Generalized Shrinkage Estimators

Bruce E. Hansen*
University of Wisconsin†

www.ssc.wisc.edu/~bhansen

This draft: October 20, 2008
Preliminary

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 unrestricted and restricted estimators.

We consider asymptotic risk defined with selected weight matrices. We construct a targeted shrinkage estimator which has global reduced risk relative to the unrestricted estimator.

We also 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 shrinkage estimators in this context, and that they inherit the globally risk reduction property of the general setting.

Our results encompass general econometric estimators defined by criterion functions, including maximum likelihood, generalized method of moments, least squares, and minimum distance.

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

1 Introduction

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

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

where $(a)_+ = a1(a \geq 0)$. It is well known that when $k \geq 3$ the estimator $\hat{\boldsymbol{\theta}}_n^*$ has smaller risk than $\hat{\boldsymbol{\theta}}_n$, 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 the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n^*$ is discontinuous at $\boldsymbol{\theta} = \mathbf{0}$, so that there is no difference between the "asymptotic" risk of the shrinkage estimator $\hat{\boldsymbol{\theta}}_n^*$ and the unconstrained estimator $\hat{\boldsymbol{\theta}}_n$. The barrier in this analysis is the asymptotic framework. To eliminate a discontinuity in an asymptotic distribution, a constructive solution is to reparameterize the model as an array so that the parameter is local to the discontinuity, thereby attaining a continuous asymptotic distribution. (A 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 $\boldsymbol{\theta}_n = n^{-1/2}\boldsymbol{\delta}$ where $\boldsymbol{\delta}$ is fixed. With this reparameterization, it is straightforward to derive an asymptotic distribution for $\hat{\boldsymbol{\theta}}_n^*$ as $n \to \infty$ which is continuous in $\boldsymbol{\delta}$. 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 $\hat{\boldsymbol{\theta}}_n^*$ is strictly less than that of the original estimator $\hat{\boldsymbol{\theta}}_n$, 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, as shrinkage towards a subspace requires a restricted parameter estimator, we allow for a wide class of restricted estimators. 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 $\boldsymbol{\theta}_n$.

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 extends the method to allow the asymptotic risk to be constructed with a targeted weight matrix, and introduces a shrinkage estimator which achieves global risk reduction in this context. Section 6 examines the robustness of risk reduction when the weight matrix is misspecified. Section 7 examines shrinkage in the presence of nuisance parameters. Section 8 shows that the methods are applicable to a broad range of criterion-based estimators. Section 9 develops a least squares shrinkage estimator using squared forecast error loss.

2 Generalized Shrinkage Estimator

Suppose that we are interested in estimating a parameter $\boldsymbol{\theta}_n$ which is an element of parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^k$. Furthermore, suppose that there is a sub-space of interest:

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

where $\mathbf{h}(\boldsymbol{\theta}) : \mathbb{R}^k \to \mathbb{R}^r$. We will call $\boldsymbol{\Theta}$ the unrestricted parameter space or model, and $\boldsymbol{\Theta}_0$ the restricted parameter space or sub-model. The purpose of the sub-model $\boldsymbol{\Theta}_0$ is not to set up a hypothesis test. Rather, it is to provide a simpler model of interest towards which to shrink. The number of restrictions r (the "shrinkage dimension") will play an important role in our analysis.

A traditional choice is to set $\boldsymbol{\Theta}_0 = \{\mathbf{0}\}$, so that estimation is shrunk towards the zero vector 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, that is, if we partition

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} \begin{matrix} k-r \\ r \end{matrix},$$

then $\mathbf{h}(\boldsymbol{\theta})$ takes the form $\mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{\theta}_2$ so that $\boldsymbol{\Theta}_0$ is the sub-model with $\boldsymbol{\theta}_2 = \mathbf{0}$.

We assume the following conditions on the probability structure.

Assumption 1 *As* $n \to \infty$,

1. $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{\delta}$ *where* $\boldsymbol{\delta} \in \mathbb{R}^k$ *and* $\mathbf{h}(\boldsymbol{\theta}_0) = 0$.
2. $\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{h}(\boldsymbol{\theta})'$ *is continuous in a neighborhood of* $\boldsymbol{\theta}_0$.
3. $\operatorname{rank}(\mathbf{H}) = r > 2$ *where* $\mathbf{H} = \mathbf{H}(\boldsymbol{\theta}_0)$.

Assumption 1.1 is the key assumption for our asymptotic theory. It states that the true parameter vector $\boldsymbol{\theta}_n$ lies in a $n^{-1/2}$ neighborhood of the restricted parameter space $\boldsymbol{\Theta}_0$. The local parameter $\boldsymbol{\delta}$ measures the discrepancy between $\boldsymbol{\Theta}_0$ and $\boldsymbol{\theta}_n$. Assumption 1.1 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 $\boldsymbol{\delta}$ and thus should not be viewed as restrictive.

The smoothness and full rank conditions of Assumptions 1.2 and 1.3 are standard. The assumption $r > 2$ is Stein's (1956) classic condition for shrinkage. Equivalently, the shrinkage dimension must be three or larger. As shown by Stein (1956) this condition is necessary in order for shrinkage to achieve global reductions in risk relative to unrestricted estimation.

Next, suppose that there are two estimators for $\boldsymbol{\theta}_n$: an unrestricted estimator $\hat{\boldsymbol{\theta}}_n$ and a restricted estimator $\tilde{\boldsymbol{\theta}}_n$, and that there is an estimator $\hat{\mathbf{V}}_n$ for the asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_n$. We impose the following high-level conditions.

Assumption 2 *As* $n \to \infty$,

1. $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{V})$ *with* $\mathbf{V} > 0$.
2. $\hat{\mathbf{V}}_n \xrightarrow{p} \mathbf{V}$.
3. *For some symmetric* $\mathbf{G}$ *such that* $\operatorname{rank}(\mathbf{GH}) = r$, $\tilde{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n - \mathbf{GH}\left(\mathbf{H}'\mathbf{GH}\right)^{-1}\mathbf{h}(\hat{\boldsymbol{\theta}}_n) + o_p\left(n^{-1/2}\right)$.

Assumptions 2.1 and 2.2 are standard, stating the the unrestricted estimator $\hat{\boldsymbol{\theta}}_n$ is asymptotically normal and its covariance matrix estimate is consistent. Assumption 2.3 states that the restricted estimator can be asymptotically constructed from the unrestricted estimator. This condition is satisfied by most restricted estimators. An efficient restricted estimator (when $\mathbf{h}(\boldsymbol{\theta}_n) = 0$) requires $\mathbf{G} = \mathbf{V}$. We do not impose this condition in order to include important estimators of interest. The assumption that $\mathbf{GH}$ is full rank is required so that $\tilde{\boldsymbol{\theta}}_n$ is well defined. It allows the matrix $\mathbf{G}$ to be deficient rank, but in this case $\mathbf{H}$ cannot lie in the null space of $\mathbf{G}$.

A simple estimator satisfying Assumption 2.3 is

$$\tilde{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n - \mathbf{G}_n\hat{\mathbf{H}}\left(\hat{\mathbf{H}}'\mathbf{G}_n\hat{\mathbf{H}}\right)^{-1}\mathbf{h}(\hat{\boldsymbol{\theta}}_n) \tag{3}$$

where $\hat{\mathbf{H}} = \mathbf{H}(\hat{\boldsymbol{\theta}}_n)$ and $\mathbf{G}_n$ is a consistent estimate of $\mathbf{G}$. In particular,

$$\tilde{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n - \hat{\mathbf{V}}_n \hat{\mathbf{H}} \left(\hat{\mathbf{H}}' \hat{\mathbf{V}}_n \hat{\mathbf{H}} \right)^{-1} \mathbf{h}(\hat{\boldsymbol{\theta}}_n) \tag{4}$$

is asymptotically efficient under $\mathbf{h}(\boldsymbol{\theta}_n) = 0$. When $\mathbf{h}(\boldsymbol{\theta})$ is linear the restricted estimator will typically take the form (3), and in this case $\mathbf{h}(\tilde{\boldsymbol{\theta}}_n) = 0$ so $\tilde{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}_0$. When $\mathbf{h}(\boldsymbol{\theta})$ is non-linear then a typical restricted estimator (one which satisfies $\tilde{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}_0$) will not satisfy (3) exactly. We still call $\tilde{\boldsymbol{\theta}}_n$ a restricted estimator in this case because it asymptotically satisfies the restriction. We discuss a range of estimators satisfying Assumption 2 in Section 8.

Given $\hat{\boldsymbol{\theta}}_n$, $\tilde{\boldsymbol{\theta}}_n$, and $\hat{\mathbf{V}}_n$, our *generalized shrinkage estimator* for $\boldsymbol{\theta}_n$ is

$$\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n - \left(\frac{r-2}{n \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right)' \hat{\mathbf{V}}_n^{-1} \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right)} \right)_1 \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right) \tag{5}$$

where

$$(a)_1 = 1(a \geq 1) + a1(0 \leq a < 1)$$

is a trimming function which trims its argument to lie between zero and one. If $\hat{\boldsymbol{\theta}}_n$ is a linear regression coefficient, (3) specializes to a classic Stein-rule estimator when $\tilde{\boldsymbol{\theta}}_n = 0$, and to Oman's (1982a, 1982b) estimator when $\mathbf{h}(\boldsymbol{\theta})$ is linear.

We can also write (5) as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_n^* &= \hat{\boldsymbol{\theta}}_n - \left(\frac{r-2}{D_n} \right)_1 \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right) \\ &= \begin{cases} \tilde{\boldsymbol{\theta}}_n & \text{if} \quad D_n < r-2 \\ \hat{\boldsymbol{\theta}}_n - \left(\frac{r-2}{D_n} \right) \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right) & \text{if} \quad D_n \geq r-2 \end{cases} \end{aligned}$$

where

$$D_n = n \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right)' \hat{\mathbf{V}}_n^{-1} \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right) \tag{6}$$

is a chi-square statistic for testing the hypothesis $\mathbf{h}(\boldsymbol{\theta}_n) = 0$. Thus $\hat{\boldsymbol{\theta}}_n^*$ can be viewed as a smoothed pre-test estimator depending on the statistic D_n. To understand this dependence, note that when $D_n \leq r-2$ then full shrinkage occurs and $\hat{\boldsymbol{\theta}}_n^* = \tilde{\boldsymbol{\theta}}_n$. As D_n increases above $r-2$ then $\hat{\boldsymbol{\theta}}_n^*$ becomes a weighted average of $\tilde{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_n$, with the weight varying smoothly with D_n. As D_n gets large, $\hat{\boldsymbol{\theta}}_n^*$ approaches the unrestricted estimator $\hat{\boldsymbol{\theta}}_n$.

When the restricted estimator satisfies (3), then (23) can be written as

$$\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n - \left(\frac{r-2}{n \mathbf{h}(\hat{\boldsymbol{\theta}}_n)' \left(\hat{\mathbf{H}}' \mathbf{G}_n \hat{\mathbf{H}} \right)^{-1} \hat{\mathbf{H}} \mathbf{G}_n \hat{\mathbf{V}}_n^{-1} \mathbf{G}_n \hat{\mathbf{H}} \left(\hat{\mathbf{H}}' \mathbf{G}_n \hat{\mathbf{H}} \right)^{-1} \mathbf{h}(\hat{\boldsymbol{\theta}}_n)} \right)_1 \mathbf{G}_n \hat{\mathbf{H}} \left(\hat{\mathbf{H}}' \mathbf{G}_n \hat{\mathbf{H}} \right)^{-1} \mathbf{h}(\hat{\boldsymbol{\theta}}_n)$$

and when it satisfies (4) this further simplifies to

$$\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n - \left(\frac{r-2}{n\mathbf{h}(\hat{\boldsymbol{\theta}}_n)' \left(\hat{\mathbf{H}}' \hat{\mathbf{V}}_n \hat{\mathbf{H}} \right)^{-1} \mathbf{h}(\hat{\boldsymbol{\theta}}_n)} \right)_1 \hat{\mathbf{V}}_n \hat{\mathbf{H}} \left(\hat{\mathbf{H}}' \hat{\mathbf{V}}_n \hat{\mathbf{H}} \right)^{-1} \mathbf{h}(\hat{\boldsymbol{\theta}}_n).$$

3 Asymptotic Distribution and Generalized Risk

We assess performance by estimation risk defined as asymptotic expected squared error loss. Specifically, for a symmetric and non-negative definite weight matrix $\mathbf{W}$ the *weighted asymptotic risk* of an estimator $\bar{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}_n$ is defined as

$$R(\bar{\boldsymbol{\theta}}_n, \mathbf{W}) = \lim_{n\to\infty} \, n \, \mathrm{E}\left(\left(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \right)' \mathbf{W} \left(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \right) \right). \tag{7}$$

If we set $\mathbf{W} = \mathbf{V}^{-1}$ then the risk is invariant to reparameterization and scaling, so this is our default choice for the weight matrix. Accordingly, we define the *generalized asymptotic risk* of $\bar{\boldsymbol{\theta}}_n$ as

$$R(\bar{\boldsymbol{\theta}}_n) = R(\bar{\boldsymbol{\theta}}_n, \mathbf{V}^{-1}) = \lim_{n\to\infty} \, n \, \mathrm{E}\left(\left(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \right)' \mathbf{V}^{-1} \left(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \right) \right).$$

We now present the asymptotic distribution and generalized asymptotic risk of the generalized shrinkage estimator $\hat{\boldsymbol{\theta}}_n^*$.

Theorem 1 *Under Assumptions 1 and 2*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \right) \xrightarrow{d} \mathrm{Z}, \tag{8}$$

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right) \xrightarrow{d} \mathbf{G}\mathbf{P}_{\mathbf{G}} \left(\mathrm{Z} + \boldsymbol{\delta} \right), \tag{9}$$

and

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n \right) \xrightarrow{d} \mathrm{Z} - \left(\frac{r-2}{\left(\mathrm{Z} + \boldsymbol{\delta} \right)' \mathbf{W}_{\mathbf{G}} \left(\mathrm{Z} + \boldsymbol{\delta} \right)} \right)_1 \mathbf{G}\mathbf{P}_{\mathbf{G}} \left(\mathrm{Z} + \boldsymbol{\delta} \right), \tag{10}$$

where $\mathrm{Z} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{V} \right)$,

$$\mathbf{W}_{\mathbf{G}} = \mathbf{P}_{\mathbf{G}} \mathbf{G} \mathbf{V}^{-1} \mathbf{G} \mathbf{P}_{\mathbf{G}} \tag{11}$$

and

$$\mathbf{P}_{\mathbf{G}} = \mathbf{H} \left(\mathbf{H}' \mathbf{G} \mathbf{H} \right)^{-1} \mathbf{H}'. \tag{12}$$

The generalized asymptotic risk of the unrestricted and generalized shrinkage estimators are

$$R(\hat{\boldsymbol{\theta}}_n) = k \tag{13}$$

and

$$R(\hat{\boldsymbol{\theta}}_n^*) = k - \mathrm{E} g_{\mathbf{G}} \left(\mathrm{Z} + \boldsymbol{\delta} \right) \tag{14}$$

where

$$g_{\mathbf{G}}(\mathbf{x}) = \frac{(r-2)^2}{\mathbf{x}'\mathbf{W}_{\mathbf{G}}\mathbf{x}} 1\left(\frac{r-2}{\mathbf{x}'\mathbf{W}_{\mathbf{G}}\mathbf{x}} \leq 1\right) + \left(2r - \mathbf{x}'\mathbf{W}_{\mathbf{G}}\mathbf{x}\right) 1\left(\frac{r-2}{\mathbf{x}'\mathbf{W}_{\mathbf{G}}\mathbf{x}} > 1\right). \tag{15}$$

Most importantly,

$$R(\hat{\boldsymbol{\theta}}_n^*) < R(\hat{\boldsymbol{\theta}}_n). \tag{16}$$

The main result is (16) which shows that the generalized shrinkage estimator has strictly smaller generalized asymptotic risk than the unrestricted estimator. This inequality holds globally in the parameter and distribution space when $r > 2$, and holds for any restriction $\mathbf{h}(\boldsymbol{\theta})$ satisfying Assumption 1. It also holds for any restricted estimator $\tilde{\boldsymbol{\theta}}_n$ satisfying Assumption 2.3, regardless of the weight matrix $\mathbf{G}$. (This is surprising as the restricted estimator $\tilde{\boldsymbol{\theta}}_n$ with $\mathbf{G} \neq \mathbf{V}$ may not have globally smaller risk than $\hat{\boldsymbol{\theta}}_n$ even when $\mathbf{h}(\boldsymbol{\theta}_n) = 0$.)

There are two critical conditions for (16). One is $r > 2$, which as mentioned in the previous section is Stein's (1956) classic condition for shrinkage. The other critical condition is that the risk is measured using the generalized asymptotic risk – expected squared loss weighted by $\mathbf{V}^{-1}$. Inequality (16) can be violated if $R(\cdot)$ is replaced by another risk function. In section 6 we explore the robustness of (16) to alternative weight matrices.

In addition, equation (10) expresses the asymptotic distribution of the generalized shrinkage estimator as a nonlinear function of a normal random vector. Equations (14)-(15) provide a formula for its generalized asymptotic risk, expressed as the expectation of the nonlinear function $g_{\mathbf{G}}(\mathbf{x})$.

4 Asymptotic Risk Reduction from Shrinkage

For simplicity, consider estimator (5) when $\mathbf{G} = \mathbf{V}$. Its generalized asymptotic risk is $k - \mathrm{E}g_{\mathbf{V}}(\mathrm{Z} + \boldsymbol{\delta})$ where

$$g_{\mathbf{V}}(\mathbf{x}) = \frac{(r-2)^2}{\mathbf{x}'\mathbf{P}_{\mathbf{V}}\mathbf{x}} 1\left(\frac{r-2}{\mathbf{x}'\mathbf{P}_{\mathbf{V}}\mathbf{x}} \leq 1\right) + \left(2r - \mathbf{x}'\mathbf{P}_{\mathbf{V}}\mathbf{x}\right) 1\left(\frac{r-2}{\mathbf{x}'\mathbf{P}_{\mathbf{V}}\mathbf{x}} > 1\right).$$

with $\mathbf{P}_{\mathbf{V}} = \mathbf{H}\left(\mathbf{H}'\mathbf{V}\mathbf{H}\right)^{-1}\mathbf{H}'$. Observe that $g_{\mathbf{V}}(\mathrm{Z} + \boldsymbol{\delta})$ depends on its argument only through the quadratic form $(\mathrm{Z} + \boldsymbol{\delta})'\mathbf{P}_{\mathbf{V}}(\mathrm{Z} + \boldsymbol{\delta})$ which has a non-central chi-square distribution with degrees of freedom r and non-centrality parameter

$$\overline{\delta} = \boldsymbol{\delta}'\mathbf{P}_{\mathbf{V}}\boldsymbol{\delta} = \boldsymbol{\delta}'\mathbf{H}\left(\mathbf{H}'\mathbf{V}\mathbf{H}\right)^{-1}\mathbf{H}'\boldsymbol{\delta}. \tag{17}$$

It follows that the generalized asymptotic risk is a function only of k, r, and $\overline{\delta}$.

The functional dependence simplifies when r is large. Casella and Hwang (1982) have shown that as $r \to \infty$ and $\overline{\delta}/r \to c$ that

$$\frac{\mathrm{E}g(\mathrm{Z} + \boldsymbol{\delta})}{r} \to \frac{1}{1+c}.$$

It follows that if $r/k \to \alpha \in (0,1]$ then

$$\frac{R(\hat{\boldsymbol{\theta}}_n^*)}{k} \to 1 - \frac{\alpha}{1+c}. \tag{18}$$

This is actually quite extraordinary. It is the lower bound on the risk among all estimators taking the weighted-average form

$$\hat{\boldsymbol{\theta}}_n^* = (1-\omega)\,\hat{\boldsymbol{\theta}}_n - \omega\hat{\boldsymbol{\theta}}_n$$

with $\omega \in \mathbb{R}$.

Since $R(\hat{\boldsymbol{\theta}}_n) = k$, another way of expressing (18) is

$$\frac{R(\hat{\boldsymbol{\theta}}_n^*)}{R(\hat{\boldsymbol{\theta}}_n)} \to 1 - \frac{\alpha}{1+c}$$

which says that $\alpha/(1+c)$ is the approximate percentage risk reduction due to shrinkage.

To understand the magnitude of the risk reduction, consider the case $\alpha = 1$ (all parameters are shrunk) so that the percentage risk improvement is $1/(1+c)$. Recall, c is the ratio of the non-centrality parameter $\overline{\delta}$ to the degrees of freedom r. Thus $c = 1$ when the non-centrality parameter $\overline{\delta}$ equals the degrees of freedom r, $c = 2$ when $\overline{\delta}$ is twice the degrees of freedom, etc. It seems reasonable to focus on values of c ranging from 0 through 4, with $c = 0$ representing the extreme case of a valid exclusion restriction, $1 \le c \le 2$ representing typical or modest values of $\overline{\delta}$, and $c = 4$ representing a very large value of $\overline{\delta}$. In these cases, the percentage risk reduction 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.

5 Targeted Shrinkage Estimator

In some cases the weighted risk (7) will be calculated with a specific weight matrix $\mathbf{W}$ not equal to $\mathbf{V}^{-1}$. In this case we will call the (7) the *targeted asymptotic risk*, to emphasize that it is calculated with a specific (e.g., targeted) weight matrix $\mathbf{W}$. In this case it will be desirable to construct an estimator which aims to minimize this targeted risk. It turns out that the ability to construct this estimator depends on the constant

$$C = \operatorname{tr}(\mathbf{A}) - 2\lambda_{\max}(\mathbf{A}) \tag{19}$$

where

$$\mathbf{A} = \left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{V}\mathbf{H}\right). \tag{20}$$

When $\mathbf{W} = \mathbf{V}^{-1}$ then $\mathbf{A} = \mathbf{I}_r$ and $C = r - 2$. Thus (19) generalizes the factor "$r-2$" appearing in (5).

Our proposed estimator will require a consistent estimate for C, requiring consistent estimates for $\mathbf{W}$ and $\mathbf{G}$.

Assumption 3 *There are estimates* $\mathbf{W}_n$ *and* $\mathbf{G}_n$ *such that as* $n \to \infty$,

1. $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$.
2. $\mathbf{G}_n \xrightarrow{p} \mathbf{G}$.
3. $C > 0$.
4. $\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H} > 0$.

The condition $C > 0$ is both a generalization and strengthening of Stein's condition since $r > 2$ is necessary for $C > 0$. Assumption 3.4 is a technical condition required for the asymptotic distribution of the shrinkage estimator to be well-behaved. If $\mathbf{W} > 0$ then Assumption 3.4 follows from Assumption 2.3. However, Assumption 3.4 allows $\mathbf{W}$ to be deficient rank, so long as $\mathbf{GH}$ does not lie in its null space.

We construct consistent estimates of $\mathbf{A}$ and C by replacing the unknowns in (19)-(20) with their point estimates:

$$\begin{aligned} \mathbf{A}_n &= \left(\hat{\mathbf{H}}'\mathbf{G}_n\hat{\mathbf{H}}\right)^{-1}\left(\hat{\mathbf{H}}'\mathbf{G}_n\mathbf{W}_n\hat{\mathbf{V}}_n\hat{\mathbf{H}}\right) \quad (21) \\ C_n &= \operatorname{tr}(\mathbf{A}_n) - 2\lambda_{\max}(\mathbf{A}_n). \quad (22) \end{aligned}$$

Our *targeted shrinkage estimator* for $\boldsymbol{\theta}_n$ is

$$\hat{\boldsymbol{\theta}}_n^{**} = \hat{\boldsymbol{\theta}}_n - \left(\frac{C_n}{n\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right)'\mathbf{W}_n\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right)}\right)_1\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right). \quad (23)$$

It is useful to note that when $C_n \leq 0$ then $\hat{\boldsymbol{\theta}}_n^{**}$ simplifies to $\hat{\boldsymbol{\theta}}_n$ and no shrinkage occurs.

We now examine targeted asymptotic risk.

Theorem 2 *Under Assumptions 1, 2, and 3,*

$$R(\hat{\boldsymbol{\theta}}_n^{**}, \mathbf{W}) < R(\hat{\boldsymbol{\theta}}_n, \mathbf{W}), \quad (24)$$

the targeted asymptotic risk of the targeted shrinkage estimator is strictly less than that of the unrestricted estimator.

Just as for the generalized shrinkage estimator, the inequality (24) holds globally in the parameter and distribution space under the condition $C > 0$. Theorem 2 shows (23) is a desirable estimator when risk is measured with a specific weight matrix $\mathbf{W}$.

This result shows that for any choice of weight matrix $\mathbf{W}$, the simple estimator $\hat{\boldsymbol{\theta}}_n^{**}$ has strictly smaller risk (defined using the weight matrix $\mathbf{W}$) than the unrestricted estimator. This risk reduction can be substantial, and holds globally in the parameter and distribution space.

6 Robustness

While Theorems 1 and 2 show that our shrinkage estimators reduce asymptotic risk when the risk function weight is known, what happens when it is unknown? Equivalently, what happens when the estimator is constucted as in (23) with a specific target weight matrix $\mathbf{W}$, but we evaluate asymptotic risk using a different weight matrix $\mathbf{K}$? We now present conditions under which the shrinkage estimator (23) still has reduced risk relative to the unrestricted estimator.

Theorem 3 *Under Assumptions 1, 2, and 3, if*

$$C < 2\left(\frac{\operatorname{tr}(\mathbf{A}^*) - 2\lambda_{\max}(\mathbf{A}^*)}{\lambda_{\max}(\mathbf{A}^{**})}\right) \tag{25}$$

where

$$\mathbf{A}^* = \left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{G}\mathbf{K}\mathbf{V}\mathbf{H}$$

and

$$\mathbf{A}^{**} = \left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H}\right)^{-1}\left(\mathbf{H}'\mathbf{G}\mathbf{K}\mathbf{G}\mathbf{H}\right), \tag{26}$$

then

$$R(\hat{\boldsymbol{\theta}}_n^{**}, \mathbf{K}) < R(\hat{\boldsymbol{\theta}}_n, \mathbf{K}). \tag{27}$$

Similarly,

$$R(\hat{\boldsymbol{\theta}}_n^{*}, \mathbf{K}) < R(\hat{\boldsymbol{\theta}}_n, \mathbf{K})$$

if (25) holds with $C = r - 2$, and $\mathbf{A}^{**}$ in (26) defined with $\mathbf{W} = \mathbf{V}^{-1}$.

Theorem 3 shows in (27) that the shrinkage estimator can have smaller risk than the unrestricted estimator, even if an alternative weight matrix $\mathbf{K}$ is used to define risk. The key condition is (25), which essentially holds as long as $\mathbf{K}$ is not too different than $\mathbf{W}$. Thus the results of Theorems 1 and 2 are robust to modest deviations in the weight matrix from that used to construct the estimators.

7 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 suggests considering weight matrices $\mathbf{W}$ which reflect this interest. To be specific, partition

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} \begin{matrix} s \\ k-s \end{matrix},$$

where $\boldsymbol{\theta}_1$ are the parameters of interest ($s \geq 3$) and $\boldsymbol{\theta}_2$ are the nuisance parameters. In this case the loss function should only put weight on estimates of $\boldsymbol{\theta}_1$, suggesting a weight matrix of the form

$$\mathbf{W} = \mathbf{R}\mathbf{W}_1\mathbf{R}' \tag{28}$$

where

$$\mathbf{R} = \begin{pmatrix} \mathbf{I}_s \\ \mathbf{0} \end{pmatrix} \tag{29}$$

and $\mathbf{W}_1 > 0$ is $s \times s$. From (23) the targeted shrinkage estimator for $\boldsymbol{\theta}_1$ is

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{1n}^{**} &= \hat{\boldsymbol{\theta}}_{1n} - \left(\frac{C_n}{n \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right)' \mathbf{R}\mathbf{W}_1\mathbf{R}' \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n \right)} \right)_1 \left(\hat{\boldsymbol{\theta}}_{1n} - \tilde{\boldsymbol{\theta}}_{1n} \right) \\
&= \hat{\boldsymbol{\theta}}_{1n} - \left(\frac{C_n}{n \left(\hat{\boldsymbol{\theta}}_{1n} - \tilde{\boldsymbol{\theta}}_{1n} \right)' \mathbf{W}_1 \left(\hat{\boldsymbol{\theta}}_{1n} - \tilde{\boldsymbol{\theta}}_{1n} \right)} \right)_1 \left(\hat{\boldsymbol{\theta}}_{1n} - \tilde{\boldsymbol{\theta}}_{1n} \right) \\
C_n &= \operatorname{tr}(\mathbf{A}_n) - 2\lambda_{\max}(\mathbf{A}_n) \\
\mathbf{A}_n &= \left(\hat{\mathbf{H}}'\mathbf{G}_n\hat{\mathbf{H}} \right)^{-1} \left(\hat{\mathbf{H}}'\mathbf{G}_n\mathbf{R}\mathbf{W}_1\mathbf{R}'\hat{\mathbf{V}}_n\hat{\mathbf{H}} \right).
\end{aligned}$$

If there is no selected choice for the $s \times s$ weight matrix $\mathbf{W}_1$, an agnostic choice is one such that C_n takes a simple form. Consider the $s \times s$ matrix

$$\mathbf{V}_{\mathbf{R}} = \mathbf{R}'\hat{\mathbf{V}}_n\hat{\mathbf{H}} \left(\hat{\mathbf{H}}'\mathbf{G}_n\hat{\mathbf{H}} \right)^{-1} \hat{\mathbf{H}}'\mathbf{G}_n\mathbf{R}$$

and notice that $\mathbf{V}_{\mathbf{R}} \geq 0$, and $\operatorname{rank}(\mathbf{V}_{\mathbf{R}}) \leq \min[r, s]$. If $r \geq s$ then $\mathbf{V}_{\mathbf{R}}$ can be full rank but if $r < s$ then it must have deficient rank. Therefore we recommend

$$\mathbf{W}_1 = \mathbf{V}_{\mathbf{R}}^+$$

the Moore-Penrose generalized inverse $\mathbf{V}_{\mathbf{R}}$. With this choice, then

$$\begin{aligned}
\operatorname{tr}(\mathbf{A}_n) &= \operatorname{tr}\left(\left(\hat{\mathbf{H}}'\mathbf{G}_n\hat{\mathbf{H}} \right)^{-1} \left(\hat{\mathbf{H}}'\mathbf{G}_n\mathbf{R}\mathbf{V}_{\mathbf{R}}^+\mathbf{R}'\hat{\mathbf{V}}_n\hat{\mathbf{H}} \right) \right) \\
&= \operatorname{tr}\left(\mathbf{V}_{\mathbf{R}}\mathbf{V}_{\mathbf{R}}^+ \right) \\
&= \operatorname{rank}\left(\mathbf{V}_{\mathbf{R}}\mathbf{V}_{\mathbf{R}}^+ \right)
\end{aligned}$$

and

$$\lambda_{\max}(\mathbf{A}_n) = \lambda_{\max}\left(\mathbf{V}_{\mathbf{R}}\mathbf{V}_{\mathbf{R}}^+ \right) = 1$$

since $\mathbf{V}_{\mathbf{R}}\mathbf{V}_{\mathbf{R}}^+$ is idempotent (e.g. Magnus and Neudecker, 1988, p. 33). Thus $C_n = \operatorname{rank}\left(\mathbf{V}_{\mathbf{R}}\mathbf{V}_{\mathbf{R}}^+ \right) - 2$. Assuming $\operatorname{rank}\left(\mathbf{V}_{\mathbf{R}}\mathbf{V}_{\mathbf{R}}^+ \right) = \min[r, s]$ we obtain the shrinkage estimator

$$\hat{\boldsymbol{\theta}}_{1n}^{**} = \hat{\boldsymbol{\theta}}_{1n} - \left(\frac{\min[r, s] - 2}{n \left(\hat{\boldsymbol{\theta}}_{1n} - \tilde{\boldsymbol{\theta}}_{1n} \right)' (\mathbf{V}_{\mathbf{R}})^+ \left(\hat{\boldsymbol{\theta}}_{1n} - \tilde{\boldsymbol{\theta}}_{1n} \right)} \right)_1 \left(\hat{\boldsymbol{\theta}}_{1n} - \tilde{\boldsymbol{\theta}}_{1n} \right).$$

By Theorem 2 a sufficient condition for this estimator to have smaller asymptotic risk than $\hat{\boldsymbol{\theta}}_{1n}$ is

$\text{rank}\,(\mathbf{V}_{\mathbf{R}}) = \min[r, s] > 2$.

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 $\hat{\boldsymbol{\theta}}_{1n}^{**}$ is a function only of the sub-estimates $\hat{\boldsymbol{\theta}}_{1n}$ and $\tilde{\boldsymbol{\theta}}_{1n}$, but the quadratic form determining the degree of shrinkage depends on the estimation method $\mathbf{G}_n$, the direction of shrinkage $\mathbf{H}$, and the sub-parameter selector matrix $\mathbf{R}$.

From this analysis we can see a potential limitation. If $\mathbf{R}$ and $\mathbf{H}$ are orthogonal, then the rank of $\mathbf{V}_{\mathbf{R}}$ will be determined by the off-diagonal block of $\hat{\mathbf{V}}_n$. In particular, if $\hat{\boldsymbol{\theta}}_{1n}$ and $\hat{\boldsymbol{\theta}}_{2n}$ are asymptotically uncorrelated, then $\text{rank}\,(\mathbf{V}_{\mathbf{R}})$ will be deficient. What we learn from this is that when the goal is to improve the precision of our estimate of $\boldsymbol{\theta}_1$, we will only gain by shrinking the parameter estimate $\hat{\boldsymbol{\theta}}_{2n}$ when $\hat{\boldsymbol{\theta}}_{1n}$ and $\hat{\boldsymbol{\theta}}_{2n}$ are asymptotically correlated. If they are correlated, then it can help to shrink $\hat{\boldsymbol{\theta}}_{2n}$, but if they are uncorrelated then there will be no impact upon the risk for estimation of $\boldsymbol{\theta}_1$. This is perfectly sensible.

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

8 Criterion-Based Estimation

In this section we show that our results apply to any set of criterion-based estimators. Specifically, suppose that the unrestricted estimator and restricted estimatores can be written as

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argmin}}\, Q_n(\boldsymbol{\theta}) \tag{30}$$

$$\tilde{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}{\text{argmin}}\, Q_n(\boldsymbol{\theta}) \tag{31}$$

where $Q_n(\boldsymbol{\theta}_n) : \mathbb{R}^k \to \mathbb{R}$ is a criterion function. This includes estimation based on maximum likelihood, GMM, minimum distance, and least squares. We focus on the case of smooth criterion functions, and define the criterion score and Hessian

$$\mathbf{S}_n(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\frac{\partial}{\partial\boldsymbol{\theta}}Q_n(\boldsymbol{\theta})$$

$$\mathbf{M}_n(\boldsymbol{\theta}) = \frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n(\boldsymbol{\theta}).$$

Estimators for the asymptotic variance of $\hat{\boldsymbol{\theta}}_n$ take the form

$$\hat{\mathbf{V}}_n = \hat{\mathbf{M}}_n^{-1}\hat{\boldsymbol{\Omega}}_n\hat{\mathbf{M}}_n^{-1}$$

where $\hat{\mathbf{M}}_n = \mathbf{M}_n(\hat{\boldsymbol{\theta}}_n)$ and $\hat{\boldsymbol{\Omega}}_n$ is a estimate of the asymptotic variance of $\mathbf{S}_n(\boldsymbol{\theta}_n)$.

The following high-level regularity conditions are conventional and are satisfied for smooth

econometric estimators.

Assumption 4

1. $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \xrightarrow{p} 0$.
2. $\mathbf{M}_n(\boldsymbol{\theta}) - \mathbf{M}(\boldsymbol{\theta}) \xrightarrow{p} 0$ *uniformly and* $\mathbf{M}(\boldsymbol{\theta})$ *is continuous, in a neighborhood of* $\boldsymbol{\theta}_0 = \lim_{n\to\infty} \boldsymbol{\theta}_n$.
3. $\mathbf{M} = \mathbf{M}(\boldsymbol{\theta}_0) > 0$.
4. $\mathbf{S}_n(\boldsymbol{\theta}_n) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$ *where* $\boldsymbol{\Omega} > 0$.
5. $\hat{\boldsymbol{\Omega}}_n \to_p \boldsymbol{\Omega}$.

Theorem 4 *Assumptions 1 and 4 imply Assumption 2 with* $\mathbf{G} = \mathbf{M}^{-1}$. *Thus Assumptions 1, 3 and 4 are sufficient for Theorem 1-3.*

It follows that for criterion-based estimators, the generalized shrinkage estimator takes the form (23)-(22) with

$$\mathbf{A}_n = \left(\hat{\mathbf{H}}'\hat{\mathbf{M}}_n^{-1}\hat{\mathbf{H}}\right)^{-1}\left(\hat{\mathbf{H}}'\hat{\mathbf{M}}_n^{-1}\mathbf{W}_n\hat{\mathbf{V}}_n\hat{\mathbf{H}}\right).$$

When both the unrestricted and restricted estimators are efficient, then the criterion function can be scaled so that $\boldsymbol{\Omega} = \mathbf{M}$. This is a type of "information matrix" equality, and holds when the criterion function $Q_n(\boldsymbol{\theta})$ is a correctly-specified log-likelihood, efficient GMM criterion, empirical likelihood criterion, efficient minimum chi-square criterion, or least-squares criterion under homoskedasticity.

In this case, an asymptotically equivalent shrinkage estimator to (5) is

$$\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n - \left(\frac{r-2}{2\left(Q_n(\tilde{\boldsymbol{\theta}}) - Q_n(\hat{\boldsymbol{\theta}})\right)}\right)_1 \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right). \tag{32}$$

This works because in this context the distance statistic

$$D_n^* = 2\left(Q_n(\tilde{\boldsymbol{\theta}}) - Q_n(\hat{\boldsymbol{\theta}})\right) \tag{33}$$

is asymptotically equivalent to the quadratic statisic (6). The shrinkage estimator (32) is particularly convenient because it is only a function of the unrestricted and restricted estimators and the criterion function at these estimates. Since (32) is equilivalent to (5), it is a generalized shrinkage estimator, not a targeted shrinkage estimator. Equivalently, it is optimal with risk is evaluated using the risk weight matrix $\mathbf{V}^{-1}$.

Specifically, this includes the following applications.

Example 1 *(MLE) When the model is a conditional density function* $f_i(\boldsymbol{\theta})$ *then* $Q_n(\boldsymbol{\theta})$ *equals the log-likelihood function*

$$Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_i(\boldsymbol{\theta}).$$

In this case (30) and (31) are the unrestricted and restricted MLE, and (33) is the likelihood ratio statistic for the test of $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}_0$.

Example 2 *(GMM) When the parameter is defined by the moment condition*

$$\mathrm{E}\mathbf{g}_i(\boldsymbol{\theta}) = \mathbf{0}$$

where $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^m$. *Then* $Q_n(\boldsymbol{\theta})$ *is one-half times the standard GMM criterion function:*

$$\begin{aligned} Q_n(\boldsymbol{\theta}) &= \frac{n}{2}\, \overline{\mathbf{g}}_n(\boldsymbol{\theta})' \hat{\boldsymbol{\Omega}}^{-1} \overline{\mathbf{g}}_n(\boldsymbol{\theta}) \\ \overline{\mathbf{g}}_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta}) \\ \hat{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\bar{\boldsymbol{\theta}})\mathbf{g}_i(\bar{\boldsymbol{\theta}})' - \overline{\mathbf{g}}_n(\bar{\boldsymbol{\theta}})\overline{\mathbf{g}}_n(\bar{\boldsymbol{\theta}})' \end{aligned}$$

and $\bar{\boldsymbol{\theta}}$ *is a preliminary estimator of* $\boldsymbol{\theta}$. *In this case (30) and (31) are the unrestricted and restricted GMM estimators, and (33) is the GMM distance statistic (Newey and West, 1987) for the test of* $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}_0$.

Example 3 *(Minimum Chi-Square) If* $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{V})$ *and* $\hat{\mathbf{V}}_n \xrightarrow{p} \mathbf{V}$ *then we can set*

$$Q_n(\boldsymbol{\theta}) = \frac{n}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\right)' \hat{\mathbf{V}}_n^{-1}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\right).$$

In this case (31) is the minimum Chi-square estimator, and (33) is the minimum Chi-square statistic for the test of $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}_0$. *(31) equals (4) when* $\mathbf{h}(\boldsymbol{\theta})$ *is linear. The minimum chi-square method is a convenient method to construct a restricted estimator* $\tilde{\boldsymbol{\theta}}_n$ *from a general estimator* $\hat{\boldsymbol{\theta}}_n$ *with covariance matrix estimate* $\hat{\mathbf{V}}_n$ *for any linear or non-linear* $\mathbf{h}(\boldsymbol{\theta})$.

Example 4 *If* $y_i = \mathbf{x}_i'\boldsymbol{\theta}_n + e_i$ *and* $\hat{\boldsymbol{\theta}}_n$ *and* $\tilde{\boldsymbol{\theta}}_n$ *are the least-squares estimates of* $\boldsymbol{\theta}_n$, *then* $Q_n(\boldsymbol{\theta}) = \sum_{i=1}^n \left(y_i - \mathbf{x}_i'\boldsymbol{\theta}\right)^2 / \hat{\sigma}^2$ *where* $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\theta}}_n\right)^2$. *In this case* $\mathbf{M} = \mathrm{E}\mathbf{x}_i\mathbf{x}_i'$ *and (33) is a scaled F statistic for for the test of* $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}_0$. *Furthermore,* $\boldsymbol{\Omega} = \mathbf{M}$ *when* $\mathrm{E}(e_i \mid \mathbf{x}_i) = 0$ *and* $\mathrm{E}\left(e_i^2 \mid \mathbf{x}_i\right) = \sigma^2$ *(homoskeastic regression).*

9 Least-Squares Estimation and Forecasting

Take the linear model

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\theta}_n + e_i \\ &= \mathbf{x}_{1i}'\boldsymbol{\theta}_{1n} + \mathbf{x}_{2i}'\boldsymbol{\theta}_{2n} + e_i \\ \mathrm{E}(\mathbf{x}_i e_i) &= 0 \end{aligned}$$

and the restriction $\boldsymbol{\theta}_{2n} = 0$. The conventional unrestricted and restricted estimators of $\boldsymbol{\theta}_n$ are the unrestricted and restricted least-squares estimators

$$y_i = \mathbf{x}_i'\hat{\boldsymbol{\theta}}_n + \hat{e}_i$$

and

$$\begin{aligned} y_i &= \mathbf{x}_{1i}'\tilde{\boldsymbol{\theta}}_{1n} + \tilde{e}_i \\ \tilde{\boldsymbol{\theta}}_n &= \begin{pmatrix} \tilde{\boldsymbol{\theta}}_{1n} \\ 0 \end{pmatrix} \end{aligned}$$

respectively.

As described in Example 4 of the previous section, these estimators satisfy the requirements of Assumption 4 with

$$\mathbf{H} = \begin{pmatrix} 0 \\ \mathbf{I} \end{pmatrix},$$

$\mathbf{V} = \mathbf{M}^{-1}\boldsymbol{\Omega}\mathbf{M}^{-1}$, $\mathbf{M} = \mathrm{E}\left(\mathbf{x}_i\mathbf{x}_i'\right)$, $\boldsymbol{\Omega} = \mathrm{E}\left(\mathbf{x}_i\mathbf{x}_i'e_i^2\right)$, and $\mathbf{G} = \mathbf{M}^{-1}$. Let $\hat{\mathbf{V}}_n = \mathbf{M}_n^{-1}\hat{\boldsymbol{\Omega}}_n\mathbf{M}_n^{-1}$ be the standard estimator for the covariance matrix of $\hat{\boldsymbol{\theta}}_n$ where $\mathbf{M}_n = n^{-1}\mathbf{X}'\mathbf{X}$ and $\hat{\boldsymbol{\Omega}}_n = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\hat{e}_i^2$, and let $\hat{\mathbf{V}}_n^0 = \mathbf{M}_n^{-1}\hat{\sigma}^2$ be a covariance matrix estimator valid under the homoskedastic regression assumption where $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \hat{e}_i^2$.

Note that $\mathbf{G} = \mathbf{M}^{-1} \neq \mathbf{V}$ and thus the restricted least-squares estimator is inefficient except when the error is homoskedastic. In contrast, an efficient restricted estimator can be found by the minimum chi-square method (31) (or equivalently GMM)

A common measure of model performance is the mean-squared forecast error (MSFE) of the point forecast $\mathbf{x}_{n+1}'\hat{\boldsymbol{\theta}}_n$ for an out-of-sample value y_{n+1}. If $\hat{\boldsymbol{\theta}}_n$ and $(\mathbf{x}_{n+1}, e_{n+1})$ are approximately independent then the mean-squared forecast error is

$$\begin{aligned} \mathrm{E}\left(y_{n+1} - \mathbf{x}_{n+1}'\hat{\boldsymbol{\theta}}_n\right)^2 &= \mathrm{E}\left(e_{n+1}^2\right) + \mathrm{E}\left(\mathbf{x}_{n+1}'\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)\right)^2 \\ &= \mathrm{E}\left(e_{n+1}^2\right) + \mathrm{E}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)' \mathbf{M}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right). \end{aligned}$$

This is equivalent to the weighted asymptotic risk, using the weight matrix $\mathbf{W} = \mathbf{M}$. Thus the forecasting criterion suggests using the targeted shrinkage estimator with $\mathbf{W}_n = \mathbf{M}_n$.

Putting this together, if $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n$ are the least-squares estimators (so $\mathbf{G}_n = \mathbf{M}_n^{-1}$) and we use MSFE risk (so $\mathbf{W}_n = \mathbf{M}_n$) then the targeted shrinkage estimator can be written as

$$\hat{\boldsymbol{\theta}}_n^{**} = \hat{\boldsymbol{\theta}}_n - \left(\frac{C_n}{F_n}\right)_1 \left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right) \tag{34}$$

where

$$F_n = \frac{\sum_{i=1}^n \hat{e}_i^2 - \sum_{i=1}^n \tilde{e}_i^2}{\hat{\sigma}^2}$$

is the F form of the Wald statistic (that is, computed under homoskedasticiy) for testing $\boldsymbol{\theta}_{2n} = 0$, and the shrinkage constant is

$$C_n = \operatorname{tr}(\mathbf{A}_n) - 2\lambda_{\max}(\mathbf{A}_n)$$

where

$$\mathbf{A}_n = \left(\hat{\mathbf{H}}'\hat{\mathbf{V}}_n^0\hat{\mathbf{H}}\right)^{-1}\left(\hat{\mathbf{H}}'\hat{\mathbf{V}}_n\hat{\mathbf{H}}\right).$$

If homoskedasticity is assumed (or believed to be a reasonable approximation) then we can replace $\mathbf{A}_n$ with the identity matrix so that C_n simplifies to $r-2$ and we obtain the shrinkage estimator

$$\hat{\boldsymbol{\theta}}_n^{**} = \hat{\boldsymbol{\theta}}_n - \left(\frac{r-2}{F_n}\right)_1\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right). \tag{35}$$

When $\mathbf{h}$ is linear this is the classic Stein-rule shrinkage estimator. Equation (35) shows that it generalizes to shrinkage towards non-linear subspaces.

Theorem 2 shows that (34) and (35) have smaller targeted asymptotic risk (MSFE) than the unrestricted least-squares estimator, globally in the parameter space. This holds generically for least squares estimation, including time-series applications. In contrast, classic Stein-rule theory has been confined to the Gaussian regression model.

10 Appendix

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

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

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

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

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

■

Lemma 2 *If* $\mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$ *and*

$$\boldsymbol{\eta}(\mathbf{x}) = \left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{W}\mathbf{B}\mathbf{x}}\right)_1 \mathbf{B}\mathbf{x}$$

where $\mathbf{W}$ *is symmetric,* $\mathrm{rank}(\mathbf{B}'\mathbf{W}\mathbf{B}) \geq 2$, $\mathbf{K} \geq 0$, *and* $C \geq 0$, *then*

$$\mathrm{E}\left((\mathrm{Z}-\boldsymbol{\eta}(\mathrm{Z}+\boldsymbol{\delta}))'\mathbf{K}(\mathrm{Z}-\boldsymbol{\eta}(\mathrm{Z}+\boldsymbol{\delta}))\right) = \mathrm{tr}(\mathbf{K}\mathbf{V}) - \mathrm{E}g(\mathrm{Z}+\boldsymbol{\delta}) \tag{36}$$

where

$$\begin{aligned}
g(\mathbf{x}) &= \frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{W}\mathbf{B}\mathbf{x}}\left(2\,\mathrm{tr}\left(\mathbf{B}'\mathbf{K}\mathbf{V}\right) - 4\frac{\mathbf{x}'\mathbf{B}'\mathbf{K}\mathbf{V}\mathbf{B}'\mathbf{W}\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{B}'\mathbf{W}\mathbf{B}\mathbf{x}} - C\frac{\mathbf{x}'\mathbf{B}'\mathbf{K}\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{B}'\mathbf{W}\mathbf{B}\mathbf{x}}\right)1\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{W}\mathbf{B}\mathbf{x}} \leq 1\right) \\
&\quad + \left(2\,\mathrm{tr}\left(\mathbf{B}'\mathbf{K}\mathbf{V}\right) - \mathbf{x}'\mathbf{B}'\mathbf{K}\mathbf{B}\mathbf{x}\right)1\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{W}\mathbf{B}\mathbf{x}} > 1\right).
\end{aligned} \tag{37}$$

Proof: First, noting that $\boldsymbol{\eta}(\mathbf{x})$ is absolutely continuous,

$$\begin{aligned}
\mathrm{E}\left((\mathrm{Z}-\boldsymbol{\eta}(\mathrm{Z}+\boldsymbol{\delta}))'\mathbf{K}(\mathrm{Z}-\boldsymbol{\eta}(\mathrm{Z}+\boldsymbol{\delta}))\right) &= \mathrm{E}\left(\mathrm{Z}'\mathbf{K}\mathrm{Z}\right) - 2\mathrm{E}\left(\eta(\mathrm{Z}+\boldsymbol{\delta})'\mathbf{K}\mathrm{Z}\right) + \mathrm{E}\left(\eta(\mathrm{Z}+\boldsymbol{\delta})'\mathbf{K}\eta(\mathrm{Z}+\boldsymbol{\delta})\right) \\
&= \mathrm{tr}(\mathbf{K}\mathbf{V}) - \mathrm{E}\left(2\,\mathrm{tr}\left(\frac{\partial}{\partial \mathbf{x}}\eta(\mathrm{Z}+\boldsymbol{\delta})'\mathbf{K}\mathbf{V}\right) - \eta(\mathrm{Z}+\boldsymbol{\delta})'\mathbf{K}\eta(\mathrm{Z}+\boldsymbol{\delta})\right) \\
&= \mathrm{tr}(\mathbf{K}\mathbf{V}) - \mathrm{E}q(\mathrm{Z}+\boldsymbol{\delta})
\end{aligned}$$

the second equality using $\mathrm{E}\left(\mathrm{Z}'\mathbf{K}\mathrm{Z}\right)=\mathrm{tr}\left(\mathbf{WV}\right)$, Lemma 1, and $\mathrm{Z}\sim\mathrm{N}\left(\mathbf{0},\mathbf{V}\right)$, where

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

We now show that $q\left(\mathbf{x}\right)=g\left(\mathbf{x}\right)$. Note that

$$\frac{\partial}{\partial\mathbf{x}}\eta\left(\mathbf{x}\right)'=\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}\right)_1\mathbf{B}'-\frac{2C}{\left(\mathbf{x}'\mathbf{B}'\mathbf{WBx}\right)^2}\mathbf{B}'\mathbf{WBx}\mathbf{x}'\mathbf{B}'1\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}>1\right)$$

so

$$\mathrm{tr}\left(\frac{\partial}{\partial\mathbf{x}}\eta\left(\mathbf{x}\right)'\mathbf{KV}\right)=\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}\right)_1\mathrm{tr}\left(\mathbf{B}'\mathbf{KV}\right)-2C\frac{\left(\mathbf{x}'\mathbf{B}'\mathbf{KVB}'\mathbf{WBx}\right)}{\left(\mathbf{x}'\mathbf{B}'\mathbf{WBx}\right)^2}1\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}>1\right)$$

and also

$$\eta\left(\mathbf{x}\right)'\mathbf{K}\eta\left(\mathbf{x}\right)=\left(\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}\right)_1\right)^2\mathbf{x}'\mathbf{B}'\mathbf{KBx}.$$

Together

$$\begin{aligned}
q\left(\mathbf{x}\right) &= \left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}\right)_1\left(2\,\mathrm{tr}\left(\mathbf{B}'\mathbf{KV}\right)-\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}\right)_1\left(\mathbf{x}'\mathbf{B}'\mathbf{KBx}\right)\right)\\
&\quad -4C\frac{\left(\mathbf{x}'\mathbf{B}'\mathbf{KVB}'\mathbf{WBx}\right)}{\left(\mathbf{x}'\mathbf{B}'\mathbf{WBx}\right)^2}1\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}>1\right)\\
&= \frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}\left(2\,\mathrm{tr}\left(\mathbf{B}'\mathbf{KV}\right)-4\frac{\mathbf{x}'\mathbf{B}'\mathbf{KVB}'\mathbf{WBx}}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}-C\frac{\mathbf{x}'\mathbf{B}'\mathbf{KBx}}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}\right)1\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}\leq 1\right)\\
&\quad +\left(2\,\mathrm{tr}\left(\mathbf{B}'\mathbf{KV}\right)-\mathbf{x}'\mathbf{B}'\mathbf{KBx}\right)1\left(\frac{C}{\mathbf{x}'\mathbf{B}'\mathbf{WBx}}>1\right)\\
&= g\left(\mathbf{x}\right)
\end{aligned}$$

as claimed. The assumption that $\mathrm{rank}\left(\mathbf{B}'\mathbf{WB}\right)\geq 2$ means that $\mathbf{x}'\mathbf{B}'\mathbf{WBx}>0$ for $\mathbf{x}\neq\mathbf{0}$ so the expression for $g\left(\mathbf{x}\right)$ is well defined. ■

Proof of Theorem 1: First, Assumptions 1.1 and 2.1 imply $\boldsymbol{\theta}_n\to\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n\xrightarrow{p}\boldsymbol{\theta}_0$. Next, by two Taylor's expansions, Assumption 1.2 implies

$$\begin{aligned}
\sqrt{n}\mathbf{h}\left(\boldsymbol{\theta}_n\right) &= \sqrt{n}\mathbf{h}\left(\boldsymbol{\theta}_0\right)+\sqrt{n}\mathbf{H}'\left(\boldsymbol{\theta}_n-\boldsymbol{\theta}_0\right)+o(1)\\
&= \mathbf{H}'\boldsymbol{\delta}+o(1).
\end{aligned}$$

Assumption 2.1 states (8). Together with another Taylor expansion we find

$$\begin{aligned}\sqrt{n}\mathbf{h}(\hat{\boldsymbol{\theta}}_n) &= \sqrt{n}\mathbf{h}\left(\boldsymbol{\theta}_n\right) + \mathbf{H}'\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) + o_p(1) \\ &= \mathbf{H}'\boldsymbol{\delta} + \mathbf{H}'\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) + o_p(1) \\ &\xrightarrow{d} \mathbf{H}'\left(\mathrm{Z} + \boldsymbol{\delta}\right). \end{aligned} \tag{38}$$

Assumption 2.3 and (38) imply

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right) &= \mathbf{GH}\left(\mathbf{H}'\mathbf{GH}\right)^{-1}\sqrt{n}\,\mathbf{h}(\hat{\boldsymbol{\theta}}_n) + o_p(1) \\ &\xrightarrow{d} \mathbf{GH}\left(\mathbf{H}'\mathbf{GH}\right)^{-1}\mathbf{H}'\left(\mathrm{Z} + \boldsymbol{\delta}\right) \\ &= \mathbf{GP}_{\mathbf{G}}\left(\mathrm{Z} + \boldsymbol{\delta}\right) \end{aligned} \tag{39}$$

which is (9). Note that the condition $\operatorname{rank}(\mathbf{GH}) = r$ in Assumption 2.3 implies that

$$\mathbf{H}'\mathbf{GH} > 0 \tag{40}$$

so the expression in (39) is well defined. It follows that

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right) &= \sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) - \left(\frac{r-2}{n\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right)'\hat{\mathbf{V}}_n^{-1}\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right)}\right)_1 \sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right) \\ &\xrightarrow{d} \mathrm{Z} - \left(\frac{r-2}{\left(\mathrm{Z}+\boldsymbol{\delta}\right)'\mathbf{P}_{\mathbf{G}}\mathbf{G}\mathbf{V}^{-1}\mathbf{G}\mathbf{P}_{\mathbf{G}}\left(\mathrm{Z}+\boldsymbol{\delta}\right)}\right)_1 \mathbf{GP}_{\mathbf{G}}\left(\mathrm{Z}+\boldsymbol{\delta}\right) \qquad (41) \\ &= \mathrm{Z} - \boldsymbol{\eta}\left(\mathrm{Z}+\boldsymbol{\delta}\right) \qquad (42) \end{aligned}$$

where

$$\boldsymbol{\eta}\left(\mathbf{x}\right) = \left(\frac{r-2}{\mathbf{x}'\mathbf{W}_{\mathbf{G}}\mathbf{x}}\right)_1 \mathbf{GP}_{\mathbf{G}}\mathbf{x}. \tag{43}$$

and $\mathbf{W}_{\mathbf{G}}$ is defined in (11). Equation (41) is (10).

To evaluate the risk for $\hat{\boldsymbol{\theta}}_n$, observe that

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}_n) &= \mathrm{E}\left(\mathrm{Z}'\mathbf{V}^{-1}\mathrm{Z}\right) \\ &= \operatorname{tr}\left(\mathbf{V}^{-1}\mathrm{E}\left(\mathrm{ZZ}'\right)\right) \\ &= \operatorname{tr}\left(\mathbf{I}_k\right) \\ &= k \end{aligned}$$

which is (13).

To evaluate the risk for $\hat{\boldsymbol{\theta}}_n^*$ we can apply Lemma 2, with $C = r-2$, $\mathbf{B} = \mathbf{GP}_{\mathbf{G}}$ and $\mathbf{W} = \mathbf{K} = \mathbf{V}^{-1}$. Note that $\mathbf{B}'\mathbf{WB} = \mathbf{W}_{\mathbf{G}}$, and under Assumption 2.3, $\operatorname{rank}(\mathbf{W}_{\mathbf{G}}) = r > 2$, so the

conditions of the Lemma are satisfied. By (36) and (42),

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}_n^*) &= \mathrm{E}\left(\mathrm{Z} - \boldsymbol{\eta}\left(\mathrm{Z} + \boldsymbol{\delta}\right)\right)' \mathbf{V}^{-1} \left(\mathrm{Z} - \boldsymbol{\eta}\left(\mathrm{Z} + \boldsymbol{\delta}\right)\right) \\ &= k - \mathrm{E} g\left(\mathrm{Z} + \boldsymbol{\delta}\right) \end{aligned}$$

with $g(\mathbf{x})$ defined in (37). We now show that $g(\mathbf{x})$ equals $g_{\mathbf{G}}(\mathbf{x})$ as defined in (15), establishing (14). Indeed, observe that $\mathbf{KV} = \mathbf{I}$, $\mathbf{P_G G P_G G} = \mathbf{P_G G}$,

$$\mathbf{B'KVB'WB} = \mathbf{P_G G P_G G V^{-1} G P_G} = \mathbf{W_G},$$

and

$$\begin{aligned} \mathrm{tr}\left(\mathbf{B'KV}\right) &= \mathrm{tr}\left(\mathbf{P_G G}\right) \\ &= \mathrm{tr}\left(\mathbf{H}\left(\mathbf{H'GH}\right)^{-1}\mathbf{H'G}\right) \\ &= r. \end{aligned}$$

Making these substitutions, we directly find that $g(\mathbf{x}) = g_{\mathbf{G}}(\mathbf{x})$ as required.

Finally, observe that using the second indicator function in the definition of (15)

$$\begin{aligned} g_{\mathbf{G}}(\mathbf{x}) &\geq \frac{(r-2)^2}{\mathbf{x'W_G x}} 1\left(\frac{r-2}{\mathbf{x'W_G x}} \leq 1\right) + (r+2)\, 1\left(\frac{r-2}{\mathbf{x'W_G x}} > 1\right) \\ &> 0 \end{aligned}$$

the second inequality since $r > 2$. The inequality $g_{\mathbf{G}}(\mathbf{x}) > 0$ and (14) imply (16), completing the proof. ■

Proof of Theorem 2: The consistency of $\hat{\boldsymbol{\theta}}_n$ and Assumption 1.2 imply

$$\hat{\mathbf{H}} = \mathbf{H}(\hat{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{H}(\boldsymbol{\theta}_0) = \mathbf{H}. \tag{44}$$

Assumptions 2.2, 3, and (44) imply $\mathbf{A}_n \xrightarrow{p} \mathbf{A}$ and $C_n \xrightarrow{p} C$. Combined with (8) and (9) we obtain

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n^{**} - \boldsymbol{\theta}_n\right) \xrightarrow{d} \mathrm{Z} - \boldsymbol{\eta}\left(\mathrm{Z} + \boldsymbol{\delta}\right) \tag{45}$$

where

$$\boldsymbol{\eta}(\mathbf{x}) = \left(\frac{C}{\mathbf{x'W^* x}}\right)_1 \mathbf{GP_G x}. \tag{46}$$

and

$$\begin{aligned} \mathbf{W}^* &= \mathbf{P_G G W G P_G} \\ &= \mathbf{H}\left(\mathbf{H'GH}\right)^{-1}\left(\mathbf{H'GWGH}\right)\left(\mathbf{H'GH}\right)^{-1}. \end{aligned} \tag{47}$$

Observe that under Assumption 3.4 and (40),

$$\operatorname{rank}(\mathbf{W}^*) = r.$$

The asymptotic risk of the unrestricted estimator is simply

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}_n, \mathbf{W}) &= \mathrm{E}\mathrm{Z}'\mathbf{W}\mathrm{Z} \\ &= \operatorname{tr}\left(\mathbf{W}\mathrm{Z}\mathrm{Z}'\right) \\ &= \operatorname{tr}(\mathbf{W}\mathbf{V}). \end{aligned} \tag{48}$$

To evaluate the risk for $\hat{\boldsymbol{\theta}}_n^{**}$ we use Lemma 2, with $\mathbf{B} = \mathbf{G}\mathbf{P}_{\mathbf{G}}$ and $\mathbf{K} = \mathbf{W}$. Note that $\mathbf{B}'\mathbf{W}\mathbf{B} = \mathbf{W}^*$, which is full rank so the conditions of the Lemma are satisfied. By (36),

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}_n^{**}, \mathbf{W}) &= \mathrm{E}\left(\mathrm{Z} - \boldsymbol{\eta}(\mathrm{Z} + \boldsymbol{\delta})\right)' \mathbf{W} \left(\mathrm{Z} - \boldsymbol{\eta}(\mathrm{Z} + \boldsymbol{\delta})\right) \\ &= \operatorname{tr}(\mathbf{W}\mathbf{V}) - \mathrm{E}g(\mathrm{Z} + \boldsymbol{\delta}) \end{aligned} \tag{49}$$

where $g(\mathbf{x})$ is (37). Observe that

$$\begin{aligned} \operatorname{tr}\left(\mathbf{B}'\mathbf{K}\mathbf{V}\right) &= \operatorname{tr}(\mathbf{P}_{\mathbf{G}}\mathbf{G}\mathbf{W}\mathbf{V}) \\ &= \operatorname{tr}\left(\mathbf{H}\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{V}\right) \\ &= \operatorname{tr}\left(\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{V}\mathbf{H}\right) \\ &= \operatorname{tr}(\mathbf{A}) \end{aligned}$$

and

$$\mathbf{B}'\mathbf{K}\mathbf{V}\mathbf{B}'\mathbf{W}\mathbf{B} = \mathbf{P}_{\mathbf{G}}\mathbf{G}\mathbf{W}\mathbf{V}\mathbf{P}_{\mathbf{G}}\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{P}_{\mathbf{G}} = \mathbf{W}^{**},$$

say. Substituting in (37) we find

$$\begin{aligned} g(\mathbf{x}) &= \frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}}\left(2\operatorname{tr}(\mathbf{A}) - 4\frac{\mathbf{x}'\mathbf{W}^{**}\mathbf{x}}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} - C\right) 1\left(\frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} \leq 1\right) \\ &\quad + \left(2\operatorname{tr}(\mathbf{A}) - \mathbf{x}'\mathbf{W}^*\mathbf{x}\right) 1\left(\frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} > 1\right) \end{aligned} \tag{50}$$

Now set $\mathbf{y} = \left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H}\right)^{1/2}\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{x}$ (which is well defined under Assumption 3.4 and (40), so that

$$\begin{aligned}
\frac{\mathbf{x}'\mathbf{W}^{**}\mathbf{x}}{\mathbf{x}'\mathbf{W}^{*}\mathbf{x}} &= \frac{\mathbf{x}'\mathbf{H}\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{V}\mathbf{H}\right)\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H}\right)\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{x}}{\mathbf{x}'\mathbf{H}\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H}\right)\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{x}} \\
&= \frac{\mathbf{y}'\left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H}\right)^{1/2}\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H}\right)^{1/2}\mathbf{y}}{\mathbf{y}'\mathbf{y}} \\
&\leq \lambda_{\max}\left(\left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H}\right)^{1/2}\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\left(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H}\right)^{1/2}\right) \\
&= \lambda_{\max}\left(\mathbf{A}\right). \qquad (51)
\end{aligned}$$

Using (50), (51), the second indicator function, and the definition (19), we see that

$$\begin{aligned}
g\left(\mathbf{x}\right) &\geq \frac{C}{\mathbf{x}'\mathbf{W}^{*}\mathbf{x}}\left(2\operatorname{tr}\left(\mathbf{A}\right) - 4\lambda_{\max}\left(\mathbf{A}\right) - C\right)1\left(\frac{C}{\mathbf{x}'\mathbf{W}^{*}\mathbf{x}} \leq 1\right) \\
&\quad + \left(2\operatorname{tr}\left(\mathbf{A}\right) - C\right)1\left(\frac{C}{\mathbf{x}'\mathbf{W}^{*}\mathbf{x}} > 1\right) \\
&\geq C\left(\frac{C}{\mathbf{x}'\mathbf{W}^{*}\mathbf{x}}\right)_1 \\
&> 0
\end{aligned}$$

the last inequality using $C > 0$. Together with (48) and (49), $g\left(\mathbf{x}\right) > 0$ implies that

$$R(\hat{\boldsymbol{\theta}}_n^{**}, \mathbf{W}) = R(\hat{\boldsymbol{\theta}}_n, \mathbf{W}) - \mathrm{E}g\left(\mathrm{Z} + \boldsymbol{\delta}\right) < R(\hat{\boldsymbol{\theta}}_n, \mathbf{W})$$

which is (24). ■

Proof of Theorem 3: The asymptotic risk of $\hat{\boldsymbol{\theta}}_n$ is $R(\hat{\boldsymbol{\theta}}_n, \mathbf{K}) = \operatorname{tr}\left(\mathbf{K}\mathbf{V}\right)$. By Lemma 2 that for $\hat{\boldsymbol{\theta}}_n^{**}$ is

$$\begin{aligned}
R(\hat{\boldsymbol{\theta}}_n^{**}, \mathbf{K}) &= \mathrm{E}\left(\mathrm{Z} - \boldsymbol{\eta}\left(\mathrm{Z} + \boldsymbol{\delta}\right)\right)'\mathbf{K}\left(\mathrm{Z} - \boldsymbol{\eta}\left(\mathrm{Z} + \boldsymbol{\delta}\right)\right) \\
&= \operatorname{tr}\left(\mathbf{K}\mathbf{V}\right) - \mathrm{E}g\left(\mathrm{Z} + \boldsymbol{\delta}\right) \\
&= R(\hat{\boldsymbol{\theta}}_n, \mathbf{K}) - \mathrm{E}g\left(\mathrm{Z} + \boldsymbol{\delta}\right)
\end{aligned}$$

with $\mathbf{B} = \mathbf{G}\mathbf{P}_{\mathbf{G}}$ in (37). It is sufficient to show that $g\left(\mathbf{x}\right) > 0$ to complete the proof.

Note that

$$\begin{aligned}
\operatorname{tr}\left(\mathbf{B}'\mathbf{K}\mathbf{V}\right) &= \operatorname{tr}\left(\mathbf{P}_{\mathbf{G}}\mathbf{G}\mathbf{K}\mathbf{V}\right) \\
&= \operatorname{tr}\left(\mathbf{H}\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{G}\mathbf{K}\mathbf{V}\right) \\
&= \operatorname{tr}\left(\left(\mathbf{H}'\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{G}\mathbf{K}\mathbf{V}\mathbf{H}\right) \\
&= \operatorname{tr}\left(\mathbf{A}^{*}\right)
\end{aligned}$$

and

$$\mathbf{B}'\mathbf{K}\mathbf{V}\mathbf{B}'\mathbf{W}\mathbf{B} = \mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{V}\mathbf{P_G}\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{P_G}.$$

Substituting in (37) we find

$$\begin{aligned} g(\mathbf{x}) &= \frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}}\left(2\operatorname{tr}(\mathbf{A}^*) - 4\frac{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{V}\mathbf{P_G}\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{P_G}\mathbf{x}}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} - C\frac{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{G}\mathbf{P_G}\mathbf{x}}{\mathbf{x}'\mathbf{W}^*\mathbf{x}}\right)1\left(\frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} \leq 1\right) \\ &\quad + \left(2\operatorname{tr}(\mathbf{A}^*) - \mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{G}\mathbf{P_G}\mathbf{x}\right)1\left(\frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} > 1\right) \end{aligned}$$

where $\mathbf{W}^*$ is defined in (47).

Again setting $\mathbf{y} = (\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H})^{1/2}(\mathbf{H}'\mathbf{G}\mathbf{H})^{-\mathbf{1}}\mathbf{H}'\mathbf{x}$, observe that

$$\begin{aligned} \frac{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{V}\mathbf{P_G}\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{P_G}\mathbf{x}}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} &= \frac{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{V}\mathbf{P_G}\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{P_G}\mathbf{x}}{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{P_G}\mathbf{x}} \\ &= \frac{\mathbf{y}'(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H})^{-1/2}(\mathbf{H}'\mathbf{G}\mathbf{K}\mathbf{V}\mathbf{H})(\mathbf{H}'\mathbf{G}\mathbf{H})^{-1}(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H})^{1/2}\mathbf{y}}{\mathbf{y}'\mathbf{y}} \\ &\leq \lambda_{\max}\left((\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H})^{-1/2}(\mathbf{H}'\mathbf{G}\mathbf{K}\mathbf{V}\mathbf{H})(\mathbf{H}'\mathbf{G}\mathbf{H})^{-1}(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H})^{1/2}\right) \\ &= \lambda_{\max}(\mathbf{A}^*) \end{aligned}$$

and

$$\begin{aligned} \frac{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{G}\mathbf{P_G}\mathbf{x}}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} &= \frac{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{G}\mathbf{P_G}\mathbf{x}}{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{P_G}\mathbf{x}} \\ &= \frac{\mathbf{y}'(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H})^{-1/2}(\mathbf{H}'\mathbf{G}\mathbf{K}\mathbf{G}\mathbf{H})(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H})^{-1/2}\mathbf{y}}{\mathbf{y}'\mathbf{y}} \\ &\leq \lambda_{\max}\left((\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H})^{-1/2}(\mathbf{H}'\mathbf{G}\mathbf{K}\mathbf{G}\mathbf{H})(\mathbf{H}'\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{H})^{-1/2}\right) \\ &= \lambda_{\max}(\mathbf{A}^{**}). \end{aligned}$$

Using these bounds, the indicator function, and (25) we find

$$\begin{aligned} g(\mathbf{x}) &= \frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}}\left(2\operatorname{tr}(\mathbf{A}^*) - 4\frac{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{V}\mathbf{P_G}\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{P_G}\mathbf{x}}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} - C\frac{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{G}\mathbf{P_G}\mathbf{x}}{\mathbf{x}'\mathbf{W}^*\mathbf{x}}\right)1\left(\frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} \leq 1\right) \\ &\quad + \left(2\operatorname{tr}(\mathbf{A}^*) - \mathbf{x}'\mathbf{W}^*\mathbf{x}\frac{\mathbf{x}'\mathbf{P_G}\mathbf{G}\mathbf{K}\mathbf{G}\mathbf{P_G}\mathbf{x}}{\mathbf{x}'\mathbf{W}^*\mathbf{x}}\right)1\left(\frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} > 1\right) \\ &\geq \frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}}\left(2\operatorname{tr}(\mathbf{A}^*) - 4\lambda_{\max}(\mathbf{A}^*) - C\lambda_{\max}(\mathbf{A}^{**})\right)1\left(\frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} \leq 1\right) \\ &\quad + \left(2\operatorname{tr}(\mathbf{A}^*) - C\lambda_{\max}(\mathbf{A}^{**})\right)1\left(\frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}} > 1\right) \\ &\geq \left(\frac{C}{\mathbf{x}'\mathbf{W}^*\mathbf{x}}\right)_1\left(2\operatorname{tr}(\mathbf{A}^*) - 4\lambda_{\max}(\mathbf{A}^*) - C\lambda_{\max}(\mathbf{A}^{**})\right) \\ &> 0. \end{aligned}$$

This completes the proof. ■

Proof of Theorem 4: Taking the first-order conditions for $\hat{\boldsymbol{\theta}}_n$, expanding in a standard Taylor expansion, and applying Assumptions 1 and 4, we find

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) &= -\mathbf{M}^{-1}\mathbf{S}_n(\boldsymbol{\theta}_n) \\ &\xrightarrow{d} \mathbf{M}^{-1}\mathrm{N}\left(0, \boldsymbol{\Omega}\right) \\ &= \mathrm{Z} \sim \mathrm{N}\left(0, \mathbf{V}\right)\end{aligned}$$

which is Assumption 2.1. Combined with Assumption 1.1 it follows that $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$. Assumption 4.2 implies $\hat{\mathbf{M}}_n \xrightarrow{p} \mathbf{M}$ and with Assumption 4.5 we obtain $\hat{\mathbf{V}}_n \xrightarrow{p} \mathbf{V}$ which is Assumption 2.2.

The estimator $\tilde{\boldsymbol{\theta}}_n$ solves the Lagrange multiplier problem

$$\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \left[Q_n(\boldsymbol{\theta}) + \sqrt{n}\boldsymbol{\lambda}'\mathbf{h}\left(\boldsymbol{\theta}\right)\right]$$

which has the first-order conditions

$$\mathbf{0} = \mathbf{S}_n(\tilde{\boldsymbol{\theta}}_n) + \tilde{\mathbf{H}}\tilde{\boldsymbol{\lambda}} \tag{52}$$

$$\mathbf{0} = \mathbf{h}(\tilde{\boldsymbol{\theta}}_n) \tag{53}$$

where $\tilde{\mathbf{H}} = \mathbf{H}(\tilde{\boldsymbol{\theta}}_n)$. We expand both equations by Taylor expansions about $\hat{\boldsymbol{\theta}}_n$ and use $\mathbf{S}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$ to find

$$\mathbf{0} = \mathbf{M}\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n\right) + \mathbf{H}\tilde{\boldsymbol{\lambda}} + o_p(1) \tag{54}$$

and

$$\mathbf{0} = \sqrt{n}\,\mathbf{h}(\hat{\boldsymbol{\theta}}_n) + \mathbf{H}'\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n\right) + o_p(1). \tag{55}$$

Premultiplying (54) by $\mathbf{H}'\mathbf{M}^{-1}$ and solving, and then using (55) we find

$$\begin{aligned}\tilde{\boldsymbol{\lambda}} &= -\left(\mathbf{H}'\mathbf{M}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n\right) + o_p(1) \\ &= \left(\mathbf{H}'\mathbf{M}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\sqrt{n}\,\mathbf{h}(\hat{\boldsymbol{\theta}}_n) + o_p(1).\end{aligned}$$

Substituting into (54) we find

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n\right) = -\mathbf{M}^{-1}\mathbf{H}\left(\mathbf{H}'\mathbf{M}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\sqrt{n}\,\mathbf{h}(\hat{\boldsymbol{\theta}}_n) + o_p(1)$$

which is Assumption 2.3 with $\mathbf{G} = \mathbf{M}^{-1}$. Since $\mathbf{M}^{-1} > 0$ and $\mathrm{rank}\left(\mathbf{H}\right) = r$ then $\mathrm{rank}\left(\mathbf{GH}\right) = r$ as required.

We have shown that Assumptions 1 and 4 imply Assumptions 2.1, 2.2 and 2.3, as stated. ■

References

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

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

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

[4] James W. and Charles M. Stein (1961): "Estimation with quadratic loss," *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, 1, 361-380.

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

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

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

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

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

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

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

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