Efficient Shrinkage in Parametric Models

Bruce E. Hansen*
University of Wisconsin†

September 2012
Revised: June 2014

Abstract

This paper introduces shrinkage for general parametric models. We show how to shrink maximum likelihood estimators towards parameter subspaces defined by general nonlinear restrictions. We derive the asymptotic distribution and risk of a shrinkage estimator using a local asymptotic framework. We show that if the shrinkage dimension exceeds two, the asymptotic risk of the shrinkage estimator is strictly less than that of the 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.

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

We also provide a new large sample minimax efficiency bound. We use the concept of local asymptotic minimax bounds, a generalization of the conventional asymptotic minimax bounds. The difference is that we consider minimax regions that are defined locally to the parametric restriction, and are thus tighter. We show that our shrinkage estimator asymptotically achieves this local asymptotic minimax bound when the shrinkage dimension is high. This theory is a combination and extension of standard asymptotic efficiency theory (Hájek, 1972) and local minimax efficiency theory for Gaussian models (Pinsker, 1980).

*Research supported by the National Science Foundation. 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 various 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 of any estimator equals the asymptotic risk of the MLE. In apparent contradiction to these results, James and Stein (1961) established that a simple shrinkage estimator in a Gaussian context has lower risk than the MLE. The reconciliation of these two branches of theory is typically based on the assertion that shrinkage is a local finite-sample effect which disappears asymptotically, and thus there are no asymptotic gains from shrinkage.

In this paper we show that this understanding is incomplete. We show that shrinkage applied to the MLE can substantially reduce the asymptotic risk (weighted mean-squared error). This demonstrates that the MLE is asymptotically inefficient. Indeed, our results show that the gains from shrinkage persist asymptotically in an appropriate local asymptotic framework.

Shrinkage necessarily depends on a shrinkage direction (in our paper treated as a parametric restriction), and the magnitude of the improvement in asymptotic risk depends on the distance between the true parameters and the shrinkage direction. If the distance is small then the reduction in risk can be quite substantial. Even when the distance is moderately large the reduction in risk can be significant.

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

To establish these results we need to develop an asymptotic (large sample) distributional theory for shrinkage estimators. This can be accomplished using the local asymptotic normality approach (e.g., van der Vaart (1998)). We model the parameter vector as being in a $n^{-1/2}$-neighborhood of the specified restriction, so that the asymptotic distributions are continuous in the localizing parameter. This approach has been used successfully for averaging estimators by Hjort and Claeskens (2003) and Liu (2013), and for Stein-type estimators by Saleh (2006).

Given the localized asymptotic parameter structure, the asymptotic distribution of the shrinkage estimator takes a James-Stein form. It follows that the asymptotic risk of the estimator can be analyzed using techniques introduced by Stein (1981). Not surprisingly, the benefits of shrinkage are maximized when the magnitude of the localizing parameter is small. What is surprising (or at least it may be to some readers) is that the numerical magnitude of the reduction in asymptotic risk (weighted mean-squared error) is quite substantial, even for relatively distant values of the localizing parameter. We can be very precise about the nature of this improvement, as we provide simple and interpretable expressions for the asymptotic risk.
We measure estimation efficiency by asymptotic risk – the large sample weighted mean-squared error. The weighted MSE necessarily depends on a weight matrix, and the optimal shrinkage estimator depends on its value. For a generic measure of fit the weight matrix can be set to the identity matrix or the inverse of the asymptotic covariance matrix, but in other cases a user may wish to select a specific weight matrix so we allow for an arbitrary user-specified weight matrix. Weighted MSE is a standard criteria in the shrinkage literature, including Bhattacharya (1966), Sclove (1968), and Berger (1976a, 1976b, 1982). What is different about our approach relative to these papers is that our estimator does not require the weight matrix to be positive definite. This may be particularly important in econometric applications where nuisance parameters are commonplace.

We benefit from the recent theory of efficient high-dimensional Gaussian shrinkage, specifically Pinkser’s Theorem (Pinsker, 1980), which gives a lower minimax bound for estimation of high dimensional normal means. We combine Pinker’s Theorem with classic large-sample minimax efficiency theory (Hájek, 1972) to provide a new asymptotic local minimax efficiency bound. We provide a minimax lower bound on the asymptotic risk, and show that the asymptotic risk of our shrinkage estimator equals this lower bound when the shrinkage dimension diverges towards infinity. This shows that the proposed shrinkage estimator is minimax efficient under high-dimensional shrinkage.

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 weighted quadratic loss functions. 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. 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 Efroimovich (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 shrinkage 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 theory concerning how to select the frequentist model averaging weights in the nonlinear likelihood context, and there is no theory concerning the optimality of such estimators. 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, while this paper focuses on likelihood models.

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 (2012). 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 a new 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 sample $\tilde{X} = \{X_1, ..., X_n\}$ of observations from a twice differentiable density $f(x, \theta)$ indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^m$. The goal is to find an estimator
\( T_n = T_n(\bar{X}) \) of \( \theta \) with small risk under quadratic loss

\[
R(\theta, T_n) = \mathbb{E}_\theta (T_n - \theta)' W (T_n - \theta) \tag{1}
\]

for some positive semi-definite weight matrix \( W \). In (1), \( \mathbb{E}_\theta \) denotes expectation with respect to the density \( f(x, \theta) \). Equivalently, (1) is weighted mean-squared error, since

\[
R(\theta, T_n) = \text{tr} \left( W \mathbb{E}_\theta (T_n - \theta)' (T_n - \theta) \right) = \sum_{j=1}^{m} \sum_{k=1}^{m} W_{jk} \mathbb{E}_\theta \left[ (T_{nj}(\bar{X}) - \theta_j) (T_{nk}(\bar{X}) - \theta_k) \right].
\]

We discuss weight matrix choices below.

The estimation setting is augmented by the belief that the true value of \( \theta \) may be close (in a sense to be made precise later) to a restricted parameter space \( \Theta_0 \subset \Theta \) defined by a differentiable parametric restriction

\[
\Theta_0 = \{ \theta \in \Theta : r(\theta) = 0 \} \tag{2}
\]

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

The goal is estimation efficiency: parameter estimation with minimum expected loss. The role of the restriction (2) will be to center a shrinkage estimator. As we show later, such shrinkage can uniformly reduce asymptotic risk in a sense to be made precise later. We will also show that such shrinkage achieves an asymptotic efficiency bound, and is therefore asymptotically efficient.

This setting is largely described by the choice of loss weight matrix \( W \) and parametric restriction \( r(\theta) \). We now describe these choices.

### 2.1 Risk Function

The risk function (1) with a general weight matrix \( W \) is useful as it includes many important special cases. We now list some examples.

**Example 1** Unweighted MSE is obtained by setting \( W = I_m \). This is a reasonable choice when the coefficients are scaled to be of roughly equal magnitude, and the coefficients are of equal importance.

**Example 2** Define the information matrix \( \mathcal{I}_\theta = \mathbb{E}_\theta \left( -\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X_{ni}, \theta) \right) \). Setting \( W = \mathcal{I}_\theta \) renders the loss function invariant to rotations of the parameter space. As we see later this choice simplifies many formulae, and consequently we call the weight matrix \( W = \mathcal{I}_\theta \) the canonical case.

**Example 3** If some elements of \( \theta \) are nuisance parameters, meaning that we do not put value on precise estimation of these components, it would make sense for the loss function to put zero weight on these elements. In this case, partition \( \theta = (\theta'_a, \theta'_b)' \) where \( \theta_a (m_a \times 1) \) are the parameters of interest and \( \theta_b (m_b \times 1) \) are nuisance parameters. Then we can set \( W = \text{diag} \{ W_{aa}, 0 \} \) where \( W_{aa} > 0 \) is \( m_a \times m_a \). By this construction, the loss function puts all the weight on \( \theta_a \), the parameters of interest.
Example 4 Consider estimation of the conditional mean $g(x) = x'\theta$ in a regression model $y_i = x_i'\theta + e_i$. Given an estimate $T_n$ of $\theta$, the estimate of the conditional mean is $\hat{g}(x) = x'T_n$ which has integrated mean-squared error

$$IMSE = \int \mathbb{E}_\theta (g(x) - \hat{g}(x))^2 w(x)dx = \mathbb{E}_\theta [(T_n - \theta)'W(T_n - \theta)]$$

where $W = \int xx'w(x)dx = \mathbb{E}(x_ix_i'w(x_i))$. This is identical to the risk function (1) with a specific weight matrix.

Example 5 Consider out-of-sample forecasting from a linear regression model $y_i = x_i'\theta + e_i$ based on an estimate $T_n$ of $\theta$. In this case the point forecast for $y_{n+1}$ is $x_{n+1}'T_n$ with out-of-sample mean-squared forecast error

$$\mathbb{E}_\theta (y_{n+1} - x_{n+1}'T_n)^2 = \mathbb{E}_\theta e_{n+1}^2 + \mathbb{E}_\theta [(T_n - \theta)'W(T_n - \theta)]$$

where $W = \mathbb{E}(x_{n+1}x_{n+1}')$. This is a translation of the risk function (1), again with a specific weight matrix.

In our first two examples (1 and 2) the weight matrix $W$ was motivated by ease of use and simplicity. In the other examples, the weight matrix was (at least partially) determined by the econometric problem. In general, the choice of weight matrix matters as it defines the risk function, and the shrinkage estimator will depend on $W$.

2.2 Shrinkage Direction

The restriction (2) defines the shrinkage direction, and its choice is critical for the construction of the shrinkage estimator. The restriction is not believed to be true, but instead represents a belief about where $\theta$ is likely to be close. The restricted space $\Theta_0$ should be a plausible simplification, centering, or “prior” about the likely value of $\theta$. 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 in applied economics 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, we believe that since restrictions are routine in applied economics, shrinkage estimators can be made similarly routine.

An important (and classical) case occurs when $\Theta_0 = \{\theta_0\}$ 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 $\Theta$ as the “kitchen-sink”, and the restricted model $\Theta_0$ as a tight parametric specification.
Quite commonly, $\Theta_0$ will take the form of an exclusion restriction. For example, if we partition

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

then an exclusion restriction takes the form $r(\theta) = \theta_2$. This is appropriate when the model with $\theta_2 = 0$ is viewed as a useful yet parsimonious approximation. For example, $\theta_2$ may the coefficients from 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, $\Theta_0$ may also be a linear subspace, in which case we can write

$$r(\theta) = R\theta - a$$

where $R$ is $m \times p$ and $a$ is $p \times 1$. One common example is to shrink the elements of $\theta$ to a common mean, in which case we would set

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

and $a = 0$. This is appropriate when $\theta$ 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, $\Theta_0$ may be a nonlinear subspace, for example one of the restrictions could be $r_j(\theta) = \theta_1\theta_2 - 1$ which would shrink the coefficients towards $\theta_1 = 1/\theta_2$. In general, nonlinear restrictions may be useful when an economic model or hypothesis implies a set of nonlinear restrictions on the coefficients.

3 Estimation

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

$$L_n(\theta) = \sum_{i=1}^{n} \log f(X_i, \theta).$$

The MLE maximizes (5) over $\theta \in \Theta$

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} L_n(\theta).$$
We assume that the maximum is unique so that \( \hat{\theta}_n \) is well defined (and similarly with the other extremum estimators defined below). Let \( V = I^{-1}_\theta \) be the asymptotic variance of the MLE and let \( \tilde{V}_n \) denote any consistent estimate of \( V \), such as

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

We wish to contract (or shrink) \( \hat{\theta}_n \) towards the restricted space \( \Theta_0 \). To do so we first define a restricted estimator \( \tilde{\theta}_n \) which (at least approximately) satisfies \( \tilde{\theta}_n \in \Theta_0 \). We consider three possible restricted estimators.

1. Restricted maximum likelihood (RML)

\[
\tilde{\theta}_n = \arg \max_{\theta \in \Theta_0} L_n(\theta).
\]

2. Efficient minimum distance (EMD)

\[
\tilde{\theta}_n = \arg \min_{\theta \in \Theta_0} \left( \tilde{\theta}_n - \theta \right)' \tilde{V}_n^{-1} \left( \tilde{\theta}_n - \theta \right).
\]

3. Projection

\[
\tilde{\theta}_n = \tilde{\theta}_n - \tilde{V}_n \tilde{R}_n \left( \tilde{R}_n' \tilde{V}_n \tilde{R}_n \right)^{-1} r(\tilde{\theta}_n)
\]

for \( \tilde{R}_n = R(\tilde{\theta}_n) \).

The RML and EMD estimators satisfy \( \tilde{\theta}_n \in \Theta_0 \), and the projection estimator satisfies \( \tilde{\theta}_n \in \Theta_0 \) for linear restrictions (3). For non-linear restrictions, however, the projection estimator only asymptotically satisfies the restriction. The three estimators are asymptotically equivalent under our assumptions. Alternatively, we could consider minimum distance or projection estimators as in (7) or (8) but with the weight matrix \( \tilde{V}_n^{-1} \) replaced with another choice. Other choices, however, would lead to asymptotically inefficient estimators, so we confine attention to the estimators (7)-(8).

Our proposed shrinkage estimator of \( \theta \) is a weighted average of the MLE and the restricted estimator

\[
\hat{\theta}_n^* = \hat{w}_n \hat{\theta}_n + (1 - \hat{w}_n) \tilde{\theta}_n
\]

where the weight takes the form

\[
\hat{w}_n = \left( 1 - \frac{\tau_p}{D_n} \right) +.
\]

In (10), \( (x)_+ = x 1(x \geq 0) \) is the “positive-part” function,

\[
D_n = n \left( \hat{\theta}_n - \tilde{\theta}_n \right)' W \left( \hat{\theta}_n - \tilde{\theta}_n \right),
\]

is a distance-type statistic for the restriction (2) in \( \Theta \), and the scalar \( \tau_p \geq 0 \) controls the degree of
shrinkage. In practice, if $W$ or $\tau_p$ is replaced with a consistent estimate $\hat{W}_n$ or $\hat{\tau}_p$ our asymptotic theory is unaffected.

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

For the choice of shrinkage parameter, we recommend

$$\tau_p = \text{tr}(A) \left(1 - \frac{2}{p}\right),$$  

(12)

where

$$A = (R'V)\frac{1}{A}(R'VWVR).$$  

(13)

This recommendation will be justified in Section 5. In the canonical case ($W = V^{-1}$) then $A = I_p$ and $\text{tr}(A) = p$, so the recommendation (12) simplifies to $\tau_p = p - 2$, which is the shrinkage parameter recommended by James and Stein (1961). In general, when $A$ is unknown, a consistent estimator is $\hat{A}_n = \left(R_n\hat{V}_n\hat{R}_n\right)^{-1}\left(\hat{R}'\hat{V}_nW\hat{V}_n\hat{R}_n\right)$ where $\hat{R}_n = R(\hat{\theta}_n)$. We can then estimate (12) with

$$\hat{\tau}_p = \text{tr}(\hat{A}_n) \left(1 - \frac{2}{p}\right)$$  

(14)

and this is our practical recommended choice for the shrinkage parameter.

The matrix $A$ plays an important role in our theory. As mentioned above, in the canonical case ($W = V^{-1}$) then $A = I_p$ and in the full shrinkage case ($R = I_m$) we have the simpler expression $A = WW$. In general, $A$ captures the divergence between the selected weight matrix $W$ and the canonical choice $V$.

Other simplifications occur in the canonical case ($W = V^{-1}$). The full shrinkage estimator is the classic James-Stein estimator. The partial shrinkage estimator with linear $r(\theta)$ is similar to Oman’s (1982a) shrinkage estimator. The difference is that Oman (1982a) uses the estimator (8) with $\hat{V}_n$ replaced by $I_m$, thus $\hat{\theta}_n = \hat{\theta}_n - \hat{R}_n (\hat{R}'_n\hat{R}_n)^{-1}(R'\hat{\theta}_n - a)$, which is an inefficient choice unless $V = I_m$. The partial shrinkage estimator is also a special case of Hansen’s (2007) Mallows Model Averaging (MMA) estimator with two models.

It is worth mentioning that in the context of nuisance parameters (as in Example 3 above, where $\theta = (\theta'_a, \theta'_b)^\prime$ and $\theta_b$ are the nuisance parameters so that $W = \text{diag}\{W_{aa}, 0\}$) then it is not necessary to compute the estimator (9) for the entire parameter vector. If only the component $\theta_a$ is of interest, then the shrinkage estimator $\hat{\theta}_{an}^* = \hat{w}_n\hat{\theta}_{an} + (1 - \hat{w}_n)\hat{\theta}_{an}$ can be calculated simply for this component of the parameter vector, and the distance statistic (11) has the simplification

$$D_n = n \left(\hat{\theta}_{an} - \hat{\theta}_{an}\right)' W_{aa} \left(\hat{\theta}_{an} - \hat{\theta}_{an}\right).$$
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

$$\theta_n = \theta_0 + n^{-1/2}h$$

where $\theta_0 \in \Theta_0$ and $h \in \mathbb{R}^m$. In this framework $\theta_n$ is the true value of the parameter, $\theta_0$ is a centering value and $h \in \mathbb{R}^m$ is a localizing parameter.

What is important about the parameter sequences (15) is that the centering value $\theta_0$ is specified to lie in the restricted parameter space $\Theta_0$. This means that the true parameter $\theta_n$ is localized to $\Theta_0$. 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 $\theta_n$ and the restricted set $\Theta_0$ is determined by the localizing parameter $h$ and the sample size $n$. For any fixed $h$ the distance $n^{-1/2}h$ shrinks as the sample size increases. However, we do not restrict the magnitude of $h$ so this does not meaningfully limit the application of our theory. We will use the symbol $\xrightarrow{\theta_n}$ to denote convergence in distribution along the parameter sequences $\theta_n$ as defined in (15).

Assumption 1

1. The observations $X_{ni}$ are independent, identically distributed draws from the density $f(x, \theta_n)$, where $\theta_n$ satisfies (15) and $\theta_0$ is in the interior of $\Theta_0$, and $\Theta$ is compact.

2. If $\theta \neq \theta'$ then $f(x, \theta) \neq f(x, \theta')$.

3. $\log f(x, \theta)$ is continuous at each $\theta \in \Theta$ with probability one.

4. $\mathbb{E}_{\theta_0} \sup_{\theta \in \Theta} |\log f(X_{ni}, \theta)| < \infty$.

5. $I_{\theta_0} = \mathbb{E}_{\theta_0} \left(-\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X_{ni}, \theta)\right)$ exists and is non-singular.

6. For some neighborhood $\mathcal{K}$ of $\theta_0$,

   (a) $f(x, \theta)$ is twice continuously differentiable,
   (b) $\int \sup_{\theta \in \mathcal{K}} \left|\frac{\partial}{\partial \theta} f(x, \theta)\right| dx < \infty$,
   (c) $\int \sup_{\theta \in \mathcal{K}} \left|\frac{\partial^2}{\partial \theta \partial \theta'} f(x, \theta)\right| dx < \infty$,
   (d) $\mathbb{E}_{\theta_0} \sup_{\theta \in \mathcal{K}} \left|\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X_{ni}, \theta)\right| < \infty$.

7. $R(\theta)$ is continuous for $\theta \in \mathcal{K}$ and rank $(R) = p$ where $R = R(\theta_0)$.
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. Assumption 1.7 allows asymptotic normality to extend to the restricted estimator. Assumption 1.1 specifies that the observations are iid, that the parameter satisfies the local sequences (15) and that the centering value is in the interior of $\Theta_0$ 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 $b_{\theta_0}$, and 1.5 defines the Fisher information. Assumption 1.6 are regularity conditions to establish asymptotic normality.

Theorem 1 Under Assumption 1, along the sequences (15),

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{\theta} Z \sim N(0, V),$$

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{\theta} Z - VR R^TV^{-1}R' (Z + h),$$

$$D_n \xrightarrow{\theta} \xi = (Z + h)' B (Z + h),$$

$$\hat{w} \xrightarrow{\theta} w = \left(1 - \frac{\tau_p}{\xi}\right)_+,$$

and

$$\sqrt{n}(\hat{\theta}_n^* - \theta_n) \xrightarrow{\theta} wZ + (1-w) \left(Z - VR R^TV^{-1}R' (Z + h)\right),$$

where

$$B = R (R'VR)^{-1} (R'VWRV) (R'VR)^{-1} R',$$

and 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 $h$. The asymptotic distribution of $\hat{\theta}_n^*$ is written as a random weighted average of the asymptotic distributions of $\hat{\theta}_n$ and $\bar{\theta}_n$.

The asymptotic distribution is obtained for parameter sequences $\theta_n$ tending towards a point in the restricted parameter space $\Theta_0$. The case of fixed $\theta \notin \Theta_0$ can be obtained by letting $h$ diverge towards infinity, in which case $\xi \rightarrow_p \infty$, $w \rightarrow_p 1$, and the distribution on the right-hand-side of (20) tends towards $Z \sim N(0, V)$.

It is important to understand that Theorem 1 does not require that the true parameter value $\theta_n$ satisfy the restriction to $\Theta_0$, only that it is in a $n^{-1/2}$-neighborhood of $\Theta_0$. The distinction is important, as $h$ can be arbitrarily large.

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

The matrix $B$ defined in (21) will play an important role in our theory. Notice that in the full shrinkage case we have the simplification $B = W$, and in the canonical case $W = V^{-1}$ we find that $B = R (R'VR)^{-1} R'$.
Equation (18) also provides the asymptotic distribution of the distance-type statistic $D_n$. The limit random variable $\xi$ controls the weight $w$ and thus the degree of shrinkage, so it is worth investigating further. Notice that its expected value is

$$\mathbb{E}\xi = h'Bh + \mathbb{E}\text{tr} (BZZ') = h'Bh + \text{tr} (A)$$

where $B$ is from (21) and $A$ was defined in (13). In the canonical case $W = V^{-1}$ we find that (22) simplifies to

$$\mathbb{E}\xi = h'Bh + p.$$  

Furthermore, in this case, $\xi \sim \chi^2_p(h'Bh)$, a non-central chi-square random variable with non-centrality parameter $h'Bh$ and degrees of freedom $p$. In general, the scalar $h'Bh$ captures how the divergence of $\theta_n$ from the restricted region $\Theta_0$ affects the distribution of $\xi$.

## 5 Asymptotic Risk

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

Specifically, let $T = \{T_n : n = 1, 2, \ldots\}$ denote a sequence of estimators. We define the asymptotic risk of the estimator sequence $T$ for the parameter sequence $\theta_n$ defined in (15) as

$$\rho(T, h) = \lim_{\zeta \to \infty} \liminf_{n \to \infty} \mathbb{E}_{\theta_n} \min \left[ n (T_n - \theta_n)' W (T_n - \theta_n), \zeta \right].$$

This is the expected scaled loss, trimmed at $\zeta$, 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, e.g.

$$\sqrt{n} (T_n - \theta_n) \xrightarrow{\theta_n} \psi$$

for some random variable $\psi$. For then (as shown by Lemma 6.1.14 of Lehmann and Casella (1998)) we have

$$\rho(T, h) = \mathbb{E} (\psi' W \psi) = \text{tr}(W \mathbb{E} (\psi \psi')).$$ 

Thus the asymptotic risk of any estimator $T_n$ satisfying (25) can be calculated using (26).

Define the largest eigenvalue of the matrix $A$ from (13)

$$\lambda_1 = \lambda_{\text{max}}(A),$$

for some random variable $\psi$. For then (as shown by Lemma 6.1.14 of Lehmann and Casella (1998)) we have

$$\rho(T, h) = \mathbb{E} (\psi' W \psi) = \text{tr}(W \mathbb{E} (\psi \psi')).$$

Thus the asymptotic risk of any estimator $T_n$ satisfying (25) can be calculated using (26).
and the ratio
\[ d_p = \frac{\text{tr}(A)}{\lambda_1}. \]  

Notice that (28) satisfies \( 1 \leq d_p \leq p \). In the canonical case \( W = V^{-1} \) then \( A = I_p \) so \( \lambda_1 = 1 \) and we have the simplification \( d_p = p \). In general, \( d_p \) can be thought of as the effective shrinkage dimension, the number of restrictions adjusted for the difference between the weight matrix \( W \) and the canonical choice \( V^{-1} \).

**Theorem 2** Under Assumption 1, if
\[ d_p > 2 \] (29)
and
\[ 0 < \tau_p \leq 2(\text{tr}(A) - 2\lambda_1), \] (30)
then
\[ \rho(\hat{\theta}^*, h) < \rho(\hat{\theta}, h) = \text{tr}(WV) \] (31)
for any \( h \). Furthermore, for any \( 0 < c < \infty \), if we define the ball
\[ H(c) = \{ h : h'Bh \leq \text{tr}(A)c \} \] (32)
then
\[ \sup_{h \in H(c)} \rho(\hat{\theta}^*, h) \leq \text{tr}(WV) - \frac{\tau_p(2(\text{tr}(A) - 2\lambda_1) - \tau_p)}{\text{tr}(A)(c+1)}. \] (33)

Equation (31) 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 \( \tau_p \) satisfies the condition (30). As (31) holds for even extremely large values of \( h \), this inequality shows that in a very real sense the shrinkage estimator strictly dominates the MLE.

The assumption (29) \( d_p > 2 \) is the critical condition needed to ensure that the shrinkage estimator can have globally smaller asymptotic risk than the MLE. As mentioned above, in the canonical case \( W = V^{-1} \) then \( d_p = p \) so in this case (29) 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. \( d_p > 2 \) is the generalization to allow for general weight matrices.

In the general weight matrix case, the assumption \( d_p > 2 \) requires \( p > 2 \) and excludes highly unbalanced matrices \( A \). Let \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \) denote the ordered eigenvalues of \( A \). \( d_p > 2 \) is equivalent to \( \lambda_2 + \cdots + \lambda_p > \lambda_1 \). This is violated only if \( \lambda_1 \) is much larger than the other eigenvalues. One sufficient condition for \( d_p > 2 \) is that for some \( j > 1 \), \( \lambda_1/\lambda_j < j - 1 \), in words, that the ratio of the largest eigenvalue to the \( j \)’th largest is not too large.

The condition \( d_p > 2 \) is necessary in order for the right-hand-side of (30) to be positive, which is necessary for the existence of a \( \tau_p \) satisfying (30). The condition (30) simplifies to \( 0 < \tau_p \leq 2(p-2) \) in the canonical case, which is a standard restriction on the shrinkage parameter.
The shrinkage parameter $\tau_p$ appears in the risk bound (33) as a quadratic expression, so there is a unique choice $\tau^*_p = \text{tr} (A) - 2\lambda_1$ which minimizes this bound. For practical implementation we recommend replacing the maximum eigenvalue $\lambda_1$ with the average $\text{tr} (A)/p$. Substituting into the expression for $\tau^*_p$ we obtain $\tau_p = \text{tr} (A) (1 - 2/p)$ which is the recommendation (12). This is simpler than $\tau^*_p$ yet likely to be close in practice (and they are equal in the canonical case). We now reexpress Theorem 2 when the shrinkage parameter is set according to the recommendation.

**Corollary 1** Under Assumption 1, if $p \geq 3$, the shrinkage parameter $\tau_p$ is set by (12) or (14), and

$$d_p \geq \frac{4p}{p + 2}$$

then (31) holds and for any $0 < c < \infty$

$$\sup_{h \in \mathbb{H}(c)} \rho(\hat{\theta}^*, h) \leq \text{tr} (WV) - \frac{\text{tr} (A) (1 - 2/p) (1 - 4/d_p + 2/p)}{c + 1}.$$  

(35)

Corollary 1 shows that when the shrinkage parameter is set according to our recommendation, then the shrinkage estimator will asymptotically dominate the MLE so long as the condition (34) holds. In the canonical case, (34) is directly satisfied (it reduces to $p \geq 2$), but otherwise it is a strengthening of (29). A sufficient condition for (34) is $d_p \geq 4$, but this is stronger than necessary for small $p$.

### 6 High Dimensional Shrinkage

From equation (35) it appears that the risk bound will simplify as the shrinkage dimension $p$ increases. The bound itself increases with $p$, however. So to investigate this further, we define the normalized asymptotic risk as the ratio as the asymptotic risk of an estimator divided by that of the MLE.

$$\pi(T, h) = \frac{\rho(T, h)}{\rho(\hat{\theta}, h)} = \frac{\rho(T, h)}{\text{tr} (WV)}.$$  

(36)

Thus the normalized risk of the MLE is unity ($\pi(\hat{\theta}, h) = 1$) and Theorem 2 shows that the normalized risk of the shrinkage estimator is less than unity under the conditions (29) and (30). This definition is useful as it will remain bounded for the shrinkage estimator as the shrinkage dimension $p$ increase.

We now investigate the behavior of the normalized risk when the shrinkage dimension $p$ is large. To do so, we need to characterize the large $p$ behavior of the constants $\text{tr} (WV)$, $\text{tr} (A)$, $d_p$ and $\tau_p$. The matrices $A$ and $WV$ are $p \times p$ and $m \times m$ so we expect $\text{tr} (WV)$ and $\text{tr} (A)$ to diverge with $p$. We only need that their ratio converges to a constant: As $p \to \infty$

$$\frac{\text{tr} (A)}{\text{tr} (WV)} \to a.$$

(37)
The limit \( a \) is a measure of the effective number of restrictions relative to the total number of parameters. Note that \( 0 \leq a \leq 1 \), with \( a = 1 \) in the full shrinkage case and \( a = 0 \) when there is no shrinkage. In the canonical case, \( a = \lim_p \frac{p}{\frac{m}{\lambda_{\max}(A)}} \), the ratio of the number of restrictions to the total number of parameters. Since \( d_p = \text{tr}(A)/\lambda_{\max}(A) \), it seems reasonable to expect \( d_p \) to grow to infinity with \( \text{tr}(A) \). Thus we assume that as \( p \to \infty \)

\[
d_p \to \infty. \tag{38}
\]

Recall that our recommended shrinkage parameter is \( \tau_p = \text{tr}(A) (1 - 2/p) \) so for large \( p \) will grow with \( \text{tr}(A) \). We consider the class of shrinkage parameters which are asymptotically equivalent to \( \text{tr}(A) \). Thus as \( p \to \infty \)

\[
\frac{\tau_p}{\text{tr}(A)} \to 1. \tag{39}
\]

**Theorem 3** 
Under Assumption 1, if as \( p \to \infty \), (37), (38), and (39) hold, then for any \( 0 < c < \infty \),

\[
\limsup_{p \to \infty} \sup_{h \in \mathcal{H}(c)} \frac{\tau(\hat{\theta}^*, h)}{c} \leq 1 - \frac{a}{c + 1}. \tag{40}
\]

Equation (40) 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 (40) 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 (40) 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, (40) simplifies to

\[
\limsup_{p \to \infty} \sup_{h \in \mathcal{H}(c)} \frac{\tau(\hat{\theta}^*, h)}{c} \leq \frac{c}{c + 1}. \tag{41}
\]

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 (11) as \( D_n = pF_n \), where \( F_n \) is an F-type statistic for (2). Using (23), this has the approximate expectation

\[
\mathbb{E}F_n \to \frac{\mathbb{E}c}{p} = 1 + \frac{h'Bh}{p} \leq 1 + c
\]

where the inequality is for \( h \in \mathcal{H}(c) \). 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 (40) indicates that substantial efficiency gains can be achieved by shrinkage for a large region of the parameter space.

![Figure 1: Asymptotic Risk of Shrinkage Estimators](image)

To illustrate these results numerically, we plot in Figure 1 the asymptotic normalized risk of the MLE \( \hat{\theta}_n \) and our shrinkage estimator \( \hat{\theta}_n^* \), along with the risk bounds, in the full shrinkage \((m = p)\) canonical case. The asymptotic risk is only a function of \( p \) and \( c \), and we plot the risk as a function of \( c \) for \( p = 4, 8, 12, \) and 20. The asymptotic normalized risk of the MLE is the upper bound of 1. The asymptotic normalized risk of the shrinkage estimator \( \hat{\theta}_n^* \) is plotted with the solid line\(^1\). The upper bound (35) is plotted using the short dashes, and the “Large \( p \)” lower bound (41) 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 asymp-

\(^{1}\)This is calculated by simulation from the asymptotic distribution using 1,000,000 simulation draws.
totic risk relative to the MLE decreases as \( p \) increases. Furthermore, we can observe that the upper bound (33) 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 (41) 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 (41) is reasonably accurate for \( p = 20 \). Thus the technical approximation \( p \to \infty \) seems to be a useful approximation even for moderate shrinkage dimensions.

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

A limitation of Theorem 3 is that the sequential limits (first taking the sample size \( n \) to infinity and then taking the dimension \( p \) to infinity) is artificial. A deeper result would employ joint limits (taking \( n \) and \( p \) to infinity jointly). 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 (40). 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. Furthermore, 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. It is insightful to contrast shrinkage and superefficient estimators, as they have distinct properties. We examine the Hodges’ estimator in the full shrinkage (\( p = m \)) case with \( \theta_0 = 0 \) and \( W = I_m \).

Setting \( D_n = n \tilde{\theta}_n' \tilde{\theta}_n \), the Hodges’ and full shrinkage estimators can be written as

\[
\hat{\theta}_n^H = \tilde{\theta}_n 1 \left(D_n \geq \sqrt{n}\right)
\]
and
\[\tilde{\theta}_n^* = \tilde{\theta}_n \left(1 - \frac{\text{tr}(V)}{D_n} \left(1 - \frac{2}{m}\right)\right)_+.\]

There are some nominal similarities. Both \(\tilde{\theta}_n^H\) and \(\tilde{\theta}_n^\ast\) shrink \(\tilde{\theta}_n\) towards the zero vector, and both estimators equal the zero vector for sufficiently small \(\Delta\), but with different thresholds. \(\tilde{\theta}_n^H = 0\) when \(D_n < \sqrt{n}\) and \(\tilde{\theta}_n^\ast = 0\) when \(D_n < \text{tr}(V) \left(1 - \frac{2}{m}\right)\). One difference is that \(\tilde{\theta}_n^\ast\) is a smooth shrinkage function (soft thresholding) but the more important difference is that the Hodges’ estimator implicitly uses a shrinkage parameter \(\tau = \sqrt{n}\) which is increasing with \(n\).

We now study the asymptotic risk of the Hodges’ estimator. Along the sequences (15) with \(\theta_0 = 0\), the estimator has the degenerate asymptotic distribution
\[
\sqrt{n} \left(\tilde{\theta}_n^H - \theta_n\right) \xrightarrow{\text{a.s.}} -h.
\]

It follows that its uniform normalized local asymptotic risk for \(H(c) = \{h : h' h \leq \text{tr}(V) c\}\) equals
\[
\sup_{h \in H(c)} \mathcal{P}(\tilde{\theta}_n^H, h) = \sup_{h \in H(c)} \frac{h' h}{\text{tr}(V)} = c.
\]

Thus the Hodges’ estimator has uniformly lower asymptotic risk than the MLE on the sets \(h \in H(c)\) for \(c < 1\). So in this sense it improved risk relative to the MLE, but only very locally. More importantly, its uniform asymptotic risk grows towards infinity as \(c\) increases. This means that the Hodges’ estimator becomes arbitrarily inefficient on arbitrarily large sets.

The difference with the shrinkage estimator is striking. Theorem 2 shows that \(\tilde{\theta}_n^\ast\) has lower asymptotic risk than the MLE for all values of \(h\), and the bound (33) holds for all \(c\). In contrast, the asymptotic risk (42) of the Hodges’ estimator becomes unbounded as \(c \to \infty\), meaning that Hodges’ estimator becomes arbitrarily inefficient on arbitrarily large sets.

Furthermore, when \(p\) is large we can see that the full shrinkage estimator dominates the Hodges’ estimator. As shown in (41)
\[
\limsup_{p \to \infty} \sup_{h \in H(c)} \mathcal{P}(\tilde{\theta}_n^\ast, h) \leq \frac{c}{c + 1} < c = \sup_{h \in H(c)} \mathcal{P}(\tilde{\theta}_n^H, h).
\]

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.

For compactness, define the quadratic loss function \(\ell(u) = u' W u\) so that the loss of an esti-
mator $T_n$ for $\theta$ is $\ell(T_n - \theta)$. Local asymptotic efficiency theory defines the asymptotic maximum risk of a sequence of estimators $T_n$ for $\theta_n = \theta_0 + n^{-1/2}h$ with arbitrary $h$ as

$$
\sup_{I \subseteq \mathbb{R}^m} \lim_{n \to \infty} \inf_{\theta \in I} \sup_{h \in I} \mathbb{E}_{\theta_n} n \ell(T_n - \theta_n)
$$

(43)

where the first supremum is taken over all finite subsets $I$ of $\mathbb{R}^m$. The local asymptotic minimax theorem (e.g. Theorem 8.11 of van der Vaart (1998)) demonstrates that under quite mild regularity conditions the asymptotic uniform risk (43) is bounded below by that of the MLE. This demonstrates that no estimator has smaller asymptotic uniform risk than the MLE over unbounded $h$. This theorem is credited to Hájek (1970, 1972), though it builds 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 $T_n$ as

$$
\sup_{I \subseteq \mathcal{H}(c)} \lim_{n \to \infty} \inf_{\theta \in I} \sup_{h \in I} \mathbb{E}_{\theta_n} n \ell(T_n - \theta_n)
$$

(44)

which replaces the supremum over all subsets of $\mathbb{R}^m$ with the supremum over all finite subsets of $\mathcal{H}(c)$.

On a side note, in the case of full shrinkage ($p = m$) then (44) is equivalent to

$$
\lim_{n \to \infty} \sup_{h \in \mathcal{H}(c)} \mathbb{E}_{\theta_n} n \ell(T_n - \theta_n).
$$

(45)

The reason for the technical difference in the ordering of the supremum between (44) and (45) is that when there is partial shrinkage the set $\mathcal{H}(c)$ is unbounded in some directions, and the interior supremum must be taken only over a compact set. In the case of full shrinkage the set $\mathcal{H}(c)$ is compact so this complicated ordering of supremum is not necessary.

The standard method to establish the efficiency bound (43) 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 (44) we need to start with a similar bound for the normal sampling model. Unfortunately, we do not have a sharp bound for this case. An important breakthrough is Pinsker’s Theorem (Pinsker, 1980) which provides a sharp bound for the normal sampling model by taking $p \to \infty$. The existing theory has established the bound for the full shrinkage canonical model (e.g., $p = m$ and $W = V^{-1}$). 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 $Z \sim N_m(h, V)$ and $d_p > 8$, where $d_p$ is defined in (28). For any estimator
\( T = T(Z) \), and any \( 0 < c < \infty \),

\[
\sup_{h \in H(c)} \mathbb{E}_h \ell(T - h) \geq \text{tr}(WV) - \left[ \frac{1}{1 + c} + \left( \frac{2}{1 + c} + 4c \right) d_p^{-1/3} \right] \text{tr}(A). \tag{46}
\]

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

\[
\bar{\ell}(u) = \frac{\ell(u)}{\text{tr}(WV)}.
\]

Then, taking the limit as \( p \to \infty \) as in Theorem 3, the normalized version of the bound (46) simplifies to

\[
\liminf_{p \to \infty} \sup_{h \in H(c)} \mathbb{E}_h \bar{\ell}(T - h) \geq 1 - \frac{a}{c + 1}. \tag{47}
\]

We do not use (47) directly, but rather use (46) 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 \( H(c) \) and those which are unconstrained.

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

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

\[
\int \left[ f(x, \theta + h)^{1/2} - f(x, \theta)^{1/2} - \frac{1}{2} h' g(x, \theta) f(x, \theta)^{1/2} \right]^2 d\mu = o \left( \|h\|^2 \right), \quad h \to 0 \tag{48}
\]

where \( g(x, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta) \). Suppose that \( \mathcal{I}_\theta = \mathbb{E}_\theta g(X_i, \theta) g(X_i, \theta)' > 0 \), set \( V = \mathcal{I}_\theta^{-1}, A \) in (13), and \( H(c) \) in (32). Finally, suppose that \( d_p > 0 \), where \( d_p \) is defined in (28). Then for any sequence of estimators \( T_n \) on the sequence \( \theta_n = \theta + n^{-1/2} h \), where \( \theta \) is in the interior of \( \Theta \), and any \( 0 < c < \infty \),

\[
\sup_{I \subset H(c)} \liminf_{n \to \infty} \sup_{h \in I} \mathbb{E}_{\theta_n} n \ell(T_n - \theta_n) \geq \text{tr}(WV) - \left[ \frac{1}{1 + c} + \left( \frac{2}{1 + c} + 4c \right) d_p^{-1/3} \right] \text{tr}(A). \tag{49}
\]

Furthermore, suppose that as \( p \to \infty \), (37) and (38) hold. Then

\[
\liminf_{p \to \infty} \sup_{I \subset H(c)} \liminf_{n \to \infty} \sup_{h \in I} \mathbb{E}_{\theta_n} n \bar{\ell}(T_n - \theta_n) \geq 1 - \frac{a}{c + 1}. \tag{50}
\]

Theorem 5 provides a lower bound on the asymptotic local minimax risk for \( h \) in the ball \( H(c) \). (49) is the case of finite \( p \), and (50) shows that the bound takes a simple form when \( p \) is large. Since this lower bound is equal to the upper bound (40) attained by our shrinkage estimator, (50) is sharp. This proves that the shrinkage estimator is asymptotically minimax efficient over the local
sets $H(c)$. To our knowledge, Theorem 5 is new. It is the first large sample local efficiency bound for shrinkage estimation.

Differentiability in quadratic mean (48) is weaker than the requirements for asymptotic normality of the MLE. (See chapter 7 of van der Vaart (1998).)

Note that the equality of (40) and (50) holds for all values of $c$. This is a very strong efficiency property. It implies that the shrinkage estimator $\hat{\theta}^*$ is simultaneously asymptotically minimax over all $c > 0$.

Classic minimax theory (e.g. Theorem 8.11 of van der Vaart (1998)) applies to all bowl-shaped loss functions $\ell(u)$, not just quadratic loss, and thus it seems reasonable to conjecture that Theorems 4 & 5 will extend beyond quadratic loss. The challenge is that Pinsker’s theorem specifically exploits the structure of the quadratic loss, and thus it is unclear how to extend Theorem 4 to allow for other loss functions. Allowing for more general loss functions would be a useful extension.

Similarly to Theorem 3, a limitation of the bound (50) is the use of the sequential limits, first taking $m$ to infinity and then $n$ to infinity. A deeper result would employ joint limits.

9 Simulation

9.1 Binary Probit

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

\[
y_i = 1 (y_i^* \geq 0) \\
y_i^* = \theta_0 + X_{1i}^T \theta_1 + X_{2i}^T \theta_2 + e_i \\
e_i \sim N(0, 1).
\]

The regressors $X_{1i}$ and $X_{2i}$ are $k \times 1$ and $p \times 1$, respectively, with $p > k$. The regressor vector $X_i$ is distributed $N(0, \Sigma)$ where $\Sigma_{jj} = 1$ and $\Sigma_{jk} = \rho$ for $j \neq k$. The goal is estimation of $\theta = (\theta_0, \theta_1, \theta_2)$ under quadratic loss and the belief that $\theta_2$ may be close to the zero vector.

The regression coefficients are set as $\theta_0 = 0$, $\theta_1 = (b, b, ..., b)'$ and $\theta_2 = (c, ..., c)'$. Consequently, the control parameters of the model are $c$, $n$, $p$, $k$, $b$, and $\rho$. We found that the results were qualitatively insensitive to the choice of $k$, $b$, and $\rho$, 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 experiment with the alternative specification $\theta_2 = (c, 0, ..., 0)'$ (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. $\hat{\theta} =$ unrestricted MLE. Probit of $y_i$ on $(1, X_{1i}, X_{2i})$

2. $\tilde{\theta} =$ restricted MLE. Probit of $y_i$ on $(1, X_{1i})$
3. \( LR_n = 2 \left( \log L(\hat{\theta}) - \log L(\tilde{\theta}) \right) \), the likelihood ratio test for the restriction \( \theta_2 = 0 \)

4. \( \hat{V} = \text{estimate of the asymptotic covariance matrix of } \sqrt{n} (\hat{\theta} - \theta) \)

The restricted estimator is selected to reflect the belief that the coefficients on \( X_{2i} \) may be close to zero.

We compare five estimators. The first is \( \hat{\theta} \), the unrestricted MLE. The second is our shrinkage estimator using the weight matrix \( W = I_{k+p} \)

\[
\hat{\theta}^* = \hat{w}_n \hat{\theta} + (1 - \hat{w}_n) \tilde{\theta} \tag{51}
\]

\[
\hat{w}_n = \left( 1 - \frac{\text{tr}(A_n)(1-2/p)}{n \left( \hat{\theta} - \tilde{\theta} \right)' \left( \hat{\theta} - \tilde{\theta} \right)} \right) + D_n = n \left( \hat{\theta} - \tilde{\theta} \right)' \left( \hat{\theta} - \tilde{\theta} \right).
\]

The third is the pretest estimator

\[
\hat{\theta}_{PT} = \begin{cases} 
\hat{\theta} & \text{if } LR_n \geq q \\
\tilde{\theta} & \text{if } LR_n < q
\end{cases}
\]

where \( q \) is the 95% quantile of the \( \chi^2_p \) distribution. The fourth estimator is the weighted AIC estimator \( \hat{\theta}_{W,AIC} \) of Burnham and Anderson (2002), which is (51) with the weight

\[
\hat{w}_n = (1 + \exp(p - LR_n/2))^{-1}.
\]

The motivation is that the weight for each estimator is proportional to the exponential of negative one-half of each model’s AIC. The fifth estimator is an approximate Bayesian model averaging (BMA) estimator \( \hat{\theta}_{BMA} \), which is (51) with the weight

\[
\hat{w}_n = (1 + \exp(p \log(n)/2 - LR_n/2))^{-1}.
\]

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.

Our primary purpose is to compare the MLE \( \hat{\theta} \) with the shrinkage estimator \( \hat{\theta}^* \). The other three 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 \( \hat{\theta} \) was unreliable. For these cases we simply set all estimates

21
equal to the restricted estimate \( \hat{\theta} \). 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 \( \hat{\theta} \),

\[
MSE(\hat{\beta}) = E \left( (\hat{\theta} - \theta)'(\hat{\theta} - \theta) \right).
\]

We calculated the MSE by simulation using 10,000 replications. 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. To simplify the presentation, we normalize the MSE of all estimators by that of the unconstrained MLE. The normalized MSE are displayed as lines. The solid line is the normalized MSE of the 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.

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 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 \( y_{ji}, i = 1, \ldots, n \), be independent draws from the exponential distribution with mean \( \theta_j \), for \( j = 0, \ldots, p \). We consider estimation of \( \theta = (\theta_0, \theta_1, \ldots, \theta_p)' \) under quadratic loss with the belief that the coefficients may be close to one another, that is, the shrinkage direction is \( \theta_0 = \theta_1 = \cdots = \theta_p \).

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
Figure 2: Finite Sample Mean Squared Error, Probit Design
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, $\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} y_{ji}$. The associated log-likelihood is $\log \mathcal{L}(\hat{\theta}) = -n \sum_{j=0}^{p} \log \hat{\theta}_j - n(p + 1)$, and the estimate of the asymptotic covariance matrix is $\hat{V} = \text{diag}\left\{\hat{\theta}_0, ..., \hat{\theta}_p\right\}$.

The restricted MLE is $\tilde{\theta} = (p + 1)^{-1} \sum_{j=0}^{p} \hat{\theta}_j$ with associated log-likelihood $\log \mathcal{L}(\tilde{\theta}) = -n(p + 1)\left(\log \tilde{\theta} + 1\right)$. The restriction is linear, with linear constraint matrix (4).

We compare the same estimators as the previous simulation experiment: unrestricted MLE, shrinkage using the weight matrix $W = I_{p+1}$, pretest, weighted AIC, and BMA. 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). (18), (19), and (20) follow by the continuous mapping theorem. ■

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

**Lemma 1** If \( Z \sim N(0, \Sigma) \) is \( m \times 1 \), \( K \) is \( m \times m \), and \( \eta(x) : \mathbb{R}^m \rightarrow \mathbb{R}^m \) is absolutely continuous, then
\[
\mathbb{E} \left( \eta(Z + h)' KZ \right) = \mathbb{E} \operatorname{tr} \left( \frac{\partial}{\partial x} \eta(Z + h)' KV \right).
\]

**Proof:** Let \( \phi_V(x) \) denote the \( N(0, \Sigma) \) density function. By multivariate integration by parts
\[
\mathbb{E} \left( \eta(Z + h)' KZ \right) = \int \eta(x + h)' KVV^{-1} x \phi_V(x) \, (dx)
= \int \operatorname{tr} \left( \frac{\partial}{\partial x} \eta(x + h)' KV \right) \phi_V(x) \, (dx)
= \mathbb{E} \operatorname{tr} \left( \frac{\partial}{\partial x} \eta(Z + h)' KV \right).
\]

**Proof of Theorem 2:** Observe that \( \sqrt{n} \left( \hat{\theta}_n - \theta_n \right) \xrightarrow{\theta_n} Z \sim N(0, \Sigma) \) under (16). Then (26) shows that
\[
\rho(\hat{\theta}, h) = \mathbb{E} \left( Z' W Z \right) = \operatorname{tr} \left( W \mathbb{E} \left( ZZ' \right) \right) = \operatorname{tr} (W \Sigma) \cdot (52)
\]

Next, \( \sqrt{n} \left( \hat{\theta}_n^* - \theta_n \right) \xrightarrow{\theta_n} \psi \), where \( \psi \) is the random variable shown in (20). The variable \( \psi \) has a classic James-Stein distribution with positive-part trimming. Define the analogous random variable without positive part trimming
\[
\psi^* = Z - \left( \frac{\tau_p}{(Z + h)' B (Z + h)} \right) V R (R' V R)^{-1} R' (Z + h) \cdot (53)
\]

Then using (26) and the fact that the pointwise quadratic risk of \( \psi \) is strictly smaller than that of \( \psi^* \) (as shown, for example, by Lemma 2 of Hansen (2014)),
\[
\rho(\hat{\theta}_n^*, h) = \mathbb{E} \left( \psi' W \psi \right) < \mathbb{E} \left( \psi^{*'} W \psi^* \right) \cdot (54)
\]
Using (53), we calculate that (54) equals
\[ \mathbb{E}(Z'WZ) + \tau_p^2 \mathbb{E} \left( \frac{(Z + h)' R (R'VR)^{-1} R'VWVR (R'VR)^{-1} R' (Z + h)}{(Z + h)' B (Z + h)} \right) \]
\[ - 2 \tau_p \mathbb{E} \left( \frac{(Z + h)' R (R'VR)^{-1} R'VWZ}{(Z + h)' B (Z + h)} \right) \]
\[ = \text{tr}(WV) + \tau_p^2 \mathbb{E} \left( \frac{1}{(Z + h)' B (Z + h)} \right) \]
\[ - 2 \tau_p \mathbb{E} \left( \eta(Z + h)' R (R'VR)^{-1} R'VWZ \right) \]
\[ = \text{tr}(WV) + \tau_p^2 \mathbb{E} \left( \frac{1}{(Z + h)' B (Z + h)} \right) \]
\[ - 2 \tau_p \mathbb{E} \left( \eta(Z + h)' R (R'VR)^{-1} R'VWZ \right) \]
(55)

where
\[ \eta(x) = \left( \frac{1}{x'Bx} \right) x. \]

Since
\[ \frac{\partial}{\partial x} \eta(x)' = \left( \frac{1}{x'Bx} \right) I - \frac{2}{(x'Bx)^2} Bxx', \]
then by Lemma 1 (Stein’s Lemma)
\[ \mathbb{E} \left( \eta(Z + h)' R (R'VR)^{-1} R'VWZ \right) = \mathbb{E} \text{tr} \left( \frac{\partial}{\partial x} \eta(Z + h)' R (R'VR)^{-1} R'VWV \right) \]
\[ = \mathbb{E} \text{tr} \left( \frac{R (R'VR)^{-1} R'VWV}{(Z + h)' B (Z + h)} \right) \]
\[ - 2 \mathbb{E} \text{tr} \left( \frac{B (Z + h) (Z + h)' R (R'VR)^{-1} R'VWV}{((Z + h)' B (Z + h))^2} \right). \]
(56)

Using (21), define \( B_1 = W^{1/2} VR (R'VR)^{-1} R' \) and \( A^* = W^{1/2} VR (R'VR)^{-1} R'VW^{1/2} \). Note that \( R (R'VR)^{-1} R'VWVB = B_1 A^* B_1, B_1 B_1 = B \), and
\[ \lambda_{\text{max}}(A^*) = \lambda_{\text{max}} \left( (R'VR)^{-1/2} R'VWVR (R'VR)^{-1/2} \right) \]
\[ = \lambda_{\text{max}} \left( (R'VR)^{-1} R'VWVR \right) \]
\[ = \lambda_1. \]

Using the inequality \( b'C b \leq (b'b) \lambda_{\text{max}}(C) \) for symmetric \( C \),
\[ \text{tr} \left( B (Z + h) (Z + h)' R (R'VR)^{-1} R'VWV \right) \]
\[ = (Z + h)' B_1 A^* B_1 (Z + h) \]
\[ \leq (Z + h)' B (Z + h) \lambda_1. \]
Thus the two terms on the right-hand-side of (56) are larger than
\[\mathbb{E}\left( \frac{\text{tr}(A) - 2\lambda_1}{(Z + h)'B(Z + h)} \right)\]
which means that (55) is smaller than
\[
\text{tr}(WV) - \tau_p \mathbb{E}\left( \frac{2(\text{tr}(A) - 2\lambda_1) - \tau_p}{(Z + h)'B(Z + h)} \right) \leq \text{tr}(WV) - \tau_p \mathbb{E}\left( \frac{2(\text{tr}(A) - 2\lambda_1) - \tau_p}{(Z + h)'B(Z + h)} \right)
\]
where the inequality is Jensen’s and uses the assumption that \(\tau_p \leq 2(\text{tr}(A) - 2\lambda_1)\).

We calculate that
\[
\mathbb{E}\left( (Z + h)'B(Z + h) \right) = h'Bh + \text{tr}(BV) \\
= h'Bh + \text{tr}(A) \\
\leq (c + 1) \text{tr}(A)
\]
where the inequality is for \(h \in H(c)\). Substituted into (57) we have established (33). As this bound is strictly less than \(\text{tr}(WV)\) for any \(c < \infty\), combined with (52) we have established (31). 

**Proof of Corollary 1.** We need to verify that (29)-(30) hold for (12) under \(p \geq 3\) and (34). First, \(p \geq 3\) and (34) imply \(d_p \geq 12/5 > 2\) so (29) holds. Second, (34) implies
\[
1 - \frac{2}{p} \leq 2 \left(1 - \frac{2}{d_p}\right)
\]
so
\[
\tau_p = \text{tr}(A) \left(1 - \frac{2}{p}\right) \leq 2 \text{tr}(A) \left(1 - \frac{2}{d_p}\right) = 2(\text{tr}(A) - 2\lambda_1)
\]
which is (30). 

**Proof of Theorem 3.** From (33) and the definition (36)
\[
\sup_{h \in H(c)} p(\hat{\theta}^*, h) \leq 1 - \frac{\text{tr}(A)}{\text{tr}(WV)} \frac{\tau_p}{\text{tr}(A)} \left( \frac{2(1 - 2/d_p) - \tau_p}{\text{tr}(A)} \right) (c + 1) \rightarrow 1 - \frac{a}{c + 1}
\]
as \(p \rightarrow \infty\), as stated. 

**Proof of Theorem 4.** Without loss of generality we can set \(V = I_m\) and \(R = \left( \begin{array}{c} 0_{m-p} \\ I_p \end{array} \right)\). To see this, start by making the transformations \(h \rightarrow V^{-1/2}h\), \(R \rightarrow V^{1/2}R\), and \(W \rightarrow V^{1/2}WV^{1/2}\)
so that \( V = I_m \). Then write \( R = Q \begin{pmatrix} 0_{m-p} \\ I_p \end{pmatrix} G \) where \( QQ' = I_p \) and \( G \) is full rank. Make the transformations \( h \mapsto Q'h, \ R \mapsto Q'RG^{-1} \) and \( W \mapsto QWQ' \). Hence \( V = I_m \) and \( R = \begin{pmatrix} 0_{m-p} \\ I_p \end{pmatrix} \) as claimed.

Partition \( h = (h_1, h_2) \), \( T = (T_1, T_2) \), \( Z = (Z_1, Z_2) \) and \( W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \) conformably with \( R \). Note that after these transformations \( \mathbf{A} = W_{22} \) and \( H(c) = \{ h : h_2'W_{22}h_2 \leq \text{tr}(W_{22})c \} \).

Set \( \eta = 1 - 2d_p^{-1/3} \) and note that \( 0 < \eta < 1 \) since \( d_p > 8 \). Fix \( \omega > 0 \). Let \( \Pi_1(h_1) \) and \( \Pi_2(h_2) \) be the independent priors \( h_1 \sim N(0, I_{m-p}\omega) \) and \( h_2 \sim N(0, I_p\eta) \). Let \( \tilde{T}_1 = \mathbb{E}(h_1 | Z) \) and \( \tilde{T}_2 = \mathbb{E}(h_2 | Z) \) be the Bayes estimators of \( h_1 \) and \( h_2 \) under these priors. By standard calculations, \( \tilde{T}_1 = \frac{\omega}{1 + \omega}Z_1 \) and \( \tilde{T}_2 = \frac{\eta}{1 + \eta}Z_2 \). Also, let \( \Pi_2^*(h_2) \) be the prior \( \Pi_2(h_2) \) truncated to the region \( \mathbf{H}_2(c) = \{ h_2 : h_2'W_{22}h_2 \leq \text{tr}(W_{22})c \} \), and let \( \tilde{T}_2^* = \mathbb{E}(h_2 | Z) \) be the Bayes estimator of \( h_2 \) under this truncated prior. Since a Bayes estimator must lie in the prior support, it follows that \( \tilde{T}_2^* \in \mathbf{H}_2(c) \) or

\[
\tilde{T}_2^*W_{22}\tilde{T}_2^* \leq \text{tr}(W_{22})c.
\] (58)

Also, since \( Z_1 \) and \( Z_2 \) are independent, and \( \Pi_1 \) and \( \Pi_2^* \) are independent, it follows that \( \tilde{T}_2^* \) is a function of \( Z_2 \) only, and \( \tilde{T}_1 - h_1 \) and \( \tilde{T}_2^* - h_2 \) are independent.

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

$$\sup_{h \in H(c)} \mathbb{E}_h \ell(T - h) \geq \int \int \mathbb{E}_h \ell(T - h) d\Pi_1(h_1) d\Pi_2^*(h_2)$$

$$\geq \int \int \mathbb{E}_h \ell(T - h) d\Pi_1(h_1) d\Pi_2^*(h_2)$$

$$= \int \int \mathbb{E}_h \left[ (\bar{T}_1 - h_1)' W_{11} (\bar{T}_1 - h_1) \right] d\Pi_1(h_1) d\Pi_2^*(h_2)$$

$$+ 2 \int \int \mathbb{E}_h \left[ (\bar{T}_1 - h_1)' W_{12} (\bar{T}_2^* - h_2) \right] d\Pi_1(h_1) d\Pi_2^*(h_2)$$

$$+ \int \int \mathbb{E}_h \left[ (\bar{T}_2^* - h_2)' W_{22} (\bar{T}_2^* - h_2) \right] d\Pi_1(h_1) d\Pi_2^*(h_2)$$

$$= \int \mathbb{E}_h \left[ (\bar{T}_1 - h_1)' W_{11} (\bar{T}_1 - h_1) \right] d\Pi_1(h_1)$$

$$+ 2 \left( \int \mathbb{E}_h (\bar{T}_1 - h_1) d\Pi_1(h_1) \right)' W_{12} \left( \int (\bar{T}_2^* - h_2) d\Pi_2^*(h_2) \right)$$

$$+ \int \mathbb{E}_h \left[ (\bar{T}_2^* - h_2)' W_{22} (\bar{T}_2^* - h_2) \right] d\Pi_2(h_2)$$

$$+ \frac{\int_{H_2(c)^c} \mathbb{E}_h \left[ (\bar{T}_2^* - h_2)' W_{22} (\bar{T}_2^* - h_2) \right] d\Pi_2(h_2)}{\int_{H_2(c)} d\Pi_2(h_2)}$$

$$- \frac{\int_{H_2(c)} \mathbb{E}_h \left[ (\bar{T}_2^* - h_2)' W_{22} (\bar{T}_2^* - h_2) \right] d\Pi_2(h_2)}{\int_{H_2(c)} d\Pi_2(h_2)}$$

where the second inequality is because the Bayes estimator $\bar{T}$ minimizes the right-hand-side of (59). The final equality uses the fact that $\bar{T}_1 - h_1$ and $\bar{T}_2^* - h_2$ are independent, and breaks the integral (60) over the truncated prior (which has support on $H_2(c)$) into the difference of the integrals over the non-truncated prior over the $\mathbb{R}^m$ and $H_2(c)^c$, respectively. We now treat the four components (61)-(64) separately.

First, since $\bar{T}_1 = \frac{\omega}{1 + \omega} Z_1$ and $\Pi_1(h_1) = N(0, I_{m-\mu})$, we calculate that

$$\int \mathbb{E}_h \left[ (\bar{T}_1 - h_1)' W_{11} (\bar{T}_1 - h_1) \right] d\Pi_1(h_1)$$

$$= \int \mathbb{E}_h \left[ (\frac{\omega}{1 + \omega} Z_1 - h_1)' W_{11} \left( \frac{\omega}{1 + \omega} Z_1 - h_1 \right) \right] d\Pi_1(h_1)$$

$$= \int \left[ \frac{1}{(1 + \omega)^2} h_1' W_{11} h_1 + \frac{\omega^2}{(1 + \omega)^2} \text{tr}(W_{11}) \right] d\Pi_1(h_1)$$

$$= \text{tr}(W_{11}) \frac{\omega}{1 + \omega}.$$  (65)

Second, since

$$\int \mathbb{E} (\bar{T}_1 - h_1) d\Pi_1(h_1) = -\frac{1}{1 + \omega} \int h_1 d\Pi_1(h_1) = 0$$

it follows that (62) equals zero.
Third, take (63). Because $\tilde{T}_2$ is the Bayes estimator under the prior $\Pi_2$,

$$
\int \mathbb{E} \left[ \left( \tilde{T}_2^* - h_2 \right)' W_{22} \left( \tilde{T}_2^* - h_2 \right) \right] d\Pi_2(h_2)
\int_{H_2(c)} d\Pi_2(h_2)
\geq \int \mathbb{E} \left[ \left( T_2 - h_2 \right)' W_{22} \left( T_2 - h_2 \right) \right] d\Pi_2(h_2)
\geq \int \mathbb{E} \left[ \left( T_2 - h_2 \right)' W_{22} \left( T_2 - h_2 \right) \right] d\Pi_2(h_2)
= \text{tr} (W_{22}) \frac{c\eta}{1 + c\eta}
(66)
= \text{tr} (W_{22}) \left( \frac{c}{1 + c} - \frac{2d_p^{-1/3}}{1 + c} \right)
(67)
$$

where (66) is a calculation similar to (65) using $\tilde{T}_2 = \frac{c\eta}{1 + c\eta} Z_2$ and $h_2 \sim N(0, I_p \cdot c\eta)$. (67) makes a simple expansion using $\eta = 1 - 2d_p^{-1/3}$.

Fourth, take (64). Our goal is to show that this term is negligible for large $p$, and our argument is based on the proof of Theorem 7.28 from Wasserman (2006). Set

$$
q = \frac{h_2' W_{22} h_2}{c \text{tr} (W_{22})}.
$$

Since $h_2 \sim N(0, I_p \cdot c\eta)$ we see that $\mathbb{E} q = \eta$. Use $(a + b)' (a + b) \leq 2a'a + 2b'b$ and (58) to find that

$$
\mathbb{E} \left[ \left( \tilde{T}_2^* - h_2 \right)' W_{22} \left( \tilde{T}_2^* - h_2 \right) \right] \leq 2 \mathbb{E} \left( \tilde{T}_2^* W_{22} \tilde{T}_2^* \right) + 2h_2' W_{22} h_2
\leq 2 \text{tr} (W_{22}) c + 2h_2' W_{22} h_2
= 2 \text{tr} (W_{22}) c (1 + q)
\leq 2 \text{tr} (W_{22}) c (2 + q - \eta).
(68)
$$

Note that $h_2 \in H_2(c)^c$ is equivalent to $q > 1$. Using (68) and the Cauchy-Schwarz inequality,

$$
\int_{H_2(c)^c} \mathbb{E}_h \left[ \left( \tilde{T}_2^* - h_2 \right)' W_{22} \left( \tilde{T}_2^* - h_2 \right) \right] d\Pi_2(h_2)
\leq 2 \text{tr} (W_{22}) c \left[ 2 \int_{H_2(c)^c} d\Pi_2(h_2) + \int_{H_2(c)^c} (q - \eta) d\Pi_2(h_2) \right]
\leq 2 \text{tr} (W_{22}) c \left[ 2\mathbb{P} (q > 1) + \text{var}(q)^{1/2} \mathbb{P} (q > 1)^{1/2} \right].
(69)
$$

31
Letting \( \mu \) denote the eigenvalues of \( W_{22} \) then we can write

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

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

\[
\text{var}(q) = \frac{\eta^2}{\left( \sum_{j=1}^{p} w_j \right)^2} \sum_{j=1}^{p} w_j^2 \text{var}(y_j^2) \leq 2d_p^{-1}
\]

(70)

since \( d_p = \frac{\sum_{j=1}^{p} w_j}{\max_j w_j} = \text{tr}(W_{22}) / \lambda_{\text{max}}(W_{22}) \). By Markov’s inequality, (70), and \( 1 - \eta = 2d_p^{-1/3} \),

\[
\mathbb{P}(q > 1) = \mathbb{P}(q > 1 - \eta) \leq \frac{\text{var}(q)}{(1 - \eta)^2} \leq \frac{d_p^{-1/3}}{2}.
\]

(71)

Furthermore, (71) and \( d_p^{-1/3} \leq 2^{-1} \) imply that

\[
\int_{H_2(c)} d\Pi_2(h_2) = 1 - \mathbb{P}(q > 1) \\
\geq 1 - \frac{d_p^{-1/3}}{2} \\
\geq \frac{3}{4}.
\]

(72)

It follows from (69), (70), (71), (72) and \( d_p^{-1/3} \leq 2^{-1} \) that

\[
\frac{\int_{H_2(c)} \mathbb{E}_h \left[ (\bar{T}_2^* - h_2)' W_{22} (\bar{T}_2^* - h_2) \right] d\Pi_2(h_2)}{\int_{H_2(c)} d\Pi_2(h_2)} \\
\leq \text{tr}(W_{22}) \frac{2c \left( d_p^{-1/3} + d_p^{-2/3} \right)}{3/4} \\
\leq \text{tr}(W_{22}) 4cd_p^{-1/3}.
\]

(73)

Together, (65) and (73) applied to (61)-(63) show that

\[
\sup_{h \in H(c)} \mathbb{E}_h \ell(T - h) \geq \frac{\omega}{1 + \omega} \text{tr}(W_{11}) + \left( \frac{c}{1 + c} - \left( \frac{2}{1 + c} + 4c \right) d_p^{-1/3} \right) \text{tr}(W_{22}).
\]
Since \( \omega \) is arbitrary we conclude that
\[
\sup_{h \in H(c)} \mathbb{E}_h \ell (T - h) \geq \text{tr} (W_{11}) + \left( \frac{c}{1 + c} - \left( \frac{2}{1 + c} + 4c \right) d_p^{-1/3} \right) \text{tr} (W_{22})
\]
\[
= \text{tr} (W) - \left( \frac{1}{1 + c} + \left( \frac{2}{1 + c} + 4c \right) d_p^{-1/3} \right) \text{tr} (W_{22})
\]
which is (46) since \( \text{tr} (W_{22}) = \text{tr} (A) \) and \( \text{tr} (W) = \text{tr} (WV) \) under the transformations made at the beginning of the proof.

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

**Proof of Theorem 5.** The proof technique is based on the arguments in Theorem 8.11 of van der Vaart (1998), with two important differences. First, van der Vaart (1998) appeals to a compactification argument from Theorem 3.11.5 of Van der Vaart and Wellner (1996), while we use a different argument which allows for possibly singular \( W \). Second, we bound the risk of the limiting experiment using Theorem 4 rather than van der Vaart’s Proposition 8.6.

Let \( Q(c) \) denote the rational vectors in \( H(c) \) placed in arbitrary order, and let \( Q_k \) denote the first \( k \) vectors in this sequence. Define \( Z_n = \sqrt{n} (T_n - \theta_n) \). There exists a subsequence \( \{n_k\} \) of \( \{n\} \) such that
\[
\sup_{I \subseteq H(c)} \liminf_{n \to \infty} \sup_{h \in I} \mathbb{E}_{\theta_n} \ell (Z_n) \geq \lim_{k \to \infty} \liminf_{n \to \infty} \sup_{h \in Q_k} \mathbb{E}_{\theta_n} \ell (Z_n)
\]
\[
= \lim_{k \to \infty} \sup_{h \in Q_k} \mathbb{E}_{\theta_{n_k}} \ell (Z_{n_k})
\]
\[
\geq \lim_{k \to \infty} \sup_{h \in Q_K} \mathbb{E}_{\theta_{n_k}} \ell (Z_{n_k})
\]
the final inequality for any \( K < \infty \).

Since we allow \( W \) to have rank \( r \leq m \), write \( W = G_1 G'_1 \) where \( G_1 \) is \( m \times r \) with rank \( r \). Set \( G = [G_1, G_2] \) where \( G_2 \) is \( m \times (m - r) \) with rank \( m - r \) and \( G'_1 G_2 = 0 \). Define
\[
Z^*_n = G^{-1} \left( \begin{array}{c} G'_1 Z_n \\ 0_{m-r} \end{array} \right)
\]
which replaces the linear combinations \( G'_2 Z_n \) with zeros. Notice that since the loss function is a quadratic in \( W = G_1 G'_1 \), then \( \ell (Z_n) = \ell (Z^*_n) \).

We next show that without loss of generality we can assume that \( Z^*_n \) is uniformly tight on a subsequence \( \{n'_k\} \) of \( \{n_k\} \). Suppose not. Then there exists some \( \varepsilon > 0 \) such that for any \( \zeta < \infty \),
\[
\liminf_{k \to \infty} P \left( Z^*_{n_k} G G' Z^*_{n_k} > \zeta \right) \geq \varepsilon.
\]
Set $\zeta = \operatorname{tr}(WWW) / \varepsilon$. Since $\ell(Z_n^*) = Z_n^\prime G G^\prime Z_n^*$, (75) implies

$$\liminf_{k \to \infty} \mathbb{E}_{\theta_k} \ell(Z_{n_k}^*) = \liminf_{k \to \infty} \mathbb{E}_{\theta_k} Z_{n_k}^\prime G G^\prime Z_{n_k}^* \geq \zeta = \operatorname{tr}(WWW)$$

which is larger than (49). Thus for the remainder we assume that $Z_n^*$ is uniformly tight on a subsequence $\{n'_k\}$ of $\{n_k\}$.

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

$$\liminf_{k \to \infty} \mathbb{E}_{\theta_{n_k}} \ell(Z_{n_k}^*) \geq \mathbb{E}_h \ell(T(Z) - h).$$

Combined with (74), the fact that the set $Q_K$ is finite, and $\ell(Z_n) = \ell(Z_n^*)$, we find that

$$\sup_{I \subset \mathcal{H}(c)} \inf_{n \to \infty} \sup_{h \in I} \mathbb{E}_{\theta_n} \ell(Z_n) \geq \sup_{h \in Q_K} \mathbb{E}_h \ell(T(Z) - h).$$

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

$$\sup_{I \subset \mathcal{H}(c)} \inf_{n \to \infty} \sup_{h \in I} \mathbb{E}_{\theta_n} \ell(Z_n) \geq \sup_{h \in \mathcal{Q}(c)} \mathbb{E}_h \ell(T(Z) - h)$$

$$= \sup_{h \in \mathcal{H}(c)} \mathbb{E}_h \ell(T(Z) - h)$$

$$\geq \operatorname{tr}(WWW) - \left[ \frac{1}{1 + c} + \left( \frac{2}{1 + c} + 4c \right) d_p^{-1/3} \right] \operatorname{tr}(A)$$

the final inequality by Theorem 4. We have shown (49).

Dividing by $\operatorname{tr}(WWW)$, and taking the limit as $p \to \infty$, we obtain

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

which is (50).
References


