Abstract

This paper introduces shrinkage for general econometric estimators satisfying a central limit theorem. We show how to shrink arbitrary estimators towards parameter subspaces defined by general nonlinear restrictions. Our simplest shrinkage estimators are functions only of the unconstrained estimator and its estimated asymptotic covariance matrix. Using a local asymptotic framework, we derive the asymptotic distribution of the generalized shrinkage estimator, and derive its asymptotic risk. We show that if the shrinkage dimension is three or larger, the asymptotic risk of the shrinkage estimator is strictly less than that of the unconstrained estimator. This reduction holds globally in the parameter and distribution space. We show that the reduction in asymptotic risk is substantial, even for moderately large values of the parameters.

Our results are quite broad, allowing for fairly general restricted estimators, and arbitrary weight matrices for the risk function.

We investigate shrinkage estimation when the parameters of interest are a strict subset of the general parameter vector, implying a risk function with a weight matrix of deficient rank. We show how to construct generalized shrinkage estimators in this context, and that they inherit the globally risk reduction property of the general setting.

This paper is currently incomplete. Work to be completed includes studies of least-squares estimation, forecasting, and specific forms of shrinkage estimators appropriate for MLE and GMM estimation.
1 Introduction

The classic James-Stein shrinkage estimator takes the following form. Suppose we have an estimator $\hat{\theta}_n$ for $\theta_n \in \Theta \subset \mathbb{R}^k$ which has the exact distribution $\hat{\theta}_n \sim N(\theta_n, V)$. The positive-part James-Stein estimator for $\theta_n$ is

$$\hat{\theta}_n^* = \left(1 - \frac{k - 2}{\hat{\theta}'_n V^{-1} \hat{\theta}_n} \right) \hat{\theta}_n$$

(1)

where $(a)_+ = a1 (a \geq 0)$. It is well known that when $k \geq 3$ the estimator $\hat{\theta}_n^*$ has smaller risk than $\hat{\theta}_n$, for a class of loss functions which includes weighted squared error. This result is finite-sample so is effectively restricted for econometric applications to the classic Gaussian regression model with exogenous regressors.

This paper extends the estimator (1) to include general econometric estimators satisfying a central limit theorem. To our knowledge, this is new. Perhaps the barrier to developing an asymptotic distribution theory has that under fixed parameter values $\theta_n = \theta$, the asymptotic distribution of $\hat{\theta}_n^*$ is discontinuous at $\theta = 0$, and there is no different between the “asymptotic” risk of the shrinkage estimator $\hat{\theta}_n^*$ and the unconstrained estimator $\hat{\theta}_n$. The error in this analysis is the asymptotic framework. To eliminate a discontinuity in an asymptotic distribution, a solution is to reparameterize the model as an array so that the parameter is local to the discontinuity, thereby attaining a continuous asymptotic distribution. The classic example of this method is Pitman drift for the study of test power, a modern example is the limit of experiments theory. For shrinkage towards the zero vector, our solution is to set $\theta_n = \theta_1$. With this reparameterization, it is straightforward to derive an asymptotic distribution for $\hat{\theta}_n^*$ as $n \to \infty$ which is continuous in $\delta$. In fact, we find that the normalized asymptotic distribution is identical to the finite sample distribution of the James-Stein estimator under exact normality. It thereby follows from classic results that the asymptotic risk (using weighted squared error loss) of $\hat{\theta}_n^*$ is strictly less than that of the original estimator $\hat{\theta}_n$, and that this result holds globally in the parameter space. This result extends classic shrinkage to econometric estimators which satisfy a central limit theorem. While mathematically simple, this extension is new and powerful.

To develop feasible econometric estimation methods, we make a number of extensions to this basic result. First, we allow for shrinkage towards any parameter subspace defined by a nonlinear restriction. Second, shrinkage towards a subspace requires a restricted parameter estimator. We allow for both efficient and inefficient restricted estimators, to incorporate potential applications of interest (e.g. least squares). Third, we focus on the positive-part shrinkage method. Fourth, we allow for weighted mean-square loss with arbitrary weight matrices. By incorporating the weight matrix into the construction of the shrinkage estimator we can obtain global risk improvements. Fifth, we provide an explicit formula for the shrinkage constant which ensures that the estimator has globally reduced risk relative to the unconstrained estimator. Sixth, we show that the risk is locally robust to misspecification of the risk weight matrix. Seventh, we investigate shrinkage estimation
when the parameters of interest are a strict subset of $\theta_n$. Mathematically this is important because it implies a rank-deficient risk weight matrix. Practically it is important because it allows researchers to focus on parameters of interest and to be explicit about the distinction between parameters of interest and nuisance parameters.

The literature on shrinkage estimation in enormous. We mention some of the most relevant contribution. Stein (1956) first observed that an unconstrained Gaussian estimator is inadmissible when $k \geq 3$. 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. Hjort and Claeskens (2003) showed that the local asymptotic framework is appropriate for the asymptotic analysis of averaging estimators. An in-depth treatment of shrinkage theory can be found in Chapter 5 of Lehmann and Casella (1998).

The organization of the paper is as follows. Section 2 presents the general framework and the generalized shrinkage estimator. Section 3 presents the main results – the asymptotic distribution of the estimator and its asymptotic risk. Section 4 uses a large-parameter approximation to the percentage risk improvement due to shrinkage, showing that the gains are substantial and broad in the parameters space. Section 5 examines shrinkage in the presence of nuisance parameters. Section 6 and beyond are yet to be written.

2 General Shrinkage Estimator

Suppose that there is an estimator $\hat{\theta}_n$ for a parameter $\theta_n \in \Theta \subset \mathbb{R}^k$. We call $\hat{\theta}_n$ the unconstrained estimator and assume that it is asymptotically normally distributed, and there is a consistent estimator for its asymptotic covariance matrix.

Assumption 1 $\text{As } n \to \infty$, 

1. $\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} \mathcal{N}(0, V)$
2. $\hat{V}_n \xrightarrow{p} V$

Next, suppose that there is a restricted parameter space defined by a restriction:

$$\Theta_0 = \{\theta \in \Theta : h(\theta) = 0\}$$

where $h(\theta) : \mathbb{R}^k \to \mathbb{R}^r$. Our goal is to shrink the unrestricted estimator $\hat{\theta}_n$ towards the subspace $\Theta_0$. The number of restrictions $r$ will play an important role in our analysis. We will call $r$ the “shrinkage dimension” as it equals the reduction in the parameter space induced by the restriction of $\Theta$ to $\Theta_0$. 
A traditional choice is to set $\Theta_0 = \{0\}$, the zero vector, so that $\hat{\theta}_n$ is shrunk towards zero and $r = k$, but it is often more sensible to shrink towards a sub-model of particular interest. The most common case is an exclusion restriction. If we partition

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

and the exclusion restriction imposes $\theta_2 = 0$, then $h(\theta) = H' \theta$ where

$$
H = \begin{pmatrix} 0 \\ I_r \end{pmatrix}.
$$

In general, a restricted estimator of $\theta_n$ can be found by constrained minimization:

$$
\tilde{\theta}_n = \arg\min_{\theta \in \Theta_n} \left( \theta - \hat{\theta}_n \right)' G_n^{-1} \left( \theta - \hat{\theta}_n \right)
$$

(3)

where $G_n > 0$ is $k \times k$. Efficient estimation suggests $G_n = \hat{V}_n$, but we allow for alternative choices to include important estimators of interest. The specific form of the estimator $\tilde{\theta}_n$ will not be important to our theory, all that is necessary is that it is first-order asymptotically equivalent to (3). One example is the linear approximation

$$
\tilde{\theta}_n = \hat{\theta}_n - G_n \hat{H} \left( \hat{H}' G_n \hat{H} \right)^{-1} h(\hat{\theta}_n)
$$

(4)

where

$$
\hat{H} = H(\hat{\theta}_n),
$$

$$
H(\theta) = \frac{\partial}{\partial \theta} h(\theta)'.
$$

When the restriction is linear $h(\theta) = H' \theta - b$ then (4) is exactly equal to (3).

Estimation performance (risk) will be assessed by asymptotic expected squared error loss. Specifically, for a symmetric and non-negative definite weight matrix $W$ the weighted asymptotic risk of an estimator $\bar{\theta}_n$ for $\theta_n$ is defined as

$$
R(\bar{\theta}_n, W) = \lim_{n \to \infty} n E \left( (\bar{\theta}_n - \theta_n)' W (\bar{\theta}_n - \theta_n) \right).
$$

It is easy to see that under Assumption 1

$$
R(\bar{\theta}_n, W) = \text{tr} (W V).
$$
Our default choice for $W$ is $V^{-1}$, as this renders the risk invariant to reparameterization and scaling. Accordingly, we define the agnostic asymptotic risk of an estimator $\hat{\theta}_n$ as
\[
R(\hat{\theta}_n) = R(\hat{\theta}_n, V^{-1}) = \lim_{n \to \infty} n \, E \left( (\hat{\theta}_n - \theta_n)' V^{-1} (\hat{\theta}_n - \theta_n) \right).
\]
We observe that
\[
R(\hat{\theta}_n) = k.
\]
In some cases, however, a specific choice for the weight matrix $W$ is desired so we allow for this possibility with the weighted risk $R(\hat{\theta}_n, W)$.

Our shrinkage estimator will be constructed as a function of an estimate $W_n$ of this weight matrix $W$. To assess robustness to this choice, we also evaluate the risk of the estimator constructed with $W$ using arbitrary alternative weight matrices $K$.

We are now ready to define our shrinkage estimators. Our generalized shrinkage estimator for $\theta_n$ is
\[
\hat{\theta}^*_n = \hat{\theta}_n - \left( \frac{C_n}{n \left( \hat{\theta}_n - \bar{\theta}_n \right)' W_n \left( \hat{\theta}_n - \bar{\theta}_n \right) + 1} \right) \left( \hat{\theta}_n - \bar{\theta}_n \right) \tag{5}
\]
\[
C_n = \text{tr} ( \hat{A}_n ) - 2\lambda_{\text{max}} ( \hat{A}_n ) \tag{6}
\]
\[
\hat{A}_n = (\hat{H}' \hat{G}_n \hat{H})^{-1} \left( \hat{H}' \hat{G}_n W_n \hat{V}_n \hat{H} \right) \tag{7}
\]
In the definition of (5)
\[
(a)_1 = 1 (a \geq 1) + a1 (a < 1)
\]
is a trimming function which trims its argument at one. Notice that $\hat{\theta}^*_n$ is a function of the matrices $G_n$, $W_n$ as well as the restriction $h(\theta)$.

The shrinkage constant $C_n$ defined in (6) is an estimate of
\[
C = \text{tr} (A) - 2\lambda_{\text{max}} (A) \tag{8}
\]
\[
A = (H'GH)^{-1} \left( H'GWVH \right) \tag{9}
\]
This plays an important role in the shrinkage theory. It is useful to notice that when the default weight matrix is used: $W_n = \hat{V}_n^{-1}$ then (9) simplifies to $\hat{A}_n = I_r$ and (6) and (8) simplify to $C_n = C = r - 2$. Thus our default generalized shrinkage estimator is
\[
\hat{\theta}^*_n = \hat{\theta}_n - \left( \frac{r - 2}{n \left( \hat{\theta}_n - \bar{\theta}_n \right)' \hat{V}_n^{-1} \left( \hat{\theta}_n - \bar{\theta}_n \right) + 1} \right) \left( \hat{\theta}_n - \bar{\theta}_n \right) \tag{10}
\]
It is also useful to note that the shrinkage estimator is a weighted function of the unconstrained and constrained estimators, with the weight depending on a quadratic in $\left( \hat{\theta}_n - \bar{\theta}_n \right)$. Note that in
the default case (10) with $G_n = \hat{V}_n$ this quadratic is

$$D_n = n \left( \hat{\theta}_n - \tilde{\theta}_n \right)' \hat{V}_n^{-1} \left( \hat{\theta}_n - \tilde{\theta}_n \right)$$

(11)

a chi-square statistic for testing the hypothesis $h(\theta_n) = 0$. Despite this dependence on $D_n$, it is important to understand that $\hat{\theta}_n^*$ is not a pre-test estimator as it is a continuous function of $D_n$.

3 Asymptotic Distribution and Risk

We impose the following conditions on the restriction and parameter space.

Assumption 2

1. $\theta_n = \theta_0 + n^{-1/2} \delta$ where $\delta \in \mathbb{R}^k$ and $h(\theta_0) = 0$

2. $H(\theta) = \frac{\partial}{\partial \theta} h(\theta)'$ is continuous in a neighborhood of $\theta_0$.

3. $G_n \overset{p}{\rightarrow} G > 0$

4. $W_n \overset{p}{\rightarrow} W$

5. $\text{rank} (H) = r \geq 3$ where $H = H(\theta_0)$.

6. $C = \text{tr} (A) - 2\lambda_{\max} (A) > 0$ where $A$ is defined in (9)

Assumption 2.1 is essential for our asymptotic theory. It states that the true parameter vector $\theta_n$ lies in a $n^{-1/2}$ neighborhood of the restricted parameter space $\Theta_0$. The local parameter $\delta$ measures the discrepancy between $\Theta_0$ and $\theta_n$. Assumption 2.1 which enables the derivation of an asymptotic distribution theory which is continuous in the parameters. We will show later that the distribution theory allows for very large values of the local parameter $\delta$ and thus should not be viewed as restrictive.

The rank conditions of Assumption 2.5 and 2.6 are also critical. The shrinkage dimension must be three or larger in order for shrinkage to uniformly dominate the unrestricted estimator (Stein, 1956). Assumption 2.6 reduces to $r \geq 3$ when $W = V^{-1}$ so only has separate relevance from Assumption 2.5 when $W \neq V^{-1}$.

We now present the asymptotic distribution and risk of the generalized shrinkage estimator $\hat{\theta}_n^*$ when the risk is assessed with an arbitrary weight matrix $K$.

**Theorem 1** Under Assumptions 1 and 2

$$\sqrt{n} \left( \hat{\theta}_n - \theta_n \right) \overset{d}{\rightarrow} Z,$$

$$\sqrt{n} \left( \hat{\theta}_n - \hat{\theta}_n \right) \overset{d}{\rightarrow} \text{GP} (Z + \delta),$$

(12)
and

\[
\sqrt{n}(\hat{\theta}_n^* - \theta_n) \xrightarrow{d} Z - \frac{C}{(Z + \delta)^{\frac{1}{2}}} \text{PGWGP}(Z + \delta),
\]

where \( Z \sim N(0, V) \) and

\[
P = H (H'GH)^{-1} H'.
\]

The asymptotic risk of the unrestricted and generalized shrinkage estimators are

\[
R(\hat{\theta}_n, K) = \text{tr}(KV)
\]

and

\[
R(\hat{\theta}_n^*, K) = \text{tr}(KV) - \text{E}g(Z + \delta)
\]

where

\[
g(x) = \frac{C}{x'W^*x} \left( 2\text{tr}(A^*) - 4 \frac{x'V^*x}{x'W^*x} - C \frac{x'K^*x}{x'W^*x} \right) 1\left( \frac{C}{x'W^*x} \leq 1 \right)
\]

\[
+ (2\text{tr}(A^*) - x'K^*x) 1\left( \frac{C}{x'W^*x} > 1 \right)
\]

where

\[
W^* = \text{PGWGP}
\]

\[
K^* = \text{PGKGP}
\]

\[
V^* = \text{PGKVPGWGP}
\]

\[
A^* = (H'GH)^{-1} (H'GKVH).
\]

Furthermore,

\[
R(\hat{\theta}_n^*, W) < R(\hat{\theta}_n, W).
\]

so the weighted risk of the generalized shrinkage estimator is strictly less than the unrestricted estimator.

Also, if

\[
H'GWGH > 0
\]

and

\[
C = \text{tr}(A) - 2\lambda_{\text{max}}(A) < 2 \left( \frac{\text{tr}(A^*) - 2\lambda_{\text{max}}(A^*)}{\lambda_{\text{max}}(A^{**})} \right)
\]

where

\[
A^{**} = (H'GWGH)^{-1} (H'GKGH),
\]

then

\[
R(\hat{\theta}_n^*, K) < R(\hat{\theta}_n, K).
\]
The main result is (17) which shows that the generalized shrinkage estimator has strictly smaller weighted asymptotic risk than the unconstrained estimator. This inequality holds globally in the parameter and distribution space. It holds for any restricted estimator \( \tilde{\theta}_n \) satisfying (3), regardless of the weight matrix \( G_n \). (This is surprising as the restricted estimator \( \tilde{\theta}_n \) does not have globally smaller risk than \( \hat{\theta}_n \) unless \( G = V^{-1} \).) It holds for any restriction \( h(\theta) \) satisfying Assumption 2. One main requirement is that the shrinkage dimension exceeds 2. Another is that the estimator (5) is constructed using the same weight matrix \( W \) as used to evaluate risk.

Theorem 1 also shows in (20) that the reduction in risk is robust to mild deviations in the weight matrix from \( W \). Inequality (20) shows that if an alternative weight matrix \( K \) is used to define risk, the generalized shrinkage estimator (5) will continue to have smaller risk than the unrestricted estimator if conditions (18) and (19) hold. The critical condition appears to be (19), which essentially states that \( K \) cannot be too different than \( W \). For the remainder of the paper we will assume \( K = W \), with the understanding that stated results are robust to modest departures of \( K \) from \( W \).

In addition, equation (13) expresses the asymptotic distribution of the generalized shrinkage estimator as a nonlinear function of a normal random vector. Equations (15)-(16) provide a general formula for its asymptotic risk, expressed as the expectation of the nonlinear function \( g(x) \).

4 Asymptotic Risk Reduction from Shrinkage

The classic (default) shrinkage estimator obtains when \( W = V^{-1} \) and \( G = V \). This is estimator (10). Its agnostic asymptotic risk is

\[
R(\tilde{\theta}_n^*) = k - \mathbb{E} g(Z + \delta)
\]

where

\[
g(x) = \frac{(r - 2)^2}{x'Px} \left( \frac{r - 2}{x'Px} \leq 1 \right) + (2r - x'Px) \left( \frac{r - 2}{x'Px} > 1 \right),
\]

with \( P = H (H'VH)^{-1} H' \).

It is also useful to observe that \( g(Z + \delta) \) depends on its argument only through the quadratic form \((Z + \delta)' P (Z + \delta)\) which has a non-central chi-square distribution with degrees of freedom \( r \) and non-centrality parameter

\[
\bar{\delta} = \delta' P \delta = \delta' H (H'VH)^{-1} H' \delta.
\]

It follows that for this estimator the agnostic asymptotic risk (15) is a function only of \( k, r, \) and \( \bar{\delta} \).

Furthermore, the ratio of the agnostic asymptotic risk of the generalized shrinkage estimator to the unrestricted estimator is

\[
\frac{R(\tilde{\theta}_n^*)}{R(\hat{\theta}_n)} = \frac{k - \mathbb{E} g(Z + \delta)}{k}.
\]

which simplifies for large \( r \). Casella and Hwang (1982) have shown that as \( r \to \infty \) and \( \bar{\delta}/r \to c \).
that
\[ \frac{\mathbb{E} g(Z + \delta)}{r} \to \frac{1}{1 + c} \]

It follows that if \( r/k \to \alpha \in (0, 1] \) then
\[ \frac{R(\hat{\theta}_n^*)}{R(\hat{\theta}_n)} \to 1 - \frac{\alpha}{1 + c}. \]

This shows that (approximately) the percentage risk improvement due to shrinkage is \( \alpha/(1 + c) \).

To understand the magnitude of the risk improvement due to shrinkage, consider the case \( \alpha = 1 \) (all parameters are shrunk) so that the percentage risk improvement is \( 1/(1 + c) \), where \( c = \delta/r \) and \( \delta \) is the non-centrality parameter in the asymptotic distribution of the statistic \( D_n \) defined in (11). Thus \( c = 1 \) when the non-centrality parameter \( \delta \) equals the degrees of freedom \( r \), \( c = 2 \) when \( \delta \) is twice the degrees of freedom, etc. It seems reasonable to consider values of \( c \) ranging from 0 through 4, with \( c = 0 \) representing the extreme case of a valid exclusion restriction, \( 1 \leq c \leq 2 \) representing typical or modest values of \( \delta \), and \( c = 4 \) representing a very large value of \( \delta \). In these cases, the percentage risk improvement ranges from 100\% (when \( c = 0 \)), to 50\% (\( c = 1 \)), to 33\% (\( c = 2 \)), to 20\% (\( c = 4 \)). For all these values, the reduction in asymptotic risk from shrinkage is substantial.

Another way to consider the situation is to examine the empirical value of the test statistic \( D_n \) relative to the degrees of freedom \( r \). Since \( D_n \overset{d}{\to} (Z + \delta)'P(Z + \delta) \) and \( E(Z + \delta)'P(Z + \delta) = r + \delta \), it follows that \( D_n/r - 1 \) is an asymptotically unbiased estimate of \( c \). Examining \( D_n \) therefore gives some information about the potential risk improvements due to shrinkage.

Strictly, this analysis is for the classic case \( W = V^{-1} \) and \( G = V \). Risk reduction due shrinkage is different for alternative choices for \( W \) and \( G \). However, the primary message is that substantial reductions in asymptotic risk are obtained, broadly in the parameter space.

5 Shrinkage in the Presence of Nuisance Parameters

It is quite common for the parameters of interest to be a strict subset of the parameter vector, and this suggest focusing on weight matrices \( W \) which reflect this interest. To be specific, partition
\[ \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \]
where \( \theta_1 \) are the parameters of interest (\( s \geq 3 \)) and \( \theta_2 \) are the nuisance parameters. In this case the loss function should only put weight on estimates of \( \theta_1 \), suggesting a weight matrix of the form
\[ W = RW_1R' \]
where
\[ R = \begin{pmatrix} I_s \\ 0 \end{pmatrix} \]  

(22)

and \( W_1 > 0 \) is \( s \times s \). From the previous section the generalized shrinkage estimator for \( \theta_1 \) is

\[ \hat{\theta}^*_1 = \hat{\theta}_1 - \frac{C_n}{n (\hat{\theta}_1 - \hat{\theta}_n)' R W_1 R' (\hat{\theta}_n - \hat{\theta}_1)} (\hat{\theta}_1 - \hat{\theta}_n) \]

\[ = \hat{\theta}_1 - \frac{C_n}{n (\hat{\theta}_1 - \hat{\theta}_1)' W_1 (\hat{\theta}_1 - \hat{\theta}_1)} (\hat{\theta}_1 - \hat{\theta}_1) \]

\[ C_n = \text{tr} (A_n) - 2\lambda_{\max} (A_n) \]

\[ A_n = (\tilde{H}' G_n \tilde{H})^{-1} (\tilde{H}' G_n R W_1 R' \tilde{V} R \tilde{H}) . \]

An agnostic choice for the \( s \times s \) weight matrix \( W_1 \) is one such that \( C_n \) takes a simple form. Consider the \( s \times s \) matrix

\[ V_R = R' G_n \tilde{H} (\tilde{H}' G_n \tilde{H})^{-1} \tilde{H}' G_n R \]

and notice that \( V_R \geq 0 \), and \( \text{rank} (V_R) \leq \min \{ r, s \} \). If \( r \geq s \) then \( V_R \) can be full rank but if \( r < s \) then it must have deficient rank. Therefore we recommend

\[ W_1 = (V_R)^+ \]

the Moore-Penrose generalized inverse \( V_R \). With this choice, then

\[ \text{tr} (A_n) = \text{tr} \left( (\tilde{H}' G_n \tilde{H})^{-1} (\tilde{H}' G_n R (V_R)^+ R' \tilde{V} \tilde{H}) \right) = \text{tr} (V_R (V_R)^+) \]

\[ = \text{rank} (V_R) \]

and

\[ \lambda_{\max} (A_n) = \lambda_{\max} (V_R (V_R)^+) = 1 \]

(For any matrix \( A \), \( AA^+ \) is idempotent (e.g. Magnus and Neudecker, 1988, p. 33) and thus \( \text{tr} (AA^+) = \text{rank} (V_R) \) and \( \lambda_{\max} (AA^+) = 1 \)). Thus \( C_n = \text{rank} (V_R) - 2 \). Assuming \( V_R = \min \{ r, s \} \) we obtain the shrinkage estimator

\[ \hat{\theta}^*_1 = \hat{\theta}_1 - \frac{\min \{ r, s \} - 2}{n (\hat{\theta}_1 - \hat{\theta}_1)' (V_R)^+ (\hat{\theta}_1 - \hat{\theta}_1)} (\hat{\theta}_1 - \hat{\theta}_1) . \]

Notice that by Theorem 1 a sufficient condition for this estimator to have smaller asymptotic risk...
than $\hat{\theta}_{1n}$ is rank $(V_R) = \min\{r, s\} > 2$.

This shows how to construct a shrinkage estimator when the parameter of interest is a strict subset of the entire parameter vector. The shrinkage estimator $\hat{\theta}_{1n}^*$ is a function only of the sub-estimates $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$, but the quadratic form determining the degree of shrinkage depends on the estimation method $G_n$, the direction of shrinkage $H$, and the sub-parameter selector matrix $R$.

From this analysis we can see a potential limitation. If $R$ and $H$ are orthogonal, then the rank of $V_R$ will be determined by the off-diagonal block of $\tilde{V}_n$. In particular, if $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ are asymptotically uncorrelated, then rank $(V_R)$ will be deficient. What we learn from this is that when the goal is to improve the precision of our estimate of $\theta_1$, we will only gain by shrinking the parameter estimate $\hat{\theta}_{2n}$ towards zero when $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ are asymptotically correlated. If they are correlated, then it can help to shrink $\hat{\theta}_{2n}$, but if they are uncorrelated then there will be no impact upon the risk for estimation of $\theta_1$. This is perfectly sensible.

However, when the estimators $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ are correlated then one useful feature of this estimator is that the choices for the matrices $H$ and $R$ have been separated. The matrix $H$ should be selected on the basis of which parameters are expected to be close to zero (and measured by the non-centrality parameter) not on the basis of intrinsic interest.

6 Sections to be Written

Criterion-based estimators
Least-Squares
Forecast Loss
MLE Shrinkage
GMM Shrinkage
7 Appendix

The following is a version of Stein’s Lemma (Stein, 1981).

Lemma 1 If \( Z \sim N(0, V) \) and \( \eta(x) : \mathbb{R}^k \rightarrow \mathbb{R}^k \) is absolutely continuous, then

\[
E(\eta(Z + \delta)'KZ) = E \left( \frac{\partial}{\partial x} \eta(Z + \delta)'KV \right). 
\]

Proof: Let \( \phi_V(x) \) denote the \( N(0, V) \) density function. By multivariate integration by parts

\[
E(\eta(Z + \delta)'KZ) = \int \eta(x + \delta)'Kx \phi_V(x) (dx) 
= \int \eta(x + \delta)'KV^{-1}x \phi_V(x) (dx) 
= \int \text{tr} \left( \frac{\partial}{\partial x} \eta(x + \delta)'KV \right) \phi_V(x) (dx) 
= E \left( \frac{\partial}{\partial x} \eta(Z + \delta)'KV \right). 
\]

\[\Box\]

Lemma 2 If \( Z^* = Z - \eta(Z + \delta) \) with \( Z \sim N(0, V) \) and

\[
\eta(x) = \left( \frac{C}{x'B'WBx} \right) Bx 
\]

where \( W \) is symmetric, \( \text{rank}(B'WB) \geq 1, K \geq 0, \) and \( C \geq 0, \) then

\[
E(Z^*KZ^*) = \text{tr}(KV) - E g(Z + \delta) \quad (23)
\]

where

\[
g(x) = \frac{C}{x'B'WBx} \left( 2\text{tr}(B'KV) - 4\frac{x'B'KV'B'WBx}{x'B'WBx} - C \frac{x'B'KBx}{x'B'WBx} \right) 1 \left( \frac{C}{x'B'WBx} \leq 1 \right) 
+ \left( 2\text{tr}(B'KV) - x'B'KBx \right) 1 \left( \frac{C}{x'B'WBx} > 1 \right) \quad (24)
\]

Proof: First, noting that \( \eta(x) \) is absolutely continuous,

\[
E(Z^*KZ^*) = E(Z'KZ) - 2E(\eta(Z + \delta)'KZ) + E(\eta(Z + \delta)'K\eta(Z + \delta)) 
= \text{tr}(KV) - E \left( 2\text{tr} \left( \frac{\partial}{\partial x} \eta(Z + \delta)'KV \right) - \eta(Z + \delta)'K\eta(Z + \delta) \right) 
= \text{tr}(KV) - E g(Z + \delta)
\]
the second equality using $E(Z'KZ) = \text{tr}(WW)$, Lemma 1, and $Z \sim N(0, V)$, where

$$q(x) = 2 \text{tr} \left( \frac{\partial}{\partial x} \eta(x)'KV \right) - \eta(x)'K\eta(x).$$

We now show that $q(x) = g(x)$. Note that

$$\frac{\partial}{\partial x} \eta(x)' = \left( \begin{array}{c} C \\ x'B'WBx \end{array} \right)_1 B' - \frac{2C}{(x'B'WBx)^2} B'WBxx'B'1 \left( \frac{C}{x'B'WBx} > 1 \right)$$

so

$$\text{tr} \left( \frac{\partial}{\partial x} \eta(x)'KV \right) = \left( \begin{array}{c} C \\ x'B'WBx \end{array} \right)_1 \text{tr} (B'KV) - 2C \left( \frac{x'B'KV'B'WBx}{(x'B'WBx)^2} \right)_1 \left( \frac{C}{x'B'WBx} > 1 \right)$$

and also

$$\eta(x)'K\eta(x) = \left( \left( \begin{array}{c} C \\ x'B'WBx \end{array} \right)_1 \right)^2 x'B'KBx.$$

Together

$$q(x) = \left( \begin{array}{c} C \\ x'B'WBx \end{array} \right)_1 \left( 2 \text{tr} (B'KV) - \left( \begin{array}{c} C \\ x'B'WBx \end{array} \right)_1 (x'B'KBx) \right)$$

$$- 4C \left( \frac{x'B'KV'B'WBx}{(x'B'WBx)^2} \right)_1 \left( \frac{C}{x'B'WBx} > 1 \right)$$

$$= C \left( \begin{array}{c} \text{tr} (B'KV) - 4 \frac{x'B'KBV'B'WBx}{x'B'WBx} - C \frac{x'B'KBx}{x'B'WBx} \end{array} \right)_1 \left( \frac{C}{x'B'WBx} \leq 1 \right)$$

$$+ \left( 2 \text{tr} (B'KV) - x'B'KBx \right)_1 \left( \frac{C}{x'B'WBx} > 1 \right)$$

$$= g(x)$$

as claimed. 

\textbf{Proof of Theorem 1:} Under Assumption 1,

$$\sqrt{n} \left( \hat{\theta}_n - \theta_n \right) \overset{d}{\rightarrow} Z_n.$$

By two Taylor's expansions

$$\sqrt{n}h(\theta_n) = \sqrt{n}h(\theta_0) + \sqrt{n}H'(\theta_n - \theta_0) + o(1)$$

$$= H'\delta + o(1)$$
and
\[
\sqrt{n} h(\hat{\theta}_n) = \sqrt{n} h(\theta_n) + \sqrt{n} H' \sqrt{n} (\hat{\theta}_n - \theta_n) + o_p(1)
\]
\[
= H' \delta + \sqrt{n} H' \sqrt{n} (\hat{\theta}_n - \theta_n) + o_p(1)
\]
\[
\xrightarrow{d} H' \delta + H' Z
\]
\[
= H' (Z + \delta).
\]

Using the estimator (4), this implies
\[
\sqrt{n} (\hat{\theta}_n - \bar{\theta}_n) = G_n \hat{H} (\hat{H}' G_n \hat{H})^{-1} \sqrt{n} h(\hat{\theta}_n)
\]
\[
\xrightarrow{d} GH (H'G)^{-1} H' (Z + \delta)
\]
\[
= GP (Z + \delta)
\]
which is (12). Since $C_n \xrightarrow{P} C$, it follows that
\[
\sqrt{n} (\hat{\theta}_n^* - \theta_n) = \sqrt{n} (\hat{\theta}_n - \theta_n) - \left( \frac{C}{n (\hat{\theta}_n - \bar{\theta}_n)' W (\hat{\theta}_n - \bar{\theta}_n)} \right) \sqrt{n} (\hat{\theta}_n - \bar{\theta}_n)
\]
\[
\xrightarrow{d} Z - \left( \frac{C}{(Z + \delta)' PGWGP (Z + \delta)} \right) GP (Z + \delta)
\]
\[
= Z - \eta (Z + \delta)
\]
where
\[
\eta (x) = \left( \frac{C}{x' PGWGPx} \right) GPx.
\]
Equation (26) is (13).

To evaluate the weighted risk for $\hat{\theta}_n$, first observe that
\[
R(\hat{\theta}_n, K) = EZ'KZ
\]
\[
= \text{tr} (KZZ')
\]
\[
= \text{tr} (KV)
\]
which is (14).

To evaluate the risk for $\hat{\theta}_n^*$ we can apply Lemma 2, with $B = GP$. By (23),
\[
R(\hat{\theta}_n^*, K) = E (Z - \eta (Z + \delta))' K (Z - \eta (Z + \delta))
\]
\[
= \text{tr} (KV) - E g (Z + \delta)
\]
where \( g(x) \) is (24) with \( B = GP \). Observing that

\[
\text{tr}(B'KV) = \text{tr}(PGKV) = \text{tr}\left(H(H'GH)^{-1}H'GKV\right) = \text{tr}\left((H'GH)^{-1}H'GKVH\right) = \text{tr}(A^*)
\]

we obtain

\[
g(x) = \frac{C}{x'B'WBx} \left(2\text{tr}(A^*) - 4x'BKV'BWBx + Cx'B'KBx\right)1\left(\frac{C}{x'B'WBx} \leq 1\right)
\]

\[
+ \left(2\text{tr}(A^*) - x'B'KBx\right)1\left(\frac{C}{x'B'WBx} > 1\right)
\]

which is (16) after substitutions.

To show (17) we set \( K = W \). Then we have the simplifications \( K^* = W^*, A^* = A \) and \( V^* = PGWVPGWGWP \). Note that if we set \( y = (H'GWGH)^{1/2}(H'GH)^{-1}H'x \), then

\[
\frac{x'V^*x}{x'W^*x} = \frac{x'PGWVPGWGWPx}{x'PGWGPx} = \frac{x'H(H'GH)^{-1}(H'GWVH)(H'GH)^{-1}(H'GWGH)(H'GH)^{-1}H'x}{y'(H'GWGH)^{1/2}(H'GH)^{-1}(H'GWGH)^{1/2}y} \leq \lambda_{\text{max}}\left((H'GWGH)^{1/2}(H'GH)^{-1}(H'GWGH)^{1/2}\right) = \lambda_{\text{max}}(A).
\]

(29)

Using these simplifications, (29), the second indicator function, (8) and Assumption 2.6,

\[
g(x) = \frac{C}{x'W^*x} \left(2\text{tr}(A) - 4x'V^*x + C\right)1\left(\frac{C}{x'W^*x} \leq 1\right)
\]

\[
+ \left(2\text{tr}(A) - x'W^*x\right)1\left(\frac{C}{x'W^*x} > 1\right)
\]

\[
\geq \frac{C}{x'W^*x} \left(2\text{tr}(A) - 4\lambda_{\text{max}}(A) - C\right)1\left(\frac{C}{x'W^*x} \leq 1\right)
\]

\[
+ \left(2\text{tr}(A) - C\right)1\left(\frac{C}{x'W^*x} > 1\right)
\]

\[
\geq C \left(\frac{C}{x'W^*x}\right)_1
\]

\[
> 0
\]

and thus \( g(x) > 0 \). Together, (14), (15) and \( g(x) > 0 \) imply (17).
Now we consider the case of general $K$. Again setting $y = (H'GWGH)^{1/2} (H'GH)^{-1} H'x$, observe that since $H'GWGH > 0$ under (18),

\[
\frac{x'V^*x}{x'W^*x} = \frac{x'PGKVPGWGPx}{x'PGWGPx} = \frac{x'H(H'GH)^{-1}(H'GKVGH)(H'GH)^{-1}(H'GWGH)(H'GH)^{-1}H'x}{x'H(H'GH)^{-1}(H'GWGH)(H'GH)^{-1}H'x} = \frac{y'(H'GWGH)^{-1/2}(H'GKVGH)(H'GH)^{-1}(H'GWGH)^{1/2}y}{y'y} \leq \lambda_{\max} \left( (H'GWGH)^{-1/2} (H'GKVGH) (H'GH)^{-1} (H'GWGH)^{1/2} \right) = \lambda_{\max} (A^*)
\]

and

\[
\frac{x'K^*x}{x'W^*x} = \frac{x'PGKGPx}{x'PGWGPx} = \frac{x'H(H'GH)^{-1}(H'GKVGH)(H'GH)^{-1}H'x}{x'H(H'GH)^{-1}(H'GWGH)(H'GH)^{-1}H'x} = \frac{y'(H'GWGH)^{-1/2}(H'GKVGH)(H'GWGH)^{-1/2}y}{y'y} \leq \lambda_{\max} \left( (H'GWGH)^{-1/2} (H'GKVGH) (H'GWGH)^{-1/2} \right) = \lambda_{\max} (A^{**}).
\]

Using these bounds, the indicator function, and (19) we find

\[
g(x) = \frac{C}{x'W^*x} \left( 2 \text{tr} (A^*) - 4 \frac{x'V^*x}{x'W^*x} - C \frac{x'K^*x}{x'W^*x} \right) 1 \left( \frac{C}{x'W^*x} \leq 1 \right) + \left( 2 \text{tr} (A^*) - x'W^*x \right) 1 \left( \frac{C}{x'W^*x} > 1 \right) \geq \frac{C}{x'W^*x} \left( 2 \text{tr} (A^*) - 4 \lambda_{\max} (A^*) - C \lambda_{\max} (A^{**}) \right) 1 \left( \frac{C}{x'W^*x} \leq 1 \right) + (2 \text{tr} (A^*) - C \lambda_{\max} (A^{**})) 1 \left( \frac{C}{x'W^*x} > 1 \right) \geq \left( \frac{C}{x'W^*x} \right) \left( 2 \text{tr} (A^*) - 4 \lambda_{\max} (A^*) - C \lambda_{\max} (A^{**}) \right) \geq 0.
\]

This completes the proof. □
References


