mathrm{Z} + \boldsymbol{h}\right) \\ &= \xi\end{aligned}$$

which is (18).

(19) and (20) follow by the continuous mapping theorem and Assumption 3. ■

Proof of Lemma 1: By Assumption 2.1, $\ell\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_n\right) = 0$. Assumption 2 implies that $\ell\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}\right)$ is

minimized and differentiable at $\boldsymbol{\beta} = \boldsymbol{\beta}_n$, so the first-order condition is $\frac{\partial}{\partial \boldsymbol{\beta}} \ell\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n} = 0$. Then by a second-order Taylor series expansion of $\ell\left(\boldsymbol{\beta}_n, \boldsymbol{T}_n\right)$ about $\boldsymbol{\beta}_n$

$$\begin{aligned} n\ell\left(\boldsymbol{\beta}_n, \boldsymbol{T}_n\right) &= n\ell\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_n\right) + n\frac{\partial}{\partial \boldsymbol{\beta}} \ell\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_n\right)' \left(\boldsymbol{T}_n - \boldsymbol{\beta}_n\right) + n\left(\boldsymbol{T}_n - \boldsymbol{\beta}_n\right)' \mathbf{W}\left(\boldsymbol{\beta}_n^*\right)\left(\boldsymbol{T}_n - \boldsymbol{\beta}_n\right) \\ &= n\left(\boldsymbol{T}_n - \boldsymbol{\beta}_n\right)' \mathbf{W}\left(\boldsymbol{\beta}_n^*\right)\left(\boldsymbol{T}_n - \boldsymbol{\beta}_n\right) \end{aligned} \tag{53}$$

for some $\boldsymbol{\beta}_n^*$ on a line segment joining $\boldsymbol{T}_n$ and $\boldsymbol{\beta}_n$. (24) and $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$ imply $\boldsymbol{T}_n \to_p \boldsymbol{\beta}_0$ and hence $\boldsymbol{\beta}_n^* \to_p \boldsymbol{\beta}_0$. By the continuity of Assumption 2.3, it follows that $\mathbf{W}\left(\boldsymbol{\beta}_n^*\right) \to_p \mathbf{W}\left(\boldsymbol{\beta}_0\right) = \mathbf{W}$. Combined with (24) we find

$$n\ell\left(\boldsymbol{\beta}_n, \boldsymbol{T}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \psi' \mathbf{W} \psi.$$

As shown by Lemma 6.1.14 of Lehmann and Casella (1998) this can be used to establish that

$$\rho(\boldsymbol{h}, \boldsymbol{T}) = \lim_{\zeta \to \infty} \liminf_{n \to \infty} \mathbb{E}_{\boldsymbol{\theta}_n} \min\left[n\ell\left(\boldsymbol{\beta}_n, \boldsymbol{T}_n\right), \zeta\right] = \mathbb{E}\left(\psi' \mathbf{W} \psi\right)$$

which is (25). ■

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

Lemma 2 *If* $\mathrm{Z} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{V}\right)$ *is* $m \times 1$, $\mathbf{K}$ *is* $m \times m$, *and* $\eta\left(\mathbf{x}\right) : \mathbb{R}^m \to \mathbb{R}^m$ *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 $\mathrm{N}\left(\mathbf{0}, \mathbf{V}\right)$ density function. By multivariate integration by parts

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

■

Proof of Theorem 2: Observe that $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \mathbf{G}'\mathrm{Z}$ with $\mathrm{Z} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{V}\right)$ under (16). Then Lemma 1 shows that

$$\rho(\boldsymbol{h}, \widehat{\boldsymbol{\beta}}) = \mathbb{E}\left(\mathrm{Z}'\mathbf{G}\mathbf{W}\mathbf{G}'\mathrm{Z}\right) = \mathrm{tr}\left(\mathbf{G}\mathbf{W}\mathbf{G}'\mathbb{E}\left(\mathrm{Z}\mathrm{Z}'\right)\right) = \mathrm{tr}\left(\mathbf{W}\mathbf{G}'\mathbf{V}\mathbf{G}\right) = \mathrm{tr}\left(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}\right). \tag{54}$$

Next, $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_n\right) \xrightarrow[\boldsymbol{\theta}_n]{} \psi$, where ψ is the random variable shown in (20). The variable ψ has a classic James-Stein distribution with positive-part trimming. Define the analogous random

variable without positive part trimming

$$\psi^* = \mathbf{G}'\mathrm{Z} - \left(\frac{\tau}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right)\mathbf{G}'\mathbf{V}\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'(\mathrm{Z}+\boldsymbol{h}). \tag{55}$$

Then using Lemma 1 and the fact that the pointwise quadratic risk of ψ is strictly smaller than that of ψ^* (as shown, for example, by Lemma 2 of Hansen (2015)),

$$\rho(\boldsymbol{h}, \widehat{\boldsymbol{\beta}}^*) = \mathbb{E}\left(\psi'\mathbf{W}\psi\right) < \mathbb{E}\left(\psi^{*\prime}\mathbf{W}\psi^*\right). \tag{56}$$

Using (55), we calculate that (56) equals

$$\begin{aligned}
&\mathbb{E}(\mathrm{Z}'\mathbf{G}\mathbf{W}\mathbf{G}'\mathrm{Z}) \\
&+ \tau^2\mathbb{E}\left(\frac{(\mathrm{Z}+\boldsymbol{h})'\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}'\mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'(\mathrm{Z}+\boldsymbol{h})}{\left((\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})\right)^2}\right) \\
&- 2\tau\mathbb{E}\left(\frac{(\mathrm{Z}+\boldsymbol{h})'\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}'\mathrm{Z}}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right) \\
&= \mathrm{tr}\left(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}\right) + \tau^2\mathbb{E}\left(\frac{1}{(\mathrm{Z}+\boldsymbol{h})'\mathbf{B}(\mathrm{Z}+\boldsymbol{h})}\right) \\
&- 2\tau\mathbb{E}\left(\eta(\mathrm{Z}+\boldsymbol{h})'\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}'\mathrm{Z}\right)
\end{aligned} \tag{57}$$

where

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

We calculate that

$$\frac{\partial}{\partial\mathbf{x}}\eta(\mathbf{x})' = \left(\frac{1}{\mathbf{x}'\mathbf{B}\mathbf{x}}\right)\mathbf{I} - \frac{2}{(\mathbf{x}'\mathbf{B}\mathbf{x})^2}\mathbf{B}\mathbf{x}\mathbf{x}'. \tag{58}$$

Using Lemma 2 (Stein's Lemma) and (58)

$$\begin{aligned}
\mathbb{E}\left(\eta(\mathrm{Z}+\boldsymbol{h})'\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}'\mathrm{Z}\right) &= \mathbb{E}\,\mathrm{tr}\left(\frac{\partial}{\partial \mathbf{x}}\eta\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}'\mathbf{V}\right) \\
&= \mathbb{E}\,\mathrm{tr}\left(\frac{\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}'\mathbf{V}}{\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)}\right) \\
&- 2\mathbb{E}\,\mathrm{tr}\left(\frac{\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}'\mathbf{V}}{\left(\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)\right)^{2}}\right) \\
&= \mathbb{E}\left(\frac{\mathrm{tr}\left(\mathbf{A}\right)}{\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)}\right) \\
&- 2\mathbb{E}\left(\frac{\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}_1'\mathbf{A}^*\mathbf{B}_1\left(\mathrm{Z}+\boldsymbol{h}\right)}{\left(\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)\right)^{2}}\right) \\
&\geq \mathbb{E}\left(\frac{\mathrm{tr}\left(\mathbf{A}\right)-2\left\|\mathbf{A}\right\|}{\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)}\right) \qquad (59)
\end{aligned}$$

where we have defined $\mathbf{B}_1 = \mathbf{W}^{1/2\prime}\mathbf{G}'\mathbf{V}\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'$. In (59), the third equality uses

$$\mathrm{tr}\left(\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}'\mathbf{V}\right) = \mathrm{tr}\left(\mathbf{A}\right)$$

and

$$\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{V}\mathbf{G}\mathbf{W}\mathbf{G}'\mathbf{V}\mathbf{B} = \mathbf{B}_1'\mathbf{A}\mathbf{B}_1.$$

The final inequality is

$$\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}_1'\mathbf{A}\mathbf{B}_1\left(\mathrm{Z}+\boldsymbol{h}\right) \leq \left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)\left\|\mathbf{A}\right\|$$

since $\mathbf{B}_1'\mathbf{B}_1 = \mathbf{B}$.

(59) implies that (57) is smaller than

$$\begin{aligned}
&\mathrm{tr}\left(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}\right) - \tau\mathbb{E}\left(\frac{2\left(\mathrm{tr}\left(\mathbf{A}\right)-2\left\|\mathbf{A}\right\|\right)-\tau}{\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)}\right) \\
&\leq \mathrm{tr}\left(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}\right) - \tau\frac{2\left(\mathrm{tr}\left(\mathbf{A}\right)-2\left\|\mathbf{A}\right\|\right)-\tau}{\mathbb{E}\left(\left(\mathrm{Z}+\boldsymbol{h}\right)'\mathbf{B}\left(\mathrm{Z}+\boldsymbol{h}\right)\right)} \qquad (60)
\end{aligned}$$

where the inequality is Jensen's and uses the assumption that $\tau \leq 2\left(\mathrm{tr}\left(\mathbf{A}\right)-2\left\|\mathbf{A}\right\|\right)$.

We calculate that

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

where the inequality is for $\boldsymbol{h} \in \mathbf{H}(c)$. Substituted into (60) we have established (32). As this bound

is strictly less than $\operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}})$ for any $c < \infty$, combined with (54) we have established (30). ■

Proof of Corollary 1. By the consistency of the parameter estimates and the continuity of the functions,

$$\widehat{\mathbf{A}}_n = \widehat{\mathbf{W}}_n^{1/2\prime}\widehat{\mathbf{G}}_n'\widehat{\mathbf{V}}_n\widehat{\mathbf{R}}_n\left(\widehat{\mathbf{R}}_n'\widehat{\mathbf{V}}_n\widehat{\mathbf{R}}_n\right)^{-1}\widehat{\mathbf{R}}_n'\widehat{\mathbf{V}}_n\widehat{\mathbf{G}}_n\widehat{\mathbf{W}}_n^{1/2} \to_p \mathbf{A}$$

and

$$\widehat{\tau}_n^* = \operatorname{tr}(\widehat{\mathbf{A}}_n) - 2\left\|\widehat{\mathbf{A}}_n\right\| \to_p \operatorname{tr}(\mathbf{A}) - 2\|\mathbf{A}\| = \tau^*.$$

Thus $\widehat{\tau}_n^*$ satisfies Assumption 3 with $\tau = \tau^*$. $\tau^* > 2$ is implied by $d > 2$, and thus $\widehat{\boldsymbol{\beta}}^*$ satisfies the conditions of Theorem 2. Thus (30) and (32) hold with $\tau = \tau^*$, and the latter bound is is

$$\begin{aligned}\operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}) - \frac{\tau^*\left(2\left(\operatorname{tr}(\mathbf{A}) - 2\|\mathbf{A}\|\right) - \tau^*\right)}{\operatorname{tr}(\mathbf{A})(c+1)} &= \operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}) - \frac{\left(\operatorname{tr}(\mathbf{A}) - 2\|\mathbf{A}\|\right)^2}{\operatorname{tr}(\mathbf{A})(c+1)} \\ &= \operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}) - \operatorname{tr}(\mathbf{A})\frac{(1-2/d)^2}{(c+1)}\end{aligned}$$

which is (35). ■

Proof of Theorem 3. Under (37) and (38), we can see that $\tau/\operatorname{tr}(\mathbf{A}) \to 1$ and $\tau/\operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}) \to a$ as $d \to \infty$. Then from (32) and the definition (36)

$$\begin{aligned}\sup_{\boldsymbol{h}\in\mathbf{H}(c)} \overline{\rho}(\boldsymbol{h}, \widehat{\boldsymbol{\theta}}^*) &\leq \frac{1}{\operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}})}\left(\operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}) - \frac{\tau\left(2\left(\operatorname{tr}(\mathbf{A}) - 2\|\mathbf{A}\|\right) - \tau\right)}{\operatorname{tr}(\mathbf{A})(c+1)}\right) \\ &= 1 - \frac{\tau}{\operatorname{tr}(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}})}\frac{\left(2\left(1-2/d\right) - \tau/\operatorname{tr}(\mathbf{A})\right)}{(c+1)} \\ &\longrightarrow 1 - \frac{a}{c+1}\end{aligned}$$

as $d \to \infty$, as stated. ■

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 $\mathbf{G} \mapsto \mathbf{V}^{1/2}\mathbf{G}$ so that $\mathbf{V} = \mathbf{I}_m$. Then write $\mathbf{R} = \mathbf{Q}\begin{pmatrix}\mathbf{0}_{m-p} \\ \mathbf{I}_p\end{pmatrix}\mathbf{D}$ where $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_p$ and $\mathbf{D}$ is full rank. Make the transformations $\boldsymbol{h} \mapsto \mathbf{Q}'\boldsymbol{h}$, $\mathbf{R} \mapsto \mathbf{Q}'\mathbf{R}\mathbf{D}^{-1}$ and $\mathbf{G} \mapsto \mathbf{Q}\mathbf{G}$. Hence $\mathbf{V} = \mathbf{I}_m$ and $\mathbf{R} = \begin{pmatrix}\mathbf{0}_{m-p} \\ \mathbf{I}_p\end{pmatrix}$ as claimed.

Partition $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2)$, $\mathbf{G} = (\mathbf{G}_1', \mathbf{G}_2')'$, and $\mathrm{Z} = (\mathrm{Z}_1', \mathrm{Z}_2')'$ comformably with $\mathbf{R}$ so that the $\boldsymbol{h}_1$, Z_1, and $\mathbf{G}_1$ have $m-p$ rows while $\boldsymbol{h}_2$, Z_2 and $\mathbf{G}_2$ have p rows. Note that after these transformations $\mathbf{A} = \mathbf{G}_2\mathbf{W}\mathbf{G}_2'$ and $\mathbf{H}(c) = \left\{\boldsymbol{h} : \boldsymbol{h}_2'\mathbf{A}\boldsymbol{h}_2 \leq \operatorname{tr}(\mathbf{A})\, c\right\}$.

Set $\eta = 1 - \psi d^{-1/3}$ for $\psi = 10^{1/3}(1+c)^{1/3}$. Note that $0 < \eta < 1$ for d sufficiently large. 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 $\widetilde{\boldsymbol{h}}_1 = \mathbb{E}(\boldsymbol{h}_1 \mid \mathrm{Z})$ and $\widetilde{\boldsymbol{h}}_2 = \mathbb{E}(\boldsymbol{h}_2 \mid \mathrm{Z})$ be the Bayes estimators of $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ under these priors. By standard calculations, $\widetilde{\boldsymbol{h}}_1 = \dfrac{\omega}{1+\omega}\mathrm{Z}_1$ and $\widetilde{\boldsymbol{h}}_2 = \dfrac{c\eta}{1+c\eta}\mathrm{Z}_2$. Also, let $\Pi_2^*(\boldsymbol{h}_2)$ be the prior $\Pi_2(\boldsymbol{h}_2)$ truncated to the region $\mathbf{H}_2(c) = \left\{\boldsymbol{h}_2 : \boldsymbol{h}_2'\mathbf{A}\boldsymbol{h}_2 \le \mathrm{tr}(\mathbf{A})\, c\right\}$, and let $\widetilde{\boldsymbol{h}}_2^* = \mathbb{E}(\boldsymbol{h}_2 \mid \mathrm{Z})$ 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 $\widetilde{\boldsymbol{h}}_2^* \in \mathbf{H}_2(c)$ or

$$\widetilde{\boldsymbol{h}}_2^{*\prime}\mathbf{A}\widetilde{\boldsymbol{h}}_2^* \le \mathrm{tr}(\mathbf{A})\, c. \tag{61}$$

Also, since Z_1 and Z_2 are independent, and Π_1 and Π_2^* are independent, it follows that $\widetilde{\boldsymbol{h}}_2^*$ is a function of Z_2 only, and $\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1$ and $\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2$ are independent. Set $\widetilde{\boldsymbol{h}} = \left(\widetilde{\boldsymbol{h}}_1', \widetilde{\boldsymbol{h}}_2^{*\prime}\right)'$ and $\widetilde{\boldsymbol{T}} = \mathbf{G}'\widetilde{\boldsymbol{h}}$, which is the Bayes estimator of $\mathbf{G}'\boldsymbol{h}$.

For any estimator $\boldsymbol{T} = \boldsymbol{T}(\mathrm{Z})$, 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}_{\boldsymbol{h}}\ell(\boldsymbol{h}, \boldsymbol{T}) &\ge \int\int \mathbb{E}_{\boldsymbol{h}}\ell(\boldsymbol{h}, \boldsymbol{T}) d\Pi_1(\boldsymbol{h}_1) d\Pi_2^*(\boldsymbol{h}_2) && (62)\\
&\ge \int\int \mathbb{E}_{\boldsymbol{h}}\ell(\boldsymbol{h}, \widetilde{\boldsymbol{T}}) d\Pi_1(\boldsymbol{h}_1) d\Pi_2^*(\boldsymbol{h}_2)\\
&= \int\int \mathbb{E}_{\boldsymbol{h}}\left(\widetilde{\boldsymbol{h}} - \boldsymbol{h}\right)' \mathbf{G}\mathbf{W}\mathbf{G}'\left(\widetilde{\boldsymbol{h}} - \boldsymbol{h}\right) d\Pi_1(\boldsymbol{h}_1) d\Pi_2^*(\boldsymbol{h}_2)\\
&= \int\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1\right)' \mathbf{G}_1\mathbf{W}\mathbf{G}_1'\left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1\right)\right] d\Pi_1(\boldsymbol{h}_1) d\Pi_2^*(\boldsymbol{h}_2)\\
&\quad + 2\int\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1\right)' \mathbf{G}_1\mathbf{W}\mathbf{G}_2'\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_1(\boldsymbol{h}_1) d\Pi_2^*(\boldsymbol{h}_2)\\
&\quad + \int\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)' \mathbf{G}_2\mathbf{W}\mathbf{G}_2'\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_1(\boldsymbol{h}_1) d\Pi_2^*(\boldsymbol{h}_2) && (63)\\
&= \int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1\right)' \mathbf{G}_1\mathbf{W}\mathbf{G}_1'\left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1\right)\right] d\Pi_1(\boldsymbol{h}_1) && (64)\\
&\quad + 2\left(\int \mathbb{E}_{\boldsymbol{h}}\left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1\right) d\Pi_1(\boldsymbol{h}_1)\right)' \mathbf{G}_1\mathbf{W}\mathbf{G}_2'\left(\int \left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right) d\Pi_2^*(\boldsymbol{h}_2)\right) && (65)\\
&\quad + \frac{\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)' \mathbf{A}\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} && (66)\\
&\quad - \frac{\int_{\mathbf{H}_2(c)^c} \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)' \mathbf{A}\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} && (67)
\end{aligned}$$

where the second inequality is because the Bayes estimator $\widetilde{\boldsymbol{T}}$ minimizes the right-hand-side of (62). The final equality uses the fact that $\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1$ and $\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2$ are independent, the equality $\mathbf{G}_2\mathbf{W}\mathbf{G}_2' = \mathbf{A}$, and breaks the integral (63) over the truncated prior (which has support on $\mathbf{H}_2(c)$) into the difference of the integrals over the non-truncated prior over $\mathbb{R}^m$ and $\mathbf{H}_2(c)^c$, respectively.

We now treat the four components (64)-(67) separately.

First, since $\widetilde{\boldsymbol{h}}_1 = \dfrac{\omega}{1+\omega}\mathrm{Z}_1$ and $\Pi_1(\boldsymbol{h}_1) = \mathrm{N}\left(\mathbf{0}, \mathbf{I}_{m-p}\omega\right)$, we calculate that (64) equals

$$\begin{aligned}
&\int \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1\right)' \mathbf{G}_1\mathbf{W}\mathbf{G}_1'\left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1\right)\right] d\Pi_1(\boldsymbol{h}_1) \\
&= \int \mathbb{E}_{\boldsymbol{h}}\left[\left(\frac{\omega}{1+\omega}\mathrm{Z}_1 - \boldsymbol{h}_1\right)' \mathbf{G}_1\mathbf{W}\mathbf{G}_1'\left(\frac{\omega}{1+\omega}\mathrm{Z}_1 - \boldsymbol{h}_1\right)\right] d\Pi_1(\boldsymbol{h}_1) \\
&= \int \left[\frac{1}{(1+\omega)^2}\boldsymbol{h}_1'\mathbf{G}_1\mathbf{W}\mathbf{G}_1'\boldsymbol{h}_1 + \frac{\omega^2}{(1+\omega)^2}\operatorname{tr}\left(\mathbf{G}_1\mathbf{W}\mathbf{G}_1'\right)\right] d\Pi_1(\boldsymbol{h}_1) \\
&= \operatorname{tr}\left(\mathbf{G}_1\mathbf{W}\mathbf{G}_1'\right)\frac{\omega}{1+\omega}. \qquad (68)
\end{aligned}$$

Second, since

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

it follows that (65) equals zero.

Third, take (66). Because $\widetilde{\boldsymbol{h}}_2$ is the Bayes estimator under the prior Π_2,

$$\begin{aligned}
&\frac{\int \mathbb{E}\left[\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)' \mathbf{A}\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} \\
&\geq \frac{\int \mathbb{E}\left[\left(\widetilde{\boldsymbol{h}}_2 - \boldsymbol{h}_2\right)' \mathbf{A}\left(\widetilde{\boldsymbol{h}}_2 - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} \\
&\geq \int \mathbb{E}\left[\left(\widetilde{\boldsymbol{h}}_2 - \boldsymbol{h}_2\right)' \mathbf{A}\left(\widetilde{\boldsymbol{h}}_2 - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2) \\
&= \operatorname{tr}\left(\mathbf{A}\right)\frac{c\eta}{1+c\eta} \qquad (69) \\
&\geq \operatorname{tr}\left(\mathbf{A}\right)\left(1 - \frac{c\eta}{1+c}\right) \\
&= \operatorname{tr}\left(\mathbf{A}\right)\left(1 - \frac{1}{1+c} - \frac{c}{1+c}\psi d^{-1/3}\right) \qquad (70)
\end{aligned}$$

where (69) is a calculation similar to (68) using $\widetilde{\boldsymbol{h}}_2 = \dfrac{c\eta}{1+c\eta}\mathrm{Z}_2$ and $\boldsymbol{h}_2 \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}_p c\eta\right)$. (70) uses $\eta = 1 - \psi d^{-1/3}$.

Fourth, take (67). Our goal is to show that this term is negligible for large d, and our argument is based on the proof of Theorem 7.28 from Wasserman (2006). Set

$$\zeta = \frac{\boldsymbol{h}_2'\mathbf{A}\boldsymbol{h}_2}{c\operatorname{tr}\left(\mathbf{A}\right)}.$$

Since $\boldsymbol{h}_2 \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_p c\eta)$ we see that $\mathbb{E}\zeta = \eta$. Use $(\boldsymbol{a}+\boldsymbol{b})'(\boldsymbol{a}+\boldsymbol{b}) \leq 2\boldsymbol{a}'\boldsymbol{a} + 2\boldsymbol{b}'\boldsymbol{b}$ and (61) to find that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)' \mathbf{A}\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)\right] &\leq 2\mathbb{E}_{\boldsymbol{h}}\left(\widetilde{\boldsymbol{h}}_2^{*\prime}\mathbf{A}\widetilde{\boldsymbol{h}}_2^*\right) + 2\boldsymbol{h}_2'\mathbf{A}\boldsymbol{h}_2 \\
&\leq 2\,\mathrm{tr}\,(\mathbf{A})\,c + 2\boldsymbol{h}_2'\mathbf{A}\boldsymbol{h}_2 \\
&= 2\,\mathrm{tr}\,(\mathbf{A})\,c\,(1+\zeta) \\
&\leq 2\,\mathrm{tr}\,(\mathbf{A})\,c\,(2+\zeta-\eta)\,. \qquad (71)
\end{aligned}$$

Note that $\boldsymbol{h}_2 \in \mathbf{H}_2(c)^c$ is equivalent to $\zeta > 1$. Using (71) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\int_{\mathbf{H}_2(c)^c} \mathbb{E}_{\boldsymbol{h}}\left[\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)' \mathbf{A}\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2\right)\right] d\Pi_2(\boldsymbol{h}_2) \\
&\leq 2\,\mathrm{tr}\,(\mathbf{A})\,c\left[2\int_{\mathbf{H}_2(c)^c} d\Pi_2(\boldsymbol{h}_2) + \int_{\mathbf{H}_2(c)^c} (\zeta-\eta)\, d\Pi_2(\boldsymbol{h}_2)\right] \\
&\leq 2\,\mathrm{tr}\,(\mathbf{A})\,c\left[2\mathbb{P}\,(\zeta>1) + \mathrm{var}\,(\zeta)^{1/2}\,\mathbb{P}\,(\zeta>1)^{1/2}\right]. \qquad (72)
\end{aligned}$$

Letting w_j denote the eigenvalues of $\mathbf{A}$ then we can write

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

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

$$\mathrm{var}\,(\zeta) = \frac{\eta^2}{\left(\sum_{j=1}^p w_j\right)^2}\sum_{j=1}^p w_j^2\,\mathrm{var}\left(y_j^2\right) \leq d^{-1} \qquad (73)$$

since $d = \mathrm{tr}\,(\mathbf{A})\,/\,\|\mathbf{A}\| = \dfrac{\sum_{j=1}^p w_j}{\max_j w_j}$. By Markov's inequality, (73), and $1-\eta = \psi d^{-1/3}$,

$$\mathbb{P}\,(\zeta > 1) = \mathbb{P}\,(\zeta - \eta > 1 - \eta) \leq \frac{\mathrm{var}\,(\zeta)}{(1-\eta)^2} \leq \frac{d^{-1/3}}{\psi^2}. \qquad (74)$$

Furthermore, (74) implies that

$$\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2) = 1 - \mathbb{P}\,(\zeta > 1) \geq 1 - \frac{d^{-1/3}}{\psi^2}. \qquad (75)$$

It follows from (72), (73), (74), and (75) that

$$\begin{aligned}
&\frac{\int_{\mathbf{H}_2(c)^c} \mathbb{E}_{\boldsymbol{h}} \left[\left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2 \right)' \mathbf{A} \left(\widetilde{\boldsymbol{h}}_2^* - \boldsymbol{h}_2 \right) \right] d\Pi_2(\boldsymbol{h}_2)}{\int_{\mathbf{H}_2(c)} d\Pi_2(\boldsymbol{h}_2)} \\
&\leq 2c \operatorname{tr}(\mathbf{A}) \frac{\left(2\frac{d^{-1/3}}{\psi^2} + \frac{d^{-2/3}}{\psi} \right)}{1 - \frac{d^{-1/3}}{\psi^2}} \\
&\leq 5c \operatorname{tr}(\mathbf{A}) \psi^{-2} d^{-1/3} \qquad (76)
\end{aligned}$$

the final inequality for sufficiently large d.

Together, (68), (70), and (76) applied to (64)-(67) and noting that ω is arbitrary show that

$$\begin{aligned}
\sup_{\boldsymbol{h} \in \mathbf{H}(c)} \mathbb{E}_{\boldsymbol{h}} \ell(\boldsymbol{h}, \boldsymbol{T}) &\geq \operatorname{tr}\left(\mathbf{G}_1 \mathbf{W} \mathbf{G}_1' \right) + \left(1 - \frac{1}{1+c} - \left(\frac{1}{1+c} \psi + 5\psi^{-2} \right) c d^{-1/3} \right) \operatorname{tr}(\mathbf{A}) \\
&\geq \operatorname{tr}\left(\mathbf{G}_1 \mathbf{W} \mathbf{G}_1' \right) + \left(1 - \frac{1}{1+c} - \frac{4c}{(1+c)^{2/3}} d^{-1/3} \right) \operatorname{tr}(\mathbf{A}) \\
&= \operatorname{tr}(\mathbf{W} \mathbf{V}_{\boldsymbol{\beta}}) - \left(\frac{1}{1+c} + \frac{4c}{(1+c)^{2/3}} d^{-1/3} \right) \operatorname{tr}(\mathbf{A})
\end{aligned}$$

which is (45). The second equality uses the fact that we had set $\psi = 10^{1/3}(1+c)^{1/3}$ and uses the fact $10^{1/3} + 5/10^{2/3} \leq 4$. The final equality uses

$$\begin{aligned}
\operatorname{tr}\left(\mathbf{G}_1 \mathbf{W} \mathbf{G}_1' \right) + \operatorname{tr}(\mathbf{A}) &= \operatorname{tr}\left(\mathbf{G}_1 \mathbf{W} \mathbf{G}_1' \right) + \operatorname{tr}\left(\mathbf{G}_2 \mathbf{W} \mathbf{G}_2' \right) \\
&= \operatorname{tr}\left(\mathbf{G}' \mathbf{W} \mathbf{G} \right) \\
&= \operatorname{tr}(\mathbf{W} \mathbf{V}_{\boldsymbol{\beta}})
\end{aligned}$$

which holds 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 $\widetilde{\boldsymbol{h}}_2^*$ 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 the main difference that 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. Then for a subsequence $\{n_k\}$ of $\{n\}$,

$$\sup_{I\subset\mathbf{H}(c)}\liminf_{n\to\infty}\sup_{\boldsymbol{h}\in I}\mathbb{E}_{\boldsymbol{\theta}_n}\, n\ell\left(\boldsymbol{\beta}_n,\boldsymbol{T}_n\right)\geq\lim_{k\to\infty}\liminf_{n\to\infty}\sup_{\boldsymbol{h}\in Q_k}\mathbb{E}_{\boldsymbol{\theta}_n}\, n\ell\left(\boldsymbol{\beta}_n,\boldsymbol{T}_n\right)$$
$$=\lim_{k\to\infty}\sup_{\boldsymbol{h}\in Q_k}\mathbb{E}_{\boldsymbol{\theta}_{n_k}}\, n_k\ell\left(\boldsymbol{\beta}_{n_k},\boldsymbol{T}_{n_k}\right). \tag{77}$$

We can assume that the lower bound (77) is finite, else (48) holds trivially. This implies that $n_k\ell\left(\boldsymbol{\beta}_{n_k},\boldsymbol{T}_{n_k}\right)=O_p(1)$. It also implies together with Assumption 4 that $\boldsymbol{\beta}_{n_k}-\boldsymbol{T}_{n_k}=o_p(1)$. To see this, suppose not. Then there exists some $\varepsilon>0$ and $\eta>0$ such that $\limsup_{n_k\to\infty}\mathbb{P}\left(\left|\boldsymbol{\beta}_{n_k}-\boldsymbol{T}_{n_k}\right|\geq\varepsilon\right)\geq\eta$, which together with Assumption 4 implies

$$\mathbb{E}_{\boldsymbol{\theta}_{n_k}}\, n_k\ell\left(\boldsymbol{\beta}_{n_k},\boldsymbol{T}_{n_k}\right)\geq n_k\inf_{|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2|\geq\varepsilon}\ell\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right)\mathbb{P}\left(\left|\boldsymbol{\beta}_{n_k}-\boldsymbol{T}_{n_k}\right|\geq\varepsilon\right)$$
$$\geq n_k c_\varepsilon\eta\to\infty$$

which contradicts the assumption that (77) is finite. Thus we conclude $\boldsymbol{\beta}_{n_k}-\boldsymbol{T}_{n_k}=o_p(1)$ as claimed.

By the Taylor expansion in (53),

$$n_k\ell\left(\boldsymbol{\beta}_{n_k},\boldsymbol{T}_{n_k}\right)=n_k\left(\boldsymbol{T}_{n_k}-\boldsymbol{\beta}_{n_k}\right)'\mathbf{W}_{n_k}\left(\boldsymbol{T}_{n_k}-\boldsymbol{\beta}_{n_k}\right) \tag{78}$$

where $\mathbf{W}_{n_k}=\mathbf{W}\left(\boldsymbol{\beta}_k^*\right)$ and $\boldsymbol{\beta}_k^*$ lies on the line segment joining $\boldsymbol{T}_{n_k}$ and $\boldsymbol{\beta}_{n_k}$. Since $\boldsymbol{\beta}_{n_k}\to\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_{n_k}-\boldsymbol{T}_{n_k}=o_p(1)$, it follows that $\boldsymbol{\beta}_k^*\to_p\boldsymbol{\beta}_0$ and $\mathbf{W}_{n_k}\to_p\mathbf{W}$ under Assumption 2. Since (78) is $O_p(1)$ and $\mathbf{W}>0$, it follows that $\sqrt{n_k}\left(\boldsymbol{T}_{n_k}-\boldsymbol{\beta}_{n_k}\right)=O_p(1)$.

As argued in the first paragraph of the proof of Theorem 8.11 of van der Vaart (1998), the tightness of $\sqrt{n_k}\left(\boldsymbol{T}_{n_k}-\boldsymbol{\beta}_{n_k}\right)$, Prohorov's theorem, differentiability in quadratic mean, differentiability of $\mathbf{g}\left(\boldsymbol{\theta}\right)$, and Le Cam's third lemma allow us to deduce that there is a subsequence $\{n'\}$ of $\{n_k\}$ such that

$$\sqrt{n'}\left(\boldsymbol{T}_{n'}-\boldsymbol{\beta}_{n'}\right)\xrightarrow[\boldsymbol{\theta}_n]{}\boldsymbol{T}(\mathrm{Z})-\mathbf{G}'\boldsymbol{h} \tag{79}$$

under $\boldsymbol{h}$ for a (possibly randomized) function of $\mathrm{Z}\sim\mathrm{N}_m\left(\boldsymbol{h},\mathbf{V}\right)$. Hence along this sequence, applying (78) and (79)

$$n'\ell\left(\boldsymbol{\theta}_{n'},\boldsymbol{T}_{n'}\right)=n'\left(\boldsymbol{T}_{n'}-\boldsymbol{\beta}_{n'}\right)'\mathbf{W}_{n'}\left(\boldsymbol{T}_{n'}-\boldsymbol{\beta}_{n'}\right)$$
$$\xrightarrow[\boldsymbol{\theta}_n]{}\left(\boldsymbol{T}(\mathrm{Z})-\mathbf{G}'\boldsymbol{h}\right)'\mathbf{W}\left(\boldsymbol{T}(\mathrm{Z})-\mathbf{G}'\boldsymbol{h}\right).$$

By the portmanteau lemma

$$\liminf_{n'\to\infty}\mathbb{E}_{\boldsymbol{\theta}_{n'}}\, n'\ell\left(\boldsymbol{\theta}_{n'},\boldsymbol{T}_{n'}\right)\geq\mathbb{E}_{\boldsymbol{h}}\left(\boldsymbol{T}(\mathrm{Z})-\mathbf{G}'\boldsymbol{h}\right)'\mathbf{W}\left(\boldsymbol{T}(\mathrm{Z})-\mathbf{G}'\boldsymbol{h}\right).$$

This implies that for any sufficiently large k (77) is larger than

$$\sup_{\boldsymbol{h}\in Q_k} \mathbb{E}_{\boldsymbol{h}} \left(\boldsymbol{T}(\mathrm{Z}) - \mathbf{G}'\boldsymbol{h}\right)' \mathbf{W} \left(\boldsymbol{T}(\mathrm{Z}) - \mathbf{G}'\boldsymbol{h}\right).$$

Since k is arbitrary and the right-hand-side is continuous in $\boldsymbol{h}$, we deduce that

$$\begin{aligned}
\sup_{I\subset\mathbf{H}(c)} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in I} \mathbb{E}_{\boldsymbol{\theta}_n} n\ell\left(\boldsymbol{\beta}_n, \boldsymbol{T}_n\right) &\geq \sup_{\boldsymbol{h}\in\mathbf{Q}(c)} \mathbb{E}_{\boldsymbol{h}} \left(\boldsymbol{T}(\mathrm{Z}) - \mathbf{G}'\boldsymbol{h}\right)' \mathbf{W} \left(\boldsymbol{T}(\mathrm{Z}) - \mathbf{G}'\boldsymbol{h}\right) \\
&= \sup_{\boldsymbol{h}\in\mathbf{H}(c)} \mathbb{E}_{\boldsymbol{h}} \left(\boldsymbol{T}(\mathrm{Z}) - \mathbf{G}'\boldsymbol{h}\right)' \mathbf{W} \left(\boldsymbol{T}(\mathrm{Z}) - \mathbf{G}'\boldsymbol{h}\right) \\
&\geq \mathrm{tr}\left(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}\right) - \left[\frac{1}{1+c} + \frac{4c}{(1+c)^{2/3}} d^{-1/3}\right] \mathrm{tr}\left(\mathbf{A}\right)
\end{aligned}$$

the final inequality by Theorem 4. We have shown (48).

Dividing by $\mathrm{tr}\left(\mathbf{W}\mathbf{V}_{\boldsymbol{\beta}}\right)$, and taking the limit as $d \to \infty$, we obtain

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

which is (49). ■

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] Burnham, Kenneth P. and David R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

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

[9] Casella, George and J.T.G. Hwang (1987): "Employing vague prior information in the construction of confidence sets," *Journal of Multivariate Analysis*, 21, 79-104

[10] Chernoff, Herman (1956): "Large-sample theory: Parametric case," *The Annals of Mathematical Statistics*, 27, 1-22.

[11] Claeskens, Gerda and Nils Lid Hjort (2003): "The focused information criterion," *Journal of the American Statistical Association*, 98, 900-945.

[12] DiTraglia, Francis J. (2014): "Using invalid instruments on purpose: Focused moment selection and averaging for GMM," University of Pennsylvania.

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

[14] 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.

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

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

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

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

[19] Hansen, Bruce E. (2015): "Shrinkage efficiency bounds," *Econometric Theory*, 31, 860-879.

[20] Hjort, Nils Lid and Gerda Claeskens (2003): "Frequentist model average estimators," *Journal of the American Statistical Association*, 98, 879-899.

[21] 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.

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

[23] Le Cam, L. (1972): "Limits of experiments," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 245-161, Berkeley: University of California Press.

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

[25] Liao, Zhipeng (2013): "Adaptive GMM shrinkage estimation with consistent moment selection," *Econometric Theory*, 29, 857-904.

[26] Liu, Chu-An (2015): "Distribution theory of the least squares averaging estimator," *Journal of Econometrics*, 186, 142-159.

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

[28] McCloskey, Adam (2015): "Bonferroni-based size-correction for nonstandard testing problems," Brown University.

[29] 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.

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

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

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

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

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

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

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

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

[38] 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.

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

[40] Tseng, Yu-Ling and Lawrence D. Brown (1997): "Good exact confidence sets and minimax estimators for the mean vector of a multivariate normal distribution," *Annals of Statistics*, 25, 2228-2258.

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

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

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