Weak versus strong dominance of shrinkage estimators *

Giuseppe De Luca
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy
E-mail: giuseppe.deluca@unipa.it

Jan R. Magnus
Department of Econometrics and Data Science, Vrije Universiteit Amsterdam,
and Tinbergen Institute, The Netherlands
E-mail: jan@janmagnus.nl

January 14, 2021

Abstract

We consider the estimation of the mean of a multivariate normal distribution with known variance. Most studies consider the risk of competing estimators, that is the trace of the mean squared error matrix. In contrast we consider the whole mean squared error matrix, in particular its eigenvalues. We prove that there are only two distinct eigenvalues and apply our findings to the James–Stein and the Thompson class of estimators. It turns out that the famous Stein paradox is no longer a paradox when we consider the whole mean squared error matrix rather than only its trace.

Keywords: Shrinkage, Dominance, James–Stein.

JEL classification: C13,C51.

*Giuseppe De Luca acknowledges financial support from the MIUR PRIN PRJ-0324.

1 Introduction

Consider p independent normally distributed observations $x_1, x_2, \ldots, x_p$ where x_i has an unknown mean θ_i and a known variance σ_i^2. Since the variances are known, there is no loss in generality by setting $\sigma_i^2 = 1$. In other words, $x \sim \mathrm{N}(\theta, I_p)$, where $x = (x_1, \ldots, x_p)'$, $\theta = (\theta_1, \ldots, \theta_p)'$, and I_p denotes the identity matrix of order p. This is the so-called (multivariate) normal location model and its relevance has been much discussed; see, for example, Johnstone (2019, Chapters 1 and 2).

Our purpose is to estimate the vector θ and we shall consider 'shrinkage' estimators of the form

$$\hat{\theta} = \lambda(w_p)x, \tag{1}$$

where λ depends on x only through $w_p = x'x$. Our paper is an extension of Hansen (2015) who studied the same class of estimators, concentrating on the efficiency bound for minimax estimators and their performance in terms of minimax regret.

We shall think of λ as a shrinkage factor ($0 \leq \lambda \leq 1$). In the cases studied below it will always be the case that $\lambda \leq 1$, but it will not always be the case that $\lambda \geq 0$ as we shall see when we define the James–Stein estimator. The univariate random variable w_p follows a noncentral χ^2 distribution with p degrees of freedom and noncentrality parameter $\theta'\theta$, which we write as $w_p \sim \chi_p^2(\theta'\theta)$. If we replace x by $y = Sx$ where S is an orthogonal matrix, then $y'y = x'x$ so that λ remains the same. Hence λ is orthogonally invariant, and the class of estimators defined by (1) is called the class of *orthogonally invariant estimators*.

The mean squared error (MSE) of $\hat{\theta}$ is the positive semidefinite $p \times p$ matrix

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'] \tag{2}$$

and its trace is called the *risk* of the estimator.

In the class of orthogonally invariant estimators we have

$$\sup_{\theta} \frac{\operatorname{tr} \mathrm{MSE}(\hat{\theta})}{p} \geq 1 \tag{3}$$

for all λ (Van der Vaart, 1998, Proposition 8.6). When equality occurs in (3) then $\hat{\theta}$ is called *minimax*. For $p = 1$ and $p = 2$ there is a unique minimax estimator, namely the maximum likelihood (ML) estimator $\hat{\theta}_{ML} = x$, sometimes called the 'usual' estimator. But for $p \geq 3$ there are many minimax estimators; see Hansen (2015) for further discussion and references.

Condition (3) refers to one aspect of the MSE matrix, namely its trace or, equivalently, its average eigenvalue. But there is more to the MSE matrix than just its trace, and in this paper

our focus is on the whole MSE matrix rather than on one aspect of it. In particular, we could be interested in $\max_i \mathrm{MSE}(\hat{\theta}_i)$ or $\max \nu_i(\mathrm{MSE}(\hat{\theta}))$, where $\nu_1() \geq \cdots \geq \nu_p()$ denote the eigenvalues. The fact that the largest individual mean squared error is relevant was well formulated by Lehmann and Casella (1998, p. 363) who write:

> "No one wants his or her blood test subjected to the possibility of large errors in order to improve a laboratory's average performance."

The largest eigenvalue is relevant too, because it provides an upper bound for arbitrary linear combinations:

$$\mathrm{MSE}(w'\hat{\theta}) = w' \,\mathrm{MSE}(\hat{\theta}) w \leq (w'w) \max_{1 \leq i \leq p} \nu_i(\mathrm{MSE}(\hat{\theta})). \tag{4}$$

To set the scene, let us first compare two estimators: the ML estimator (where $\lambda \equiv 1$) and the 'silly' estimator $\hat{\theta}_0 = 0$ (where $\lambda \equiv 0$). The ML estimator has zero bias and variance I_p so that its MSE matrix is $\mathrm{MSE}_{ML} = I_p$, while the silly estimator has bias $-\theta$ and zero variance so that $\mathrm{MSE}_0 = \theta\theta'$. We are interested in the difference

$$\Delta = \mathrm{MSE}_{ML} - \mathrm{MSE}_0 = I_p - \theta\theta'.$$

The eigenvalues of Δ are $1 - \theta'\theta$ (multiplicity 1) and 1 (multiplicity $p-1$). Thus, $\operatorname{tr}\Delta \geq 0$ if and only if $\theta'\theta \leq p$, while $\Delta \geq 0$ (Δ is positive semidefinite) if and only if $\theta'\theta \leq 1$. Put differently, the average eigenvalue of Δ is nonnegative if and only if $\theta'\theta \leq p$, while all eigenvalues are nonnegative if and only if $\theta'\theta \leq 1$.

This is illustrated in the left panel of Figure 1. For each p the line segment AB contains the points where $\Delta \geq 0$, while the larger line segment AC contains the points where $\operatorname{tr}\Delta \geq 0$. On the line segment BC the weaker condition holds but not the stronger. The region where the silly estimator performs better than the ML estimator thus depends on which criterion is used, in other words what we mean by 'better'.

The situation is rather different with the James–Stein (JS) estimator

$$\hat{\theta}_{JS} = \left(1 - \frac{c}{x'x}\right) x \qquad (p \geq 3,\, c \geq 0), \tag{5}$$

graphed in the right panel of Figure 1 for $c = p - 2$. In this case,

$$\Delta = \mathrm{MSE}_{ML} - \mathrm{MSE}_{JS} = I_p - \mathrm{MSE}_{JS}$$

and $\operatorname{tr}\Delta \geq 0$ for all θ so that the point C is at infinity. This is the Stein paradox which states that for $p \geq 3$ the ML estimator is not admissible because there is another estimator, namely the JS

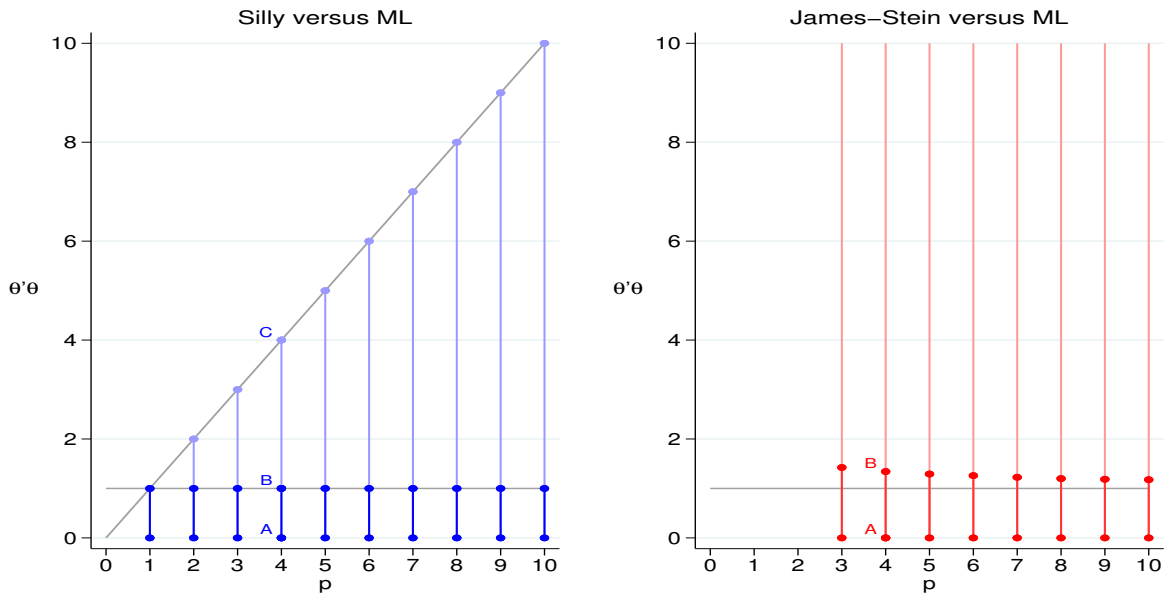

Figure 1: Comparison of silly versus ML (left) and James–Stein versus ML (right) estimators

estimator, which is uniformly (that is, for every $\theta'\theta$) better than the ML estimator. But, again, it depends on what we mean by 'better'. If we use the stronger criterion $\Delta \geq 0$, then the JS estimator is only better on the interval AB, that is for small values of $\theta'\theta$, but not otherwise. In other words, according to the stronger criterion, the JS estimator is not uniformly better than the ML estimator, or vice versa.

In this paper we shall explore the difference between the weaker and the stronger criterion. In Section 2 we formally define and discuss weak and strong dominance. In Section 3 we show that, under mild regularity conditions on λ, the MSE matrix of any orthogonally invariant estimator has only two distinct eigenvalues: ν_1 (multiplicity 1) and ν_2 (multiplicity $p-1$). In Section 4 we derive alternative expressions for the MSE matrix and the two eigenvalues, using Stein's lemma. In Sections 5 and 6 we specialize these results to two important classes of shrinkage estimators: the JS estimator (5) and the Thompson estimator

$$\hat{\theta}_{Th} = \frac{x'x}{c + x'x}\, x \qquad (c \geq 0). \tag{6}$$

We compare the properties of these two classes in Section 7. In Section 8 we confront the (joint) Thompson estimator with its separate counterpart where each component is estimated separately. In other words, we ask whether joint estimation is indeed useful, which is the essence of the Stein

paradox. Section 9 concludes. There are two appendices. Appendix A states two results concerning idempotent matrices, while Appendix B provides three versions of Stein's lemma (in increasing generality).

2 Weak and strong dominance

Thus motivated, let x be a single observation from the p-variate normal distribution with mean θ and variance I_p, and let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of θ with mean squared error matrices MSE_1 and MSE_2, respectively. Let

$$\Delta(\theta) = \mathrm{MSE}_2 - \mathrm{MSE}_1$$

denote the difference of the two MSE matrices. If $\operatorname{tr}\Delta(\theta) \geq 0$ for all $\theta \in \Theta$ with strict inequality for at least one value of $\theta \in \Theta$, then we say that $\hat{\theta}_1$ *weakly* dominates $\hat{\theta}_2$ on Θ. If $\hat{\theta}_1$ weakly dominates $\hat{\theta}_2$ for all θ, then we say that $\hat{\theta}_1$ weakly dominates $\hat{\theta}_2$.

Similarly, if $\Delta(\theta)$ is positive semidefinite for all $\theta \in \Theta$ with $\Delta(\theta) \neq 0$ for at least one $\theta \in \Theta$, then we say that $\hat{\theta}_1$ *strongly* dominates $\hat{\theta}_2$ on Θ. If $\hat{\theta}_1$ strongly dominates $\hat{\theta}_2$ for all θ, then we say that $\hat{\theta}_1$ strongly dominates $\hat{\theta}_2$.

Strong dominance implies weak dominance, but not vice versa. Weak dominance requires that the *average* eigenvalue of $\Delta(\theta)$ is ≥ 0 for all θ, while strong dominance requires that *all* eigenvalues of $\Delta(\theta)$ are ≥ 0.

Since the trace is a linear operator we can equivalently say that $\hat{\theta}_1$ weakly dominates $\hat{\theta}_2$ if $\operatorname{tr}\mathrm{MSE}_1 \leq \operatorname{tr}\mathrm{MSE}_2$ for all θ with strict inequality for at least one θ or, written differently, if

$$\mathrm{E}[(\hat{\theta}_1 - \theta)'(\hat{\theta}_1 - \theta)] \leq \mathrm{E}[(\hat{\theta}_2 - \theta)'(\hat{\theta}_2 - \theta)]$$

for all θ with strict inequality for at least one θ.

In contrast to the trace, eigenvalues are not linear operators and so it is, in general, not true that $\nu_i(\Delta) = \nu_i(\mathrm{MSE}_2) - \nu_i(\mathrm{MSE}_1)$. However, in the special case

$$\mathrm{MSE}_1 = \nu_1 J + \nu_2(I_p - J), \qquad \mathrm{MSE}_2 = \xi_1 J + \xi_2(I_p - J),$$

where $J = \theta\theta'/\theta'\theta$ for $\theta \neq 0$, it follows from Lemma 1 in Appendix A that the eigenvalues of MSE_1 are ν_1 (multiplicity 1) and ν_2 (multiplicity $p-1$), the eigenvalues of MSE_2 are ξ_1 (multiplicity 1) and ξ_2 (multiplicity $p-1$), and the eigenvalues of $\Delta = \mathrm{MSE}_2 - \mathrm{MSE}_1$ are $\xi_1 - \nu_1$ (multiplicity 1) and $\xi_2 - \nu_2$ (multiplicity $p-1$). This special case is the typical situation in the current paper (except in Section 8) because the MSE matrices that we shall encounter have only two distinct eigenvalues

where the larger eigenvalue has multiplicity 1 and the smaller eigenvalue has multiplicity $p-1$; see Proposition 1. Weak dominance is then essentially determined by the smaller eigenvalue, while strong dominance is essentially determined by the larger eigenvalue.

Although the trace criterion (the sum or equivalently the arithmetic mean of the eigenvalues of Δ) for weak dominance is widely used, an alternative would be the determinant of Δ (the product of the eigenvalues) or its pth root (geometric mean of the eigenvalues). One could also extend the definition of weak dominance by introducing a positive semidefinite weight matrix W, and say that $\hat{\theta}_1$ weakly dominates $\hat{\theta}_2$ with respect to W if $\operatorname{tr}(W\Delta(\theta)) \geq 0$ for all θ with strict inequality for at least one θ, that is, if

$$\mathrm{E}[(\hat{\theta}_1 - \theta)'W(\hat{\theta}_1 - \theta)] \leq \mathrm{E}[(\hat{\theta}_2 - \theta)'W(\hat{\theta}_2 - \theta)]$$

for all θ with strict inequality for at least one θ. If $\hat{\theta}_1$ weakly dominates $\hat{\theta}_2$ with respect to W *for all* W, then $\hat{\theta}_1$ strongly dominates $\hat{\theta}_2$; see also Saleh (2006, Section 1.3).

Most authors only study weak dominance, but strong dominance is important too, particularly if we are not only interested in the estimator $\hat{\theta}$ but also (or primarily) in linear combinations, say $w'\hat{\theta}$ for some given vector w. This situation occurs whenever the normal location model is obtained after preliminary transformations of the original model. Two recent examples are weighted-average least squares (Magnus and De Luca, 2016) and wavelet shrinkage estimators (Johnstone, 2019, Chapter 7). In such cases we want to know whether $\hat{\theta}_1$ dominates $\hat{\theta}_2$ weakly with respect to ww'. In many cases w is not known in advance, in which case we want to know whether $\hat{\theta}_1$ dominates $\hat{\theta}_2$ weakly with respect to all ww', that is, whether $\hat{\theta}_1$ strongly dominates $\hat{\theta}_2$. While for weak dominance the average eigenvalue matters, for strong dominance it is the largest eigenvalue which matters.

When $p \geq 3$ there are many orthogonally invariant estimators which weakly dominate the ML estimator (or, put differently, are minimax), for example the JS estimator. But are there also orthogonally invariant estimators which strongly dominate the ML estimator? Probably not.

Conjecture *In the class of orthogonally invariant estimators no estimator $\hat{\theta}$ strongly dominates the ML estimator.*

If an estimator $\hat{\theta}_*$ exists which strongly dominates the ML estimator, then all eigenvalues of $\mathrm{MSE}(\hat{\theta}_*)$ must be ≤ 1 for all θ. Also, from (3), $\sup_\theta \nu_1(\theta) \geq \sup_\theta \bar{\nu}(\theta) \geq 1$, where $\nu_1(\theta)$ denotes the largest eigenvalue of $\mathrm{MSE}(\hat{\theta}_*)$ and $\bar{\nu}(\theta)$ the average eigenvalue. Hence, $\sup_\theta \nu_1(\theta) = 1$. The conjecture claims that the only estimator satisfying $\sup_\theta \nu_1(\theta) = 1$ is the ML estimator.

The conjecture clearly holds for the JS estimator as shown in the right panel of Figure 1 because the point B is finite for all values of p. We shall provide further evidence (but no proof) supporting the conjecture as we proceed.

3 Structure and eigenvalues of the MSE matrix

To make further progress, we shall place the following restrictions on λ.

Assumption A *The function $\lambda(w_p)$ is absolutely continuous and nondecreasing on $[0, \infty)$, and $\mathrm{E}\,|\lambda'(w_p)|$ is finite.*

The requirement of absolute continuity is stronger than (uniform) continuity but weaker than continuous differentiability, and allows kinks (soft thresholding) but not jumps (hard thresholding); see Candès, Sing-Long, and Trzasko (2013), Tibshirani (2015), and Mikkelsen and Hansen (2018) for details. The assumption that λ is nondecreasing is intuitive when we go back to the ML estimator $\hat{\theta}_{ML} = x$ (where $\lambda \equiv 1$) and the silly estimator $\hat{\theta}_0 = 0$ (where $\lambda \equiv 0$). The silly estimator dominates weakly for $\theta'\theta \leq p$ and strongly for $\theta'\theta \leq 1$. This suggest a 'pretest' estimator where

$$\lambda_{pt} = \begin{cases} 0 & \text{if } w_p \leq c, \\ 1 & \text{if } w_p > c, \end{cases}$$

for some $c \geq 0$. The pretest estimator is not a satisfactory estimator, but continuous versions of it might be. These continuous versions would be a weighted average of the ML and silly estimators such that the larger is w_p the more weight is given to the ML estimator. In our case, both the James–Stein and the Thompson class of estimators satisfy the requirement that λ is nondecreasing; see Casella (1990) for a discussion in the context of the JS estimator. We note that Proposition 1 does not depend on the assumption that λ is nondecreasing, except to show that $\nu_1 \geq \nu_2$.

If λ satisfies Assumption A, we say that is belongs to the $\mathcal{L}$-class. Well-known examples in the $\mathcal{L}$-class are $\lambda(w_p) = 1$, the ML estimator, and $\lambda(w_p) = 1 - c/w_p$, the JS estimator (when $c = p - 2$).

From Bock (1975, Theorems A and B) we know that for any function $\psi : [0, \infty) \to (-\infty, \infty)$ we have

$$\mathrm{E}[\psi(w_p)x_i] = \theta_i\,\mathrm{E}[\psi(w_{p+2})], \tag{7a}$$

$$\mathrm{E}[\psi^2(w_p)x_i^2] = \mathrm{E}[\psi^2(w_{p+2})] + \theta_i^2\,\mathrm{E}[\psi^2(w_{p+4})], \tag{7b}$$

$$\mathrm{E}[\psi^2(w_p)x_ix_j] = \theta_i\theta_j\,\mathrm{E}[\psi^2(w_{p+4})] \quad (i \neq j), \tag{7c}$$

where w_p, w_{p+2}, and w_{p+4} denote noncentral χ^2 random variables with common noncentrality $\theta'\theta$ and degrees of freedom p, $p+2$, and $p+4$, respectively. The third equality is in fact not proved in Bock (1975) but the proof is similar to the proof of the second equality; see also Saleh (2006, Section 2.2, Theorem 7).

These facts lead to the following result.

Proposition 1 *The bias and MSE of* $\hat{\theta}$ *are*

$$\text{bias}(\hat{\theta}) = (\text{E}[\lambda(w_{p+2})] - 1)\,\theta$$

and

$$\text{MSE}(\hat{\theta}) = \nu_1 J + \nu_2 (I_p - J),$$

where $J = \theta\theta'/\theta'\theta$ *and* $I_p - J$ *are idempotent matrices of rank* 1 *and* $p-1$ *respectively. The eigenvalues of* $\text{MSE}(\hat{\theta})$ *are* $\nu_1 \geq \nu_2 \geq 0$*, where*

$$\nu_1 = \text{E}[\lambda^2(w_{p+2})] + (\theta'\theta)\left(1 - 2\,\text{E}[\lambda(w_{p+2})] + \text{E}[\lambda^2(w_{p+4})]\right)$$

and

$$\nu_2 = \text{E}[\lambda^2(w_{p+2})]$$

have multiplicities 1 *and* $p-1$*, respectively. The two eigenvalues coincide if and only if* $\theta = 0$ *or* $\lambda \equiv 1$*, in which case the bias is zero and* $\text{MSE}(\hat{\theta}) = \nu I_p$*, where* $\nu = \text{E}[\lambda^2(w_{p+2})]$ *when* $\theta = 0$ *and* $\nu = 1$ *when* $\lambda \equiv 1$*.*

Proof From (7a) we find the bias as

$$\text{bias}(\hat{\theta}) = \text{E}[\hat{\theta} - \theta] = \text{E}[\lambda(w_p)x - \theta] = (\text{E}[\lambda(w_{p+2})] - 1)\,\theta.$$

Similarly, from (7b) and (7c),

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \text{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'] \\
&= \text{E}[\lambda^2(w_p)xx'] - \text{E}[\lambda(w_p)x]\theta' - \theta\,\text{E}[\lambda(w_p)x'] + \theta\theta' \\
&= \text{E}[\lambda^2(w_{p+2})]I_p + \text{E}[\lambda^2(w_{p+4})]\theta\theta' - 2\,\text{E}[\lambda(w_{p+2})]\theta\theta' + \theta\theta' \\
&= \nu_1 J + \nu_2 (I_p - J).
\end{aligned}$$

Since J is symmetric idempotent with rank $r(J) = 1$, it follows from Lemma 1 in Appendix A that ν_1 and ν_2 are the eigenvalues of $\text{MSE}(\hat{\theta})$ with multiplicities 1 and $p-1$, respectively.

Finally, $\nu_1 \geq \nu_2$ if and only if

$$\begin{aligned}
&\mathrm{E}\left[1 - 2\lambda(w_{p+2}) + \lambda^2(w_{p+4})\right] \\
&\quad = \mathrm{E}\left[(1 - \lambda(w_{p+4}))^2\right] + 2\,\mathrm{E}[\lambda(w_{p+4}) - \lambda(w_{p+2})] \\
&\quad = \mathrm{E}\left[(1 - \lambda(w_{p+4}))^2\right] + 4\,\mathrm{E}[\lambda'(w_{p+4})] \geq 0,
\end{aligned}$$

where the second equality follows from (11) (to be proved later). Since λ is nondecreasing, we have $\lambda' \geq 0$ and hence $\mathrm{E}[\lambda'] \geq 0$.

Proposition 1 generalizes Saleh (2006, Eq. 4.3.20) who obtained the MSE matrix of the JS estimator (in a slightly different but equivalent form), but not the eigenvalues. The proposition shows that any estimator in the $\mathcal{L}$-class has a MSE matrix with only two distinct eigenvalues: ν_1 with multiplicity 1 and ν_2 with multiplicity $p-1$. It also gives explicit expressions for these two eigenvalues. From the two eigenvalues it is easy to obtain the trace of the MSE matrix as

$$\begin{aligned}
\operatorname{tr}\mathrm{MSE}(\hat{\theta}) &= \nu_1 + (p-1)\nu_2 \\
&= p\,\mathrm{E}[\lambda^2(w_{p+2})] + (\theta'\theta)\left(1 - 2\,\mathrm{E}[\lambda(w_{p+2})] + \mathrm{E}[\lambda^2(w_{p+4})]\right) \qquad (8)
\end{aligned}$$

and the determinant as $\nu_1\nu_2^{p-1}$.

The two eigenvalues $\nu_1 \geq \nu_2 \geq 0$ depend only on p and $\theta'\theta$, and they completely characterize the MSE matrix. This is consistent with (and more general than) the well-known fact that the risk (i.e. the trace of the MSE matrix) of $\mathcal{L}$-class estimators depends on θ only through p and $\theta'\theta$ (see, e.g., Hansen, 2015, Theorem 1).

4 Alternative route using Stein's lemma

There is an alternative route based on Stein's lemma (see Appendix B). Starting from the basic equality

$$\begin{aligned}
(\hat{\theta} - x)(\hat{\theta} - x)' = (\hat{\theta} - \theta)(\hat{\theta} - \theta)' + (x - \theta)(x - \theta)' \\
- (\hat{\theta} - \theta)(x - \theta)' - (x - \theta)(\hat{\theta} - \theta)',
\end{aligned}$$

we can write the MSE matrix as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - x)(\hat{\theta} - x)'] - I_p + \mathrm{E}[\hat{\theta}(x - \theta)'] + \mathrm{E}[(x - \theta)\hat{\theta}'].$$

The difficulty lies in the expression $\mathrm{E}[\hat{\theta}(x-\theta)']$, that is, the covariance between $\hat{\theta}$ and x. Since we have assumed that $\lambda(w_p)$ is absolutely continuous and that $\mathrm{E}\,|\lambda'(w_p)|$ is finite (Assumption A), we can invoke Lemma 5 from Appendix B which gives $\mathrm{E}[\hat{\theta}(x-\theta)'] = \mathrm{E}[\partial\hat{\theta}/\partial x']$. Now,

$$d\hat{\theta} = (d\lambda(w_p))\,x + \lambda(w_p)\,dx = \lambda'(w_p)(2x'dx)\,x + \lambda(w_p)\,dx$$

and we thus obtain

$$\frac{\partial\hat{\theta}}{\partial x'} = \lambda(w_p)I_p + 2\lambda'(w_p)\,xx'.$$

Then, letting

$$\phi(w_p) = (1-\lambda(w_p))^2 + 4\lambda'(w_p), \tag{9}$$

the MSE matrix takes the form

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}[(1-\lambda(w_p))^2 xx'] - I_p + 2\,\mathrm{E}[\lambda(w_p)]I_p + 4\,\mathrm{E}[\lambda'(w_p)xx'] \\ &= I_p - 2\,\mathrm{E}[1-\lambda(w_p)]I_p + \mathrm{E}[\phi(w_p)xx'],\end{aligned} \tag{10}$$

which, using again (7b) and (7c), we can write as

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= I_p - 2\,\mathrm{E}[1-\lambda(w_p)]I_p + \mathrm{E}[\phi(w_{p+2})]I_p + \mathrm{E}[\phi(w_{p+4})]\theta\theta' \\ &= \nu_1 J + \nu_2(I_p - J),\end{aligned}$$

where $J = \theta\theta'/\theta'\theta$ is the idempotent matrix of Proposition 1. In summary, we have proved

Proposition 2 *The eigenvalues ν_1 and ν_2 in Proposition 1 can equivalently be written as*

$$\nu_1 = 2\,\mathrm{E}[\lambda(w_p)] - 1 + \mathrm{E}[\phi(w_{p+2})] + (\theta'\theta)\,\mathrm{E}[\phi(w_{p+4})],$$

and

$$\nu_2 = 2\,\mathrm{E}[\lambda(w_p)] - 1 + \mathrm{E}[\phi(w_{p+2})].$$

Since the eigenvalues must be the same as in Proposition 1, we have proved as a by-product that

$$2\,\mathrm{E}[\lambda'(w_{p+2})] = \mathrm{E}[\lambda(w_{p+2})] - \mathrm{E}[\lambda(w_p)], \tag{11}$$

which generalizes Saleh (2006, Eq. 2.2.13e).

Notice that we now have three equivalent expressions for the trace of the MSE matrix, namely

$$\begin{aligned}\operatorname{tr}\operatorname{MSE}(\hat{\theta}) &= p\,\mathrm{E}[\lambda^2(w_{p+2})] + (\theta'\theta)\left(1 - 2\,\mathrm{E}[\lambda(w_{p+2})] + \mathrm{E}[\lambda^2(w_{p+4})]\right) \\ &= p - 2p\,\mathrm{E}[1 - \lambda(w_p)] + \mathrm{E}[w_p\phi(w_p)] \\ &= p - 2p\,\mathrm{E}[1 - \lambda(w_p)] + p\,\mathrm{E}[\phi(w_{p+2})] + (\theta'\theta)\,\mathrm{E}[\phi(w_{p+4})], \end{aligned} \tag{12}$$

where the first expression follows from (8), the second follows from (10), and the third follows by adding up the eigenvalues $\nu_1 + (p-1)\nu_2$ in Proposition 2 or alternatively from (7b).

5 The James–Stein class

The well-known James–Stein (JS) estimator (Stein, 1956; James and Stein, 1961) is defined in (5) through

$$\lambda_{JS}(w_p) = 1 - \frac{c}{w_p} \qquad (p \geq 3,\, c \geq 0).$$

The traditional JS estimator has $c = p-2$, but we leave c undetermined for now. The derivative of λ is $\lambda'_{JS}(w_p) = c/w_p^2$ and the function ϕ defined in (9) becomes

$$\phi_{JS}(w_p) = (1 - \lambda(w_p))^2 + 4\lambda'(w_p) = c(c+4)/w_p^2.$$

Letting $\mu_{k,p} = \mathrm{E}[(1/w_p)^k]$, we shall employ the following identities (see Saleh, 2006, Eq. 2.2.13):

$$\begin{aligned}(\theta'\theta)\mu_{1,p+2} &= 1 - (p-2)\mu_{1,p}, \\ (\theta'\theta)\mu_{2,p+4} &= \mu_{1,p+2} - (p-2)\mu_{2,p+2}, \\ 2\mu_{2,p+2} &= \mu_{1,p} - \mu_{1,p+2}. \end{aligned}$$

This gives

$$\nu_1 = 1 - \left(\frac{c[(c+4)(p-3)+4]}{2}\right)\mu_{1,p} + \left(\frac{c(c+4)(p-1)}{2}\right)\mu_{1,p+2}, \tag{13}$$

$$\nu_2 = 1 + \frac{c^2}{2}\mu_{1,p} - \frac{c(c+4)}{2}\mu_{1,p+2}, \tag{14}$$

so that

$$\operatorname{tr}\operatorname{MSE}(\hat{\theta}_{JS}) = \nu_1 + (p-1)\nu_2 = p - c(2p - c - 4)\mu_{1,p}. \tag{15}$$

The JS estimator thus weakly dominates the ML estimator if and only if $0 < c < 2(p-2)$ and $\operatorname{tr}\operatorname{MSE}(\hat{\theta}_{JS})$ reaches a minimum for $c = p-2$, which is therefore the obvious choice if one is only interested in weak dominance. For $c = 0$ the JS estimator equals the ML estimator, but for

$c = 2(p-2)$ the JS estimator does not equal the ML estimator and the two MSE matrices are not the same even though the two traces are the same (namely p) for every value of $\theta'\theta$.

The fact that the traditional JS estimator (with $c = p-2$) weakly dominates the ML estimator is the so-called Stein paradox, which caused and still causes much comment and disbelief. Thompson (1989, pp. 182–183) writes:

> "When estimating, simultaneously, the density of mosquitoes in Houston, the average equatorial temperature of Mars, and the gross national product of ancient Persia, we ought not believe that some mathematical quirk demands that we multiply our usual (separable) estimates by a finagle factor which artificially combines all three estimates. [...] When the use of a particular criterion function yields results that are completely contrary to our intuitions, we should question the criterion function before disregarding our intuitions."

This is precisely right: we should question the criterion function. If we choose the trace of the MSE matrix as our criterion (weak dominance) then we get the Stein paradox, but if we consider the whole MSE matrix then the paradox disappears, as shown in Figure 1.

According to Proposition 1 we have, at $\theta = 0$,

$$\nu_1 = \nu_2 = 1 - \frac{c(2p-c-4)}{p(p-2)},$$

so that both eigenvalues are smaller than 1 at $\theta = 0$ for any $0 < c < 2(p-2)$. Hence, for small values of $\theta'\theta$, the JS estimator dominates the ML estimator strongly but not for large values, and hence the JS estimator does not strongly dominate the ML estimator; see also the discussion in Saleh (2006, Eqs 4.3.31 and 4.3.32).

5.1 The positive JS estimator

The traditional JS estimator (where $c = p-2$) has the disadvantage that the associated λ function is negative when $w_p < p-2$. For small values of $\theta'\theta$ the probability of this happening is nontrivial. In particular, at $\theta'\theta = 0$ we find that $\Pr(w_p < p-2) = 0.30$ for $p = 5$, 0.37 for $p = 10$, and 0.41 for $p = 20$. At $\theta'\theta = 10$ these probabilities reduce to 0.01 ($p = 5$), 0.03 ($p = 10$), and 0.07 ($p = 20$); and at larger values of $\theta'\theta$ the probabilities become negligible.

The positive James–Stein (JS+) estimator (Baranchik, 1964) forces the shrinkage factor λ to lie between zero and one:

$$\lambda_{JS+}(w_p) = \begin{cases} 0 & \text{if } w_p \le p-2, \\ 1 - \frac{p-2}{w_p} & \text{if } w_p > p-2. \end{cases} \tag{16}$$

Note that the constant c in the definition of the JS estimator is no longer a constant but rather a concave function of w_p. The positive JS estimator is continuous but not differentiable at $w_p = p-2$; it is however absolutely continuous and hence satisfies Assumption A. The derivative of λ is

$$\lambda'_{JS+}(w_p) = \begin{cases} 0 & \text{if } w_p < p-2, \\ (p-2)/w_p^2 & \text{if } w_p > p-2. \end{cases}$$

The positive JS estimator dominates the traditional JS estimator not only weakly but even strongly; see, e.g., Saleh (2006, Section 4.3.3). The positive JS estimator, like the JS estimator, dominates the ML estimator, but only weakly in accordance with our conjecture. Since JS+ dominates JS, the JS estimator is not admissible. But the JS+ is also inadmissible because it is kinked. Hence there exists an estimator that weakly dominates it; see Lehmann and Casella (1998, Chapter 5, Example 7.3). Such an estimator was in fact found by Shao and Strawderman (1994), but the improvement over $\hat{\theta}_{JS+}$ is negligible.

5.2 Hansen's trimmed linear shrinkage estimator

The trimmed linear shrinkage (TLS) estimator proposed by Hansen (2015) was designed to have good performance in terms of minimax regret. Hansen (2015) first derives the efficiency bound (the lowest achievable trace of the MSE matrix) for the class of minimax orthogonally invariant estimators satisfying the conditions of Efron and Morris (1976, Theorem 3). Then he obtains the TLS estimator by numerically approximating the smallest possible maximum regret over the class of continuous linear splines. This is the TLS estimator with shrinkage function

$$\lambda_{TLS}(w_p) = \begin{cases} 0 & \text{if } w_p \leq \tau_1, \\ 1 - b - \frac{a}{w_p} & \text{if } \tau_1 < w_p \leq \tau_2, \\ 1 - \frac{2(p-2)}{w_p} & \text{if } w_p > \tau_2, \end{cases} \tag{17}$$

where a and b depend on p (see Hansen, 2015, Table 3) and satisfy

$$0 < b < 1, \qquad 0 < a < 2(p-2)(1-b),$$

with $\tau_1 = a/(1-b)$ and $\tau_2 = (2(p-2)-a)/b$. As with the positive JS estimator, the constant c in the definition of the TLS estimator is a concave function of w_p. The TLS estimator is continuous but not differentiable at $w_p = \tau_1$ and $w_p = \tau_2$. When $a = p-2$ and $b = 0$ we obtain the positive JS estimator as a special case.

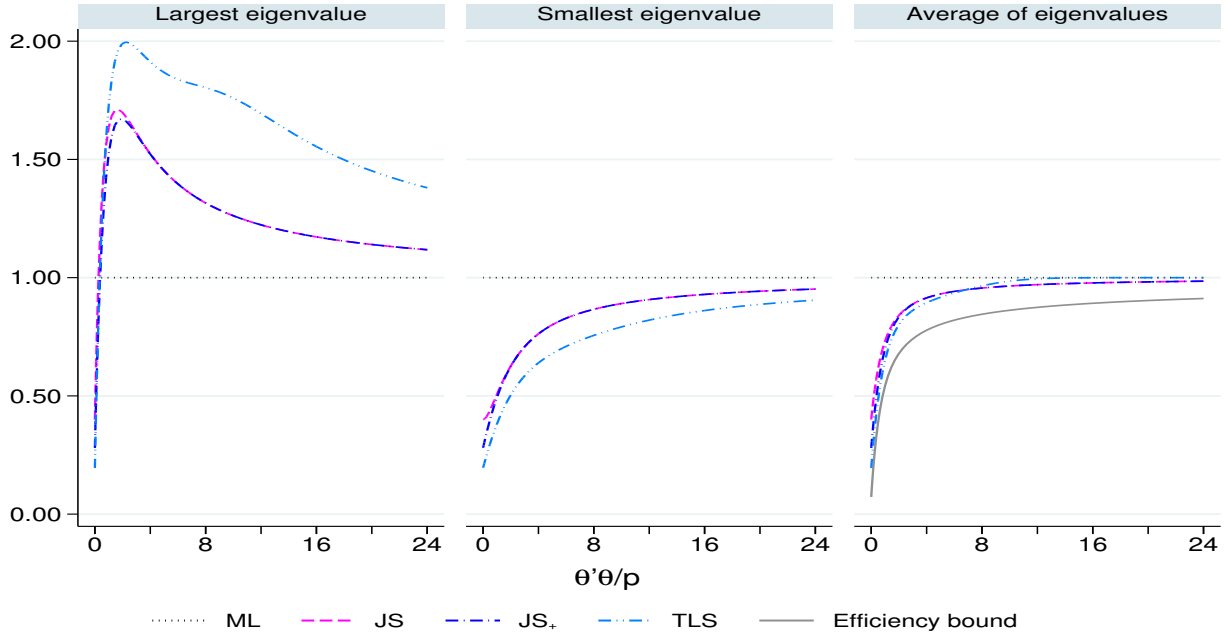

Figure 2: James–Stein class eigenvalues for $p = 5$

The derivative of λ is

$$\lambda'_{TLS}(w_p) = \begin{cases} 0 & \text{if } w_p < \tau_1, \\ a/w_p^2 & \text{if } \tau_1 < w_p < \tau_2, \\ 2(p-2)/w_p^2 & \text{if } w_p > \tau_2. \end{cases}$$

Even though the TLS estimator is inadmissible (because kinked), numerical comparisons in Hansen (2015, Table 2) show that this estimator leads to a substantial reduction of maximum regret relative to eleven other shrinkage estimators.

5.3 Comparison of JS class estimators

In Figure 2 we present the largest, smallest, and average eigenvalue of the MSE matrix for each of the three JS-class estimators: JS, JS+, and TLS, and as a benchmark also of the ML estimator. The results are given for $p = 5$; a comparison for different values of p will be discussed in Figures 4 and 5. The expectations involved in the two eigenvalues of the JS-class estimators are approximated numerically by Monte Carlo simulations based on forty million pseudo-random draws.

In the third panel we plot the average eigenvalue which, for each of the three JS-class estimators, is monotonically increasing in $\theta'\theta$ and converges to the minimax bound 1 (the unique eigenvalue of the ML estimator) as $\theta'\theta \to \infty$. In this panel we also plot Hansen's efficiency bound which, by

definition, lies below each of the three curves. It is clear that each of the three JS-class estimators weakly dominate the ML estimator. The JS+ estimator weakly (even strongly) dominates the JS estimator, but the TLS estimator does not weakly dominate either the JS or the JS+ estimator. The TLS estimator has lower risk for small values of $\theta'\theta$, but slightly larger risk at large values of $\theta'\theta$. In particular, at $\theta'\theta = 0$, the average risk is 0.40 (JS), 0.28 (JS+) and 0.20 (TLS), while the Hansen bound is 0.07. In terms of risk performance the three estimators are close except for very small values of $\theta'\theta$.

In the second panel we plot the smallest eigenvalue ν_2, which is close to average, because the smallest eigenvalue has multiplicity $p-1$. The TLS estimator performs particularly well.

Things are different for the largest eigenvalue ν_1, plotted in the first panel. The JS-class estimators strongly dominate the ML estimator for very small values of $\theta'\theta$, but clearly not uniformly. The good performance of the TLS estimator for the smallest and average eigenvalue comes at the expense of a poor performance for the largest eigenvalue. Since JS+ strongly dominates JS, we see that the largest eigenvalue associated with the JS+ estimator is uniformly smaller than the largest eigenvalue associated with the JS estimator, especially for small values of $\theta'\theta$. At $\theta'\theta = 0$ the largest, smallest, and average eigenvalue coincide, and hence the largest eigenvalue is 0.40 for the JS estimator, 0.28 for the JS+ estimator, and 0.20 for the TLS estimator. The shape of the largest eigenvalue is characterized by a maximum larger than 1 at some $\theta'\theta > 0$, and so none of these three minimax estimators strongly dominates the ML estimator. Hence our conjecture appears to be true for the JS class.

6 The Thompson class

As an alternative to the JS class we present what we call the Thompson class, defined in (6):

$$\lambda_{Th}(w_p) = \frac{w_p}{c + w_p}$$

for some $c \geq 0$. The estimator was introduced by Thompson (1968) and we shall see that it also weakly dominates the ML estimator.

If we allow c in (5) to depend on w_p, then the Thompson class can be obtained as a special case of the JS class by replacing c in (5) by the concave function $c(w_p) = cw_p/(c+w_p)$; see also Baranchik (1970, Example 2). As with the JS+ estimator (but unlike the JS estimator), the function λ_{Th} of the Thompson estimator satisfies $0 \leq \lambda_{Th} \leq 1$, and hence is a proper shrinkage factor. The shrinkage

function is continuously differentiable and its derivative is

$$\lambda'_{Th}(w_p) = \frac{c}{(c+w_p)^2},$$

so that the function ϕ defined in (9) takes the form

$$\phi_{Th}(w_p) = (1-\lambda_{Th}(w_p))^2 + 4\lambda'_{Th}(w_p) = \frac{c(c+4)}{(c+w_p)^2}.$$

Using Propositions 1 and 2 we have

$$\nu_1 = \nu_2 + \theta'\theta\, Q_p \tag{18}$$

and

$$\nu_2 = 1 - 2\,\mathrm{E}\left[\frac{c}{c+w_p}\right] + \mathrm{E}\left[\frac{c(c+4)}{(c+w_{p+2})^2}\right] = \mathrm{E}\left[\frac{w_{p+2}^2}{(c+w_{p+2})^2}\right], \tag{19}$$

where

$$Q_p = 1 - 2\,\mathrm{E}\left[\frac{w_{p+2}}{c+w_{p+2}}\right] + \mathrm{E}\left[\frac{w_{p+4}^2}{(c+w_{p+4})^2}\right] = \mathrm{E}\left[\frac{c(c+4)}{(c+w_{p+4})^2}\right]. \tag{20}$$

In particular, it follows from (12) that

$$\begin{aligned}\operatorname{tr}\mathrm{MSE}(\hat\theta_{Th}) &= p\nu_2 + \theta'\theta\, Q_p \\ &= p - (2p-c-4)\,\mathrm{E}[1-\lambda_{Th}(w_p)] - (c+4)\,\mathrm{E}[1-\lambda_{Th}(w_p)]^2,\end{aligned} \tag{21}$$

from which we see that a necessary and sufficient condition that the Thompson estimator weakly dominates the ML estimator is

$$0 < c \le 2(p-2) + \frac{2p\,\mathrm{E}[1-\lambda_{Th}(w_p)]^2}{\mathrm{E}[\lambda_{Th}(w_p)(1-\lambda_{Th}(w_p))]},$$

so that a sufficient condition is given by $0 < c \le 2(p-2)$. The Thompson estimator, like the JS estimator, is inadmissible (Strawderman and Cohen, 1971, p. 278), which means that there exists another estimator which weakly dominates the Thompson estimator, but almost certainly with negligible benefits.

In Figure 3 we present, again for $p = 5$, the largest, smallest, and average eigenvalue of the MSE matrix for four variants of the Thompson estimator (labeled Th_1–Th_4). In Th_1 we take $c = 2(p-2)$, which is the minimax regret solution obtained through Hansen's efficiency bound, while in Th_2 we take $c = p-2$, the central value in the interval $[0, 2(p-2)]$ for which we know that $\operatorname{tr}\mathrm{MSE}(\hat\theta_{Th}) \le p$. In Th_3 and Th_4 we let c depend on w_p. The above formulas in this section are then no longer valid

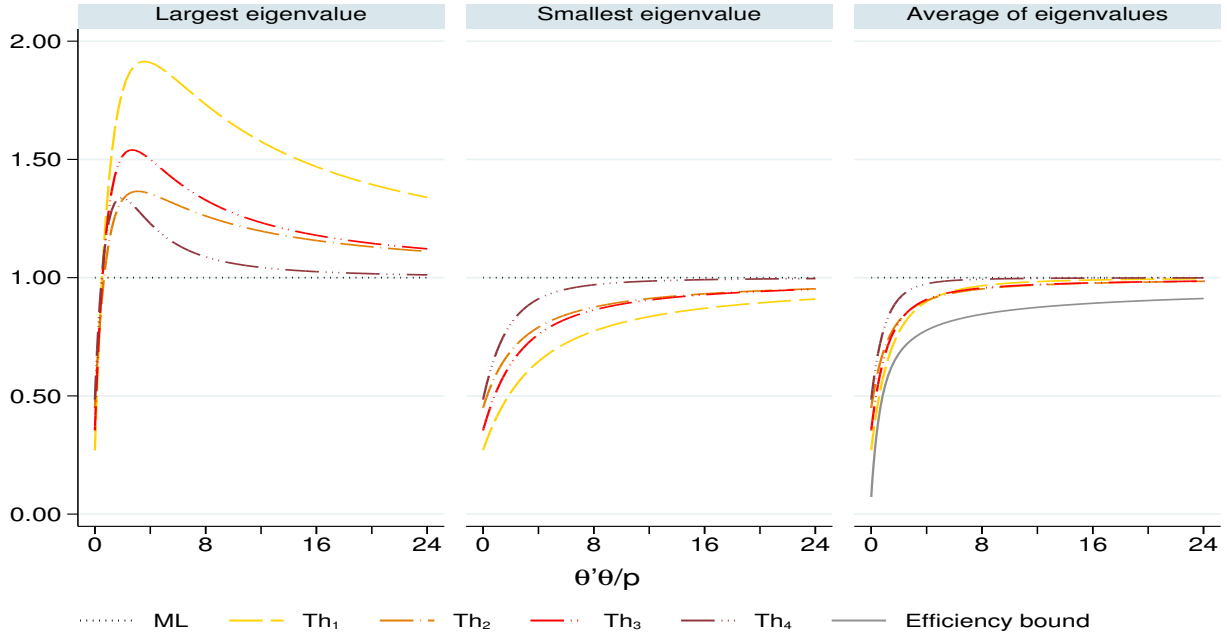

Figure 3: Thompson class eigenvalues for $p = 5$

because λ'_{Th} will have an additional term, but it will still be the case that λ_{Th} depends on x only through w_p and hence Propositions 1 and 2 still apply. More specifically, when $c = c(w_p)$ we have

$$\phi_{Th}(w_p) = \frac{c(c+4) - 4w_p c'(w_p)}{(c+w_p)^2}.$$

In Th_3 we let

$$c = \frac{(w_p + 2p)(p-2)}{w_p + p}, \qquad c' = \frac{-p(p-2)}{(w_p+p)^2},$$

motivated by the fact that $c = 2(p-2)$ performs well at $\theta'\theta = 0$, while $c = p-2$ performs well for large values of $\theta'\theta$. The c function is a compromise which decreases monotonically from $2(p-2)$ at $w_p = 0$ to $p-2$ when $w_p \to \infty$.

In Th_4 we let

$$c = \frac{2p(p-2)}{w_p + p}, \qquad c' = \frac{-2p(p-2)}{(w_p+p)^2},$$

motivated by the fact that the minimax regret solution for the average eigenvalue is $c = 2(p-2)$, while the minimax solution for the largest eigenvalue is $c = 0$ (this is also true in the class of JS estimators). The c function is monotonically decreasing from $c = 2(p-2)$ at $w_p = 0$ to $c = 0$ when $w_p \to \infty$.

At $\theta'\theta = 0$ the largest, smallest, and average eigenvalue are all equal, namely 0.27 in Th_1, 0.45 in Th_2, 0.35 in Th_3, and 0.48 in Th_4, while the efficiency bound is 0.07. The estimator Th_1 has the

smallest ν_2 (the smaller of the two eigenvalues) and therefore performs well in terms of risk (average eigenvalue), especially for small values of $\theta'\theta$. But there is a cost, namely that ν_1 (the larger of the two eigenvalues) is the largest of the four estimators. As $\theta'\theta$ increases, the average eigenvalue of Th_1 converges more rapidly to the minimax bound 1 than Th_2 and Th_3. In contrast, the estimator Th_4 has the largest ν_2 and therefore performs relatively poorly in terms of risk (average eigenvalue), but it has the smallest ν_1 of the four estimators and therefore safeguards against high risk for linear combinations of $\hat{\theta}_{Th}$ according to the inequality in (4). The estimators Th_2 and Th_3 are in-between. In particular, the eigenvalues of Th_3 are (mostly) in-between Th_1 and Th_2, by construction.

These results emphasize again the trade-off between the smallest and largest eigenvalue: the risk of shrinkage estimators in the $\mathcal{L}$-class can be made smaller but at the cost of increasing the largest eigenvalue. All four estimators weakly dominate the ML estimator (third panel), but none of them dominates the ML estimator strongly (first panel), in accordance with our conjecture.

7 Comparison of James–Stein and Thompson class estimators

One may wonder which of the two classes performs better: the celebrated James–Stein class or the much less well-known Thompson class. In the previous two sections we reported results only for $p = 5$. Let us now compare the two classes for $p = 5$, 10, and 20. In Figures 4 and 5 we compare, respectively, the average and the largest eigenvalue of the ML, JS+, TLS, Th_3, and Th_4 estimators.

In Figure 4 we see that the average eigenvalue gets closer to the efficiency bound as p increases, so that the improvement relative to the ML estimator becomes larger. Also, as p increases, Th_4 is not performing well, JS+ is preferred over TLS, and the difference between JS+ and Th_3 becomes negligible. At small values of $\theta'\theta$ the differences between the estimators are more pronounced. For example, for $p = 20$, the eigenvalues at $\theta'\theta = 0$ are equal to 0.05 for the TLS estimator, 0.06 for the JS+ estimator, 0.20 for the Th_3 estimator, and 0.31 for the Th_4 estimator. Thus, we conclude that the JS+ estimator is to be preferred when the average eigenvalue is our criterion, because for small values of $\theta'\theta$ it outperforms Th_3 and for large values of p it outperforms TLS.

While the behavior of the average eigenvalue is important, the behavior of the largest eigenvalue is important too, especially if we wish to protect ourselves against high risk for linear combinations of the p components of our estimator. This is because we know from (4) that the risk of a linear combination $w'\hat{\theta}$ in the $\mathcal{L}$-class can be as large as $(\theta'\theta)\nu_1$.

If we choose the largest eigenvalue as our criterion, then things are rather different as shown

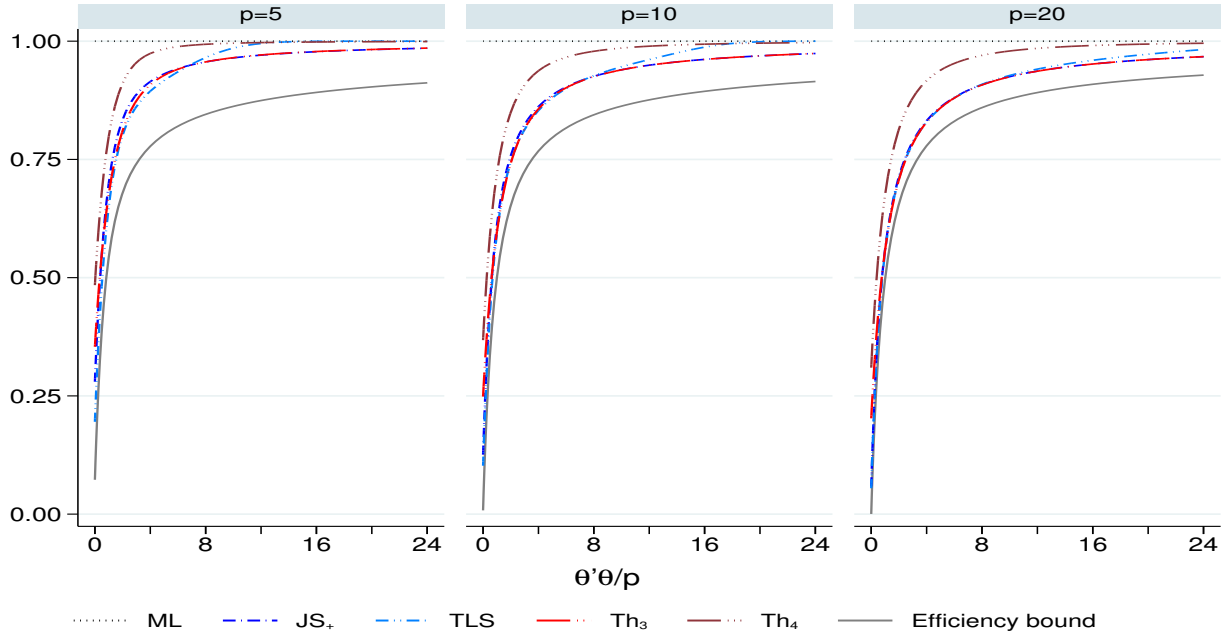

Figure 4: James–Stein versus Thompson class estimators: average of eigenvalues

in Figure 5. While the average eigenvalue is rather stable in p (except at $\theta'\theta = 0$), the largest eigenvalue is an increasing function of p (and, not shown, the smallest eigenvalues is a decreasing function of p). This implies that, as p increases, the risk for linear combinations of $\hat{\theta}$ increases, so that the difference with the (constant) risk of the ML estimator can become large. From the point of view of the largest eigenvalue the Thompson-class estimators perform better than the JS-class estimators, Th_4 being the best choice and TLS the worst.

A good compromise is the Thompson estimator Th_3 which performs well in both criteria. It weakly dominates the ML estimator without increasing the largest eigenvalue too much.

8 The Thompson estimator: joint versus separate

So far we have studied the $\mathcal{L}$-class of shrinkage estimators and in particular the James–Stein and the Thompson classes. For these estimators, Propositions 1 and 2 apply so that the MSE matrix has only two distinct eigenvalues $\nu_1 \geq \nu_2 \geq 0$. Many estimators in the $\mathcal{L}$-class weakly dominate the ML estimator, for example the JS estimator (Stein paradox). The fact that combining things that have nothing to do with each other can provide an advantage remains difficult to understand, see the quote by Thompson (1989) in Section 5. We have tried to explain the paradox by emphasizing that the trace criterion is only one possible criterion, and that we could equally well (perhaps even

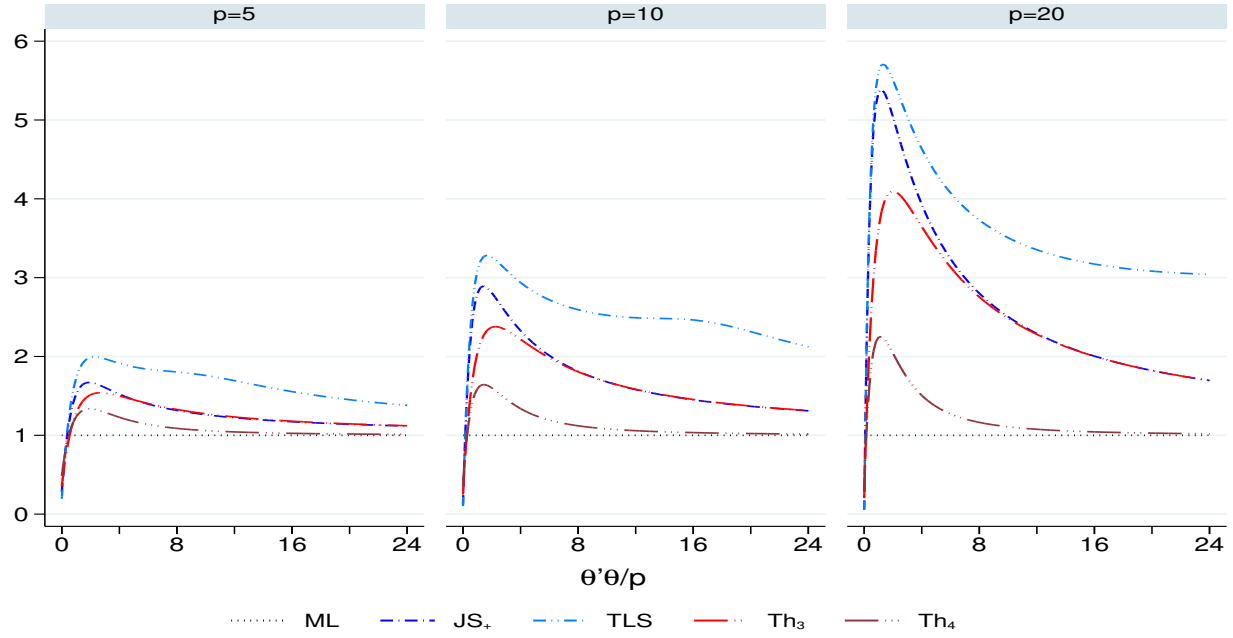

Figure 5: James–Stein versus Thompson class estimators: largest eigenvalue

better) choose the maximum eigenvalue criterion. Then there is no paradox.

In this section we confront joint estimation with separate (component-wise) estimation and we do this in the context of the Thompson estimator because the Thompson estimator is defined for all p (also $p = 1$) in contrast to the JS estimator.

The observations are still given by $x \sim \mathrm{N}(\theta, I_p)$. But now we estimate each θ_i separately by

$$\hat{\theta}_i^* = \lambda^*(x_i)x_i \qquad (i = 1, \ldots, p), \tag{22}$$

where λ^* is the common shrinkage function. Notice that $\hat{\theta}_i^*$ is different from the ith component $\hat{\theta}_i = \lambda(w_p)x_i$ of the joint estimator, because the shrinkage function λ^* depends only on x_i while λ depends on $w_p = x'x$. Since the x_i are independent, the $\hat{\theta}_i^*$ are also independent while the $\hat{\theta}_i$ are not.

In our case, the shrinkage factor for the (separate) Thompson estimator is

$$\lambda^*(x_i) = \frac{x_i^2}{c + x_i^2} \tag{23}$$

for some $c \geq 0$. The MSE matrix is now

$$\mathrm{MSE}(\hat{\theta}^*) = \Sigma + bb',$$

where

$$\Sigma = \text{diag}(\text{var}(\hat{\theta}_1^*), \ldots, \text{var}(\hat{\theta}_p^*)), \qquad b = \text{E}[\hat{\theta}^* - \theta]$$

represent the variance and the bias of $\hat{\theta}^*$. The MSE matrix is not diagonal because $\hat{\theta}^*$ is biased and hence, for $i \neq j$,

$$\text{E}[(\hat{\theta}_i^* - \theta_i)(\hat{\theta}_j^* - \theta_j)] = b_i b_j \neq 0.$$

in general. In contrast to the MSE matrix for the joint estimator (where there are only two distinct eigenvalues), the eigenvalues of the MSE matrix for the separate estimator are, in general, all distinct.

The Hansen efficiency bound does not apply to the separate estimator, but each $\hat{\theta}_i^*$ satisfies the univariate efficiency bound $\theta_i^2/(1+\theta_i^2)$ derived by Magnus (2002, Theorem A.7). Using the univariate efficiency bound we find that minimax regret solution for the generic component of the separate Thompson estimator $\hat{\theta}_i^* = x_i^3/(c+x_i^2)$ is $c = 3.7213$ with maximum regret equal to 0.1815.[1]

While the properties of the MSE matrix for the joint estimator $\hat{\theta}$ depend on θ only through p and the noncentrality parameter $\theta'\theta$, this is no longer the case for the separate estimator $\hat{\theta}^*$. To analyze the properties of $\text{MSE}(\hat{\theta}^*)$ we need to make assumptions on the components of θ. We shall assume that either all θs are the same, $\theta = \delta \imath_p$ for some scalar δ, or that there are two sets of θs: two large values and $p-2$ small values. Thus, we assume that

$$\theta_1 = \theta_2 = \delta, \qquad \theta_3 = \theta_4 = \cdots = \theta_p = (1-\alpha)\delta, \tag{24}$$

where the parameter $0 \leq \alpha \leq 1$ controls the degree of sparsity. When $\alpha = 0$ there is no sparsity (all θs are the same) and when $\alpha = 1$ there is maximum sparsity. Our treatment of the θs is similar to the simulation setup of Hansen (2016) who compares the risk bounds of the JS and the least absolute shrinkage and selection operator (LASSO). Here, instead of performing numerical MSE comparisons, we provide explicit expressions for the eigenvalues of $\text{MSE}(\hat{\theta}^*)$ which are thus directly comparable with those of $\text{MSE}(\hat{\theta})$ in Propositions 1 and 2.

When all θs are the same ($\alpha = 0$), then the x_i are not only independent but also identically distributed. The MSE matrix then takes the form

$$\text{MSE}(\hat{\theta}^*) = \nu_1^* J + \nu_2^*(I_p - J), \qquad J = \imath_p \imath_p'/p, \tag{25}$$

where

$$\nu_1^* = \nu_2^* + p\,(\text{E}[\hat{\theta}_1^*] - \delta)^2, \qquad \nu_2^* = \text{var}(\hat{\theta}_1^*). \tag{26}$$

[1] The maximum regret 0.1815 is much smaller than the maximum regret 0.4251 for the same estimator reported in Magnus (2002, p. 230). Apparently a computational or typographical error.

When there are two sets of θs ($\alpha > 0$), then the MSE matrix is given by (31) in Appendix A,

$$\mathrm{MSE}(\hat{\theta}^*) = \begin{pmatrix} \nu_1 J_2 + \nu_2(I_2 - J_2) & \gamma \imath_2 \imath_{p-2}' \\ \gamma \imath_{p-2} \imath_2' & \xi_1 J_{p-2} + \xi_2(I_{p-2} - J_{p-2}) \end{pmatrix}, \tag{27}$$

where $\imath_2 = (1,1)'$ and $\imath_{p-2} = (1,1,\ldots,1)'$ have dimensions 2 and $p-2$ respectively, and $J_2 = \imath_2 \imath_2'/2$ and $J_{p-2} = \imath_{p-2}\imath_{p-2}'/(p-2)$, and

$$\nu_1 = \sigma_1^2 + 2b_1^2, \qquad \nu_2 = \sigma_1^2, \qquad \xi_1 = \sigma_p^2 + (p-2)b_p^2, \qquad \xi_2 = \sigma_p^2, \qquad \gamma = b_1 b_p,$$

with

$$b_1 = \mathrm{E}[\hat{\theta}_1^*] - \delta, \qquad \sigma_1^2 = \mathrm{var}(\hat{\theta}_1^*), \qquad b_p = \mathrm{E}[\hat{\theta}_p^*] - (1-\alpha)\delta, \qquad \sigma_p^2 = \mathrm{var}(\hat{\theta}_p^*).$$

The MSE matrix depends on four parameters and its behavior is completely determined by its four eigenvalues ξ_2 (multiplicity $p-3$), and ν_2, ν_1^*, and ξ_1^* (each with multiplicity 1), where

$$\nu_1^* = \frac{\nu_1 + \xi_1}{2} + \frac{1}{2}\sqrt{(\nu_1 - \xi_1)^2 + 8(p-2)\gamma^2}, \tag{28}$$

$$\xi_1^* = \frac{\nu_1 + \xi_1}{2} - \frac{1}{2}\sqrt{(\nu_1 - \xi_1)^2 + 8(p-2)\gamma^2}. \tag{29}$$

The largest eigenvalue is ν_1^* because $\nu_1^* \geq \xi_1^*$ and $\nu_1^* \geq \max(\nu_1, \xi_1) \geq \max(\nu_2, \xi_2)$, and the trace is given by

$$\mathrm{tr}\,\mathrm{MSE}(\hat{\theta}^*) = (p-3)\xi_2 + \nu_2 + \nu_1^* + \xi_1^* = 2(\sigma_1^2 + b_1^2) + (p-2)(\sigma_p^2 + b_p^2). \tag{30}$$

In Figure 6 we present the average eigenvalue of the joint Thompson estimator Th_3 and four versions of the separate Thompson estimator $\hat{\theta}^*$, depending on the sparsity α. We consider three values of c: one value of $c < 1$, one value $c = 1$, and one value $c > 1$. For the largest value we select the minimax regret solution $c = 3.72$. The selected degrees of sparsity are: $\alpha = 0$ (all θs the same, no sparsity), $\alpha = 0.75$, $\alpha = 0.90$, and $\alpha = 1$ (maximum sparsity). We set $p = 10$. Plots for alternative values of p are qualitatively similar.

For $\alpha = 0$ and $\alpha = 0.75$, the joint estimator outperforms the separate estimator. However, when there is a stronger degree of sparsity ($\alpha = 0.90$ and $\alpha = 1$), the separate estimator can perform better than the joint estimator. Thus, we conclude that the joint Thompson estimator is to be preferred when the coefficients are roughly comparable in magnitude, while the separate Thompson estimator is to be preferred when few coefficients are large in magnitude and the others are relatively small. These conclusions agree with earlier comparisons between JS and LASSO (Hansen, 2016) and between joint and component-wise JS estimators (Lehmann and Casella, 1998, p. 365),

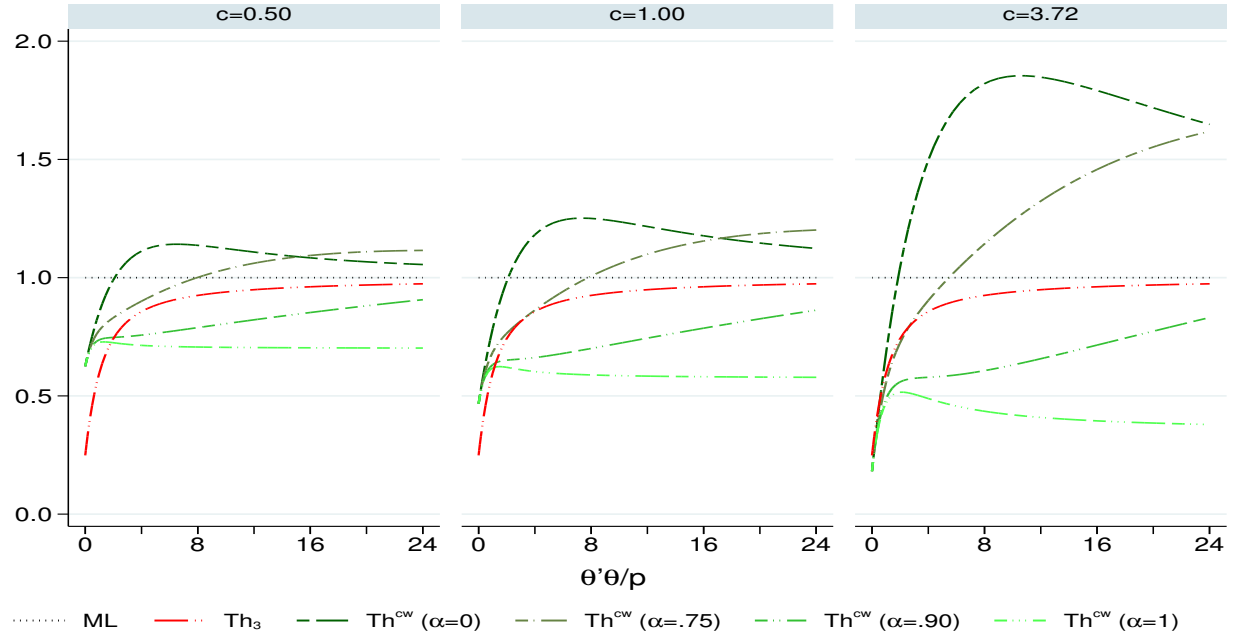

Figure 6: Joint versus separate Thompson estimator for $p = 10$: average of eigenvalues

When there is a strong degree of sparsity (α close to one), the separate Thompson estimator based on the minimax regret solution $c = 3.72$ has uniformly lower risk than at smaller values of c. In general, however, this 'optimal' choice of c ensures low risk only around $\theta'\theta = 0$ where regret is at its maximum. Minimax regret is a local optimality criterion which may not perform well in regions of the parameter space that are far from the optimal solution. For small values of α and large values of $\theta'\theta$, the separate Thompson estimator with $c = 3.72$ performs poorly.

Notice that the curvature at $\alpha = 1$ is different than the curvature at $\alpha < 1$. When $\alpha = 1$ we have $b_p = 0$ and $\xi_1 = \xi_2$, and the curvature of the average eigenvalue depends solely on $\text{MSE}(\hat{\theta}_1^*)$. But when $\alpha < 1$ the curvature of the average eigenvalue depends on both $\text{MSE}(\hat{\theta}_1^*)$ and $\text{MSE}(\hat{\theta}_p^*)$.

In Figure 7 we present the corresponding figure for the largest eigenvalue. Here is becomes clear that the 'optimal' value (in terms of minimax regret) of $c = 3.72$ may not be a good choice. To safeguard against large values of the largest eigenvalue we need to restrict c to $c \leq 1$, in which case the separate estimators perform well compared to the joint estimator. Choosing $c \leq 1$ thus limits the maximum risk of the separate estimator by sacrificing a small amount of efficiency in terms of the trace criterion, in the spirit of the limited translation empirical Bayes estimators proposed by Efron and Morris (1972).

Summarizing, the joint estimator performs well when the θs are close to each other in magnitude,

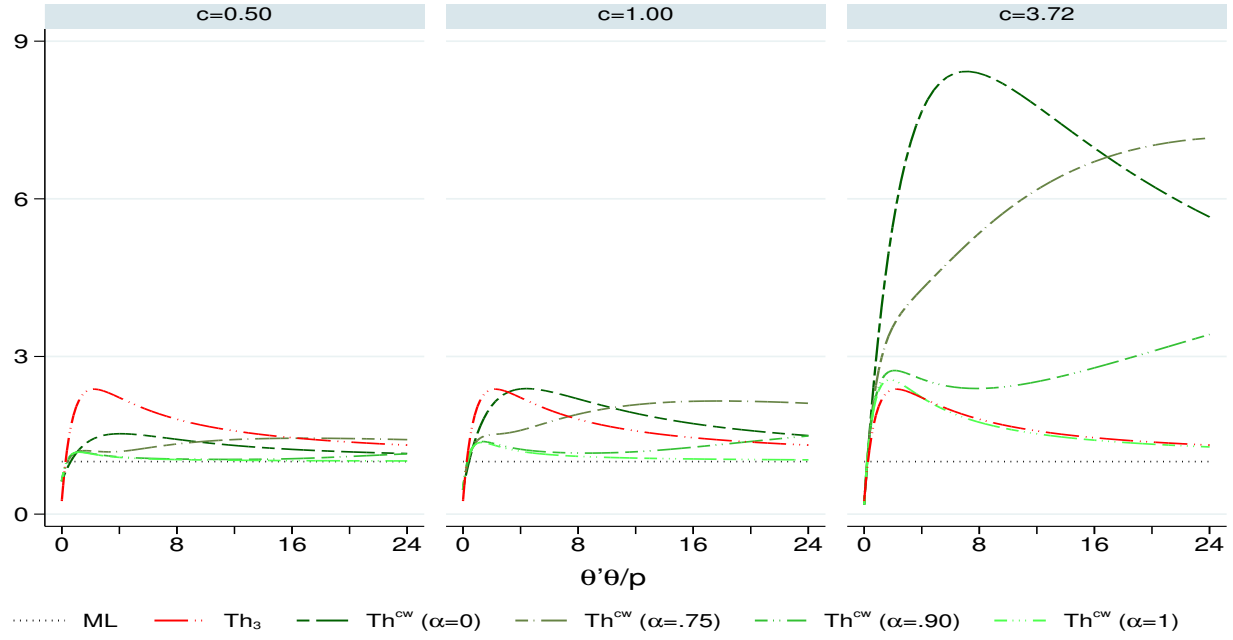

Figure 7: Joint versus separate Thompson estimator for $p = 10$: largest eigenvalue

while the separate estimator is better when there is a large degree of sparsity. In addition, the 'optimal' value for c in terms of minimax regret may not work well over the whole parameter space, because maximum regret occurs at $\theta'\theta = 0$ so that the optimal c value works for small values of $\theta'\theta$ but not necessarily for large values.

9 Conclusions

We have considered the estimation of the p-variate normal mean by means of the class of orthogonally invariant shrinkage estimators, which includes as special cases the James-Stein and the Thompson classes of estimators. We have shown that the mean squared error matrix of these estimators has only two distinct eigenvalues: the larger with multiplicity 1 and the smaller with multiplicity $p-1$. This result has important implications because the two eigenvalues fully characterize the mean squared error matrix.

We have applied our findings to study the well-known concept of weak dominance (trace criterion = average eigenvalue criterion) and the less well-known concept of strong dominance (largest eigenvalue criterion) in the James-Stein and the Thompson classes of estimators. Our conjecture is that, while there are several shrinkage estimators which weakly dominate the ML estimator (Stein paradox), none of them dominates the ML estimator strongly. The intuition is that there exists

a trade-off between the smaller and the larger of the two eigenvalues, as the risk of our shrinkage estimators can only be made smaller at the expense of increasing the larger eigenvalue. The validity of the Stein paradox thus depends crucially on the choice of the criterion function used to rank estimators.

Although the trace criterion is important, we have emphasized that the largest eigenvalue criterion is important too, particularly when we are concerned with the risk of linear combinations of the coefficients or (as a special case) with the maximum risk in estimating single components of the unknown mean vector. It turns out that in these situations we can often reduce the maximum risk by sacrificing a little efficiency in terms of average risk. For example, we find that the James-Stein class of shrinkage estimators is slightly preferred in terms of average risk, but that the Thompson class of shrinkage estimators weakly dominates the ML estimator with only a slight increase in maximum risk.

To gain additional insight about the Stein paradox we have also compared the (joint) Thompson estimator with its separate (component-wise) counterpart in the case when the unknown mean vector consists of two sets of coefficients. Our findings suggest that the joint estimation approach is to be preferred when the coefficients are roughly comparable in magnitude, while the separate estimation approach is to be preferred when there is a strong degree of sparsity with few large coefficients and many small or zero coefficients. Again, there is a clear trade-off between the four possible eigenvalues of the mean squared error matrix of the separate estimator, so that it may be desirable to sacrifice a little efficiency in terms of average risk to limit maximum risk on the single components of the shrinkage estimator.

Our conjecture is valid for a large class of shrinkage estimators, but we are not sure how large this class precisely is. We claim that it is the complete $\mathcal{L}$-class, and we invite the reader to prove it.

Appendices

A Some results involving idempotent matrices

Our first result is not new.

Lemma 1 *Let A be a symmetric idempotent $p \times p$ matrix of rank $r(A) = r$. Then, $r(I_p - A) = p - r$ and $A(I_p - A) = 0$. Next, let*

$$V = \nu_1 A + \nu_2 (I_p - A).$$

Then the eigenvalues of V are ν_1 (multiplicity r) and ν_2 (multiplicity $p-r$), its determinant is $|V| = \nu_1^r \nu_2^{p-r}$, and its inverse is $V^{-1} = (1/\nu_1)A + (1/\nu_2)(I_p - A)$ when $\nu_1 \neq 0$ and $\nu_2 \neq 0$.

Proof This is a simple version of a much more general result, see Abadir and Magnus (2005, Exercises 8.72 and 8.73).

Now let's consider an extension where the $p \times p$ matrix V is partitioned into blocks of dimensions p_1 and p_2 as follows:

$$V = \begin{pmatrix} \nu_1 J_1 + \nu_2(I_{p_1} - J_1) & \gamma \imath_1 \imath_2' \\ \gamma \imath_2 \imath_1' & \xi_1 J_2 + \xi_2(I_{p_2} - J_2) \end{pmatrix}, \tag{31}$$

where $\imath_1 = (1, 1, \ldots, 1)'$ and $\imath_2 = (1, 1, \ldots, 1)'$ have dimensions p_1 and p_2 respectively, and $J_1 = \imath_1 \imath_1'/p_1$ and $J_2 = \imath_2 \imath_2'/p_2$.

In the special case $p_1 = p_2 = p$ we can write $V = V_1 \otimes J + V_2 \otimes (I_p - J)$, where

$$V_1 = \begin{pmatrix} \nu_1 & \gamma p \\ \gamma p & \xi_1 \end{pmatrix}, \qquad V_2 = \begin{pmatrix} \nu_2 & 0 \\ 0 & \xi_2 \end{pmatrix}, \qquad J = (1/p)\, \imath \imath',$$

which implies that the eigenvalues of V are given by ν_2 ($p-1$ times), ξ_2 ($p-1$ times), and the two eigenvalues of V_1 (Magnus, 1982, Lemma 2.1).

When $p_1 \neq p_2$ we cannot write V in terms of Kronecker matrices, but the result is still essentially the same.

Lemma 2 *The eigenvalues of V are ν_2 ($p_1 - 1$ times), ξ_2 ($p_2 - 1$ times) and two additional eigenvalues ν_1^* and ξ_1^* given by*

$$\frac{\nu_1 + \xi_1}{2} \pm \frac{1}{2}\sqrt{(\nu_1 - \xi_1)^2 + 4\gamma^2 p_1 p_2}.$$

The sum of the eigenvalues is

$$\operatorname{tr} V = \nu_1 + (p_1 - 1)\nu_2 + \xi_1 + (p_2 - 1)\xi_2,$$

and V is positive semidefinite if and only if ν_1, ν_2, ξ_1, and ξ_2 are all ≥ 0 and in addition $\nu_1 \xi_1 \geq \gamma^2 p_1 p_2$.

Proof We write

$$V - \lambda I_p = \begin{pmatrix} \bar{\nu}_1 J_1 + \bar{\nu}_2(I_{p_1} - J_1) & \gamma \imath_1 \imath_2' \\ \gamma \imath_2 \imath_1' & \bar{\xi}_1 J_2 + \bar{\xi}_2(I_{p_2} - J_2) \end{pmatrix},$$

where

$$\bar{\nu}_1 = \nu_1 - \lambda, \qquad \bar{\nu}_2 = \nu_2 - \lambda, \qquad \bar{\xi}_1 = \xi_1 - \lambda, \qquad \bar{\xi}_2 = \xi_2 - \lambda.$$

The determinant is

$$\begin{aligned}
|V - \lambda I_p| &= |\bar{\nu}_1 J_1 + \bar{\nu}_2(I_{p_1} - J_1)| \\
&\quad \times |\bar{\xi}_1 J_2 + \bar{\xi}_2(I_{p_2} - J_2) - \gamma^2 \imath_2 \imath_1' [\bar{\nu}_1 J_1 + \bar{\nu}_2(I_{p_1} - J_1)]^{-1} \imath_1 \imath_2'| \\
&= \bar{\nu}_1 \bar{\nu}_2^{p_1 - 1} \left| \bar{\xi}_1 J_2 + \bar{\xi}_2(I_{p_2} - J_2) - \gamma^2 \imath_2 \imath_1' \left(\frac{1}{\bar{\nu}_1} J_1 + \frac{1}{\bar{\nu}_2}(I_{p_1} - J_1) \right) \imath_1 \imath_2' \right| \\
&= \bar{\nu}_1 \bar{\nu}_2^{p_1 - 1} |\bar{\xi}_1 J_2 + \bar{\xi}_2(I_{p_2} - J_2) - (\gamma^2/\bar{\nu}_1) p_1 p_2 J_2| \\
&= \bar{\nu}_1 \bar{\nu}_2^{p_1 - 1} \left| \frac{\bar{\nu}_1 \bar{\xi}_1 - \gamma^2 p_1 p_2}{\bar{\nu}_1} J_2 + \bar{\xi}_2(I_{p_2} - J_2) \right| \\
&= \bar{\nu}_2^{p_1 - 1} \bar{\xi}_2^{p_2 - 1} \left(\bar{\nu}_1 \bar{\xi}_1 - \gamma^2 p_1 p_2 \right) \\
&= (\nu_2 - \lambda)^{p_1 - 1} (\xi_2 - \lambda)^{p_2 - 1} \left((\nu_1 - \lambda)(\xi_1 - \lambda) - \gamma^2 p_1 p_2 \right).
\end{aligned}$$

Hence the eigenvalues of V are ν_2 ($p_1 - 1$ times), ξ_2 ($p_2 - 1$ times), and the two solutions ν_1^* and ξ_1^* of the quadratic equation $(\nu_1 - \lambda)(\xi_1 - \lambda) - \gamma^2 p_1 p_2 = 0$. Note that $\nu_1^* + \xi_1^* = \nu_1 + \xi_1$. In the special case $\gamma = 0$ we have $\nu_1^* = \nu_1$ and $\xi_1^* = \xi_1$.

The sum of the eigenvalues is

$$(p_1 - 1)\nu_2 + (p_2 - 1)\xi_2 + \nu_1 + \xi_1,$$

and it easy to verify that this equals the trace of V, as of course it should.

B Stein's lemma

Stein's lemma (Stein, 1981) is a rather surprising and strong result. We first consider the univariate, then the multivariate case. The generalization is not straightforward.

Lemma 3 *Let $x \sim \mathrm{N}(\theta, 1)$ and let $h : \Re \to \Re$ be an absolutely continuous function with derivative h'. Assume that $\mathrm{E}\,|h'(x)| < \infty$. Then,*

$$\mathrm{cov}(h(x), x) = \mathrm{E}[h(x)(x - \theta)] = \mathrm{E}[h'(x)].$$

Proof We write

$$\begin{aligned}
[h(x)\phi(x - \theta)]' &= h'(x)\phi(x - \theta) + h(x)\phi'(x - \theta) \\
&= h'(x)\phi(x - \theta) - h(x)(x - \theta)\phi(x - \theta),
\end{aligned}$$

where ϕ denotes the standardnormal density. Integrating gives

$$0 = h(x)\phi(x - \theta)\Big|_{-\infty}^{\infty} = \mathrm{E}[h'(x)] - \mathrm{E}[h(x)(x - \theta)].$$

Note the requirement of absolute continuity, which imposes a smoothness property on h that is stronger than (uniform) continuity, but weaker than continuous differentiability. It guarantees that h is differentiable almost everywhere. In the applications it is important to place minimal restrictions on the function h, for example that it may be kinked.

The univariate version of Stein's lemma is a powerful result with many nontrivial applications. As a simple example, let $h(x) = x^m$. Then we immediately obtain all moments of the normal distribution through the recursion $\mathrm{E}[x^{m+1}] = \theta\,\mathrm{E}[x^m] + m\,\mathrm{E}[x^{m-1}]$.

In the multivariate case we have $x \sim \mathrm{N}(\theta, I_p)$ with $p \geq 2$ and we need the concept of 'almost differentiability' (in Stein's terminology). We write $x = (x_i, x_{-i})$ to decompose a point $x \in \Re^p$ in terms of its ith component x_i and all other components x_{-i}. Thus, $h(\cdot, x_{-i})$ refers to h as a function of its ith argument with all other arguments fixed at x_{-i}. Then h is 'almost differentiable' if for each $i = 1, \ldots, p$ and almost every $x_{-i} \in \Re^{p-1}$ the function $h(\cdot, x_{-i}) : \Re \to \Re$ is absolutely continuous. An almost differentiable function h has partial derivatives almost everywhere.

Given this multivariate extension of the concept of absolute continuity, Stein's lemma reads as follows.

Lemma 4 (Stein) *Let $x \sim \mathrm{N}(\theta, I_p)$ with $p \geq 2$ and let $h : \Re^p \to \Re$ be almost differentiable with $\mathrm{E}\,\|\nabla h(x)\| < \infty$, where $\nabla h(x)$ denotes the gradient of $h(x)$. Then,*

$$\mathrm{E}\left[h(x)(x-\theta)\right] = \mathrm{E}[\nabla h(x)].$$

Proof See Stein (1981, Lemma 2).

Stein's result can be generalized straightforwardly to the case where h is a vector function.

Lemma 5 *Let $x \sim \mathrm{N}(\theta, I_p)$ with $p \geq 2$ and let $h : \Re^p \to \Re^q$. If $h_j : \Re^p \to \Re$ is almost differentiable with $\mathrm{E}\,\|\nabla h_j(x)\| < \infty$ for all $j = 1, \ldots, q$, then*

$$\mathrm{E}\left[h(x)(x-\theta)'\right] = \mathrm{E}\left[\frac{\partial h(x)}{\partial x'}\right].$$

Proof This follows from Lemma 4 by considering each row of $h(x)(x-\theta)'$ separately. Then,

$$\mathrm{E}\left[h_j(x)(x-\theta)'\right] = \mathrm{E}\left[\frac{\partial h_j(x)}{\partial x'}\right]$$

for each j and the result follows.

References

Abadir, K. M. and J. R. Magnus (2005). *Matrix Algebra*. Cambridge University Press.

Baranchik, A. J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. *Technical report*, No. 51, Department of Statistics, Stanford University.

Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Annals of Mathematical Statistics*, 41, 642–645.

Bock, M. E. (1975). Minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, 3, 209–218.

Candès, E. J., C.A. Sing-Long, and J.D. Trzasko (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61(19), 4643–4657.

Casella, G. (1990). Estimators with nondecreasing risk: Application of a chi-squared identity. *Statistics & Probability Letters*, 10, 107–109.

Efron, B. E. and C. Morris (1972). Limiting the risk of Bayes and empirical Bayes estimators–part II: the empirical Bayes case. *Journal of the American Statistical Association*, 67, 130–139.

Efron, B. E. and C. Morris (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *Annals of Statistics* 4, 11–21.

James, W. and C. M. Stein (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 361–379. University of California Press.

Hansen, B. E. (2015). Shrinkage efficiency bounds. *Econometric Theory*, 31, 860–879.

Hansen, B. E. (2016). The risk of James–Stein and Lasso shrinkage. *Econometric Reviews*, 35, 1456–1470.

Johnstone, I. M. (2019). *Gaussian Estimation: Sequence and Wavelet Models*. Draft version, available from statweb.stanford.edu.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation*, Second Edition. Springer-Verlag.

Magnus, J. R. (1982). Multivariate error components analysis of linear and nonlinear regression models by maximum likelihood. *Journal of Econometrics*, 19, 239–285.

Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal*, 5, 225–236.

Magnus, J.R. and G. De Luca (2016). Weighted-average least squares: A review. *Journal of Economic Surveys*, 30, 117–148.

Mikkelsen, F. R. and N. R. Hansen (2018). Degrees of freedom for piecewise Lipschitz estimators. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 54, 819–841.

Saleh, A. K. M. E. (2006). *Theory of Preliminary Test and Stein-Type Estimation With Applications*. Wiley.

Shao, P. Y.-S. and W. E. Strawderman (1994). Improving on the James-Stein positive-part estimator. *Annals of Statistics*, 22, 1517–1538.

Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 197–206. University of California Press.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9, 1135–1151.

Strawderman, W. E. and A. Cohen (1971). Admissibility of estimators of the mean vector of a multivariate normal distribution with quadratic loss. *Annals of Mathematical Statistics*, 42, 270–96.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, 63, 113–122.

Thompson, J. R. (1989). *Empirical Model Building*. Wiley.

Tibshirani, R. J. (2015). Degrees of freedom and model search. *Statistica Sinica*, 25, 1265–1296.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.