Submitted to the Annals of Statistics

IMPROVED MULTIVARIATE NORMAL MEAN ESTIMATION WITH UNKNOWN COVARIANCE WHEN p IS GREATER THAN n

By Didier Chételat and Martin T. Wells *

Cornell University

We consider the problem of estimating the mean vector of a p-variate normal (θ, Σ) distribution under invariant quadratic loss, $(\delta - \theta)'\Sigma^{-1}(\delta - \theta)$, when the covariance is unknown. We propose a new class of estimators that dominate the usual estimator $\delta^0(X) = X$. The proposed estimators of θ depend upon X and an independent Wishart matrix S with n degrees of freedom, however S is singular almost surely when $p > n$. The proof of domination involves the development of some new unbiased estimators of risk for the $p > n$ setting. We also find some relationships between the amount of domination and the magnitudes of n and p.

1. Introduction. Suppose a p-dimensional random vector X is observed which is normally distributed, with mean vector θ and unknown positive definite covariance matrix Σ, and we wish to estimate θ under the invariant quadratic loss

$$L(\theta, \delta) = (\delta - \theta)'\Sigma^{-1}(\delta - \theta). \tag{1.1}$$

Since the covariance matrix Σ is unknown, a random matrix S is observed along with X, which is assumed to be independent of X, and has a Wishart distribution with n degrees of freedom, where $p > n$. In high-dimensional estimation problems, where p, the number of features, is nearly as large as, or larger than, n, the number of observations, the ordinary least squares estimator does not typically provide a satisfactory estimate of θ.

Modern data sets are increasingly becoming characterized by a number of features that is much larger than the number of sample units (large-p, small-n) in contrast to classical data sets where the number of sample units is often much larger than the number of random variables (small-p, large-n). Modern applications in the $p > n$ setting include examples from microarrays, association mapping, proteomics, radiology, biomedical imaging, signal

*This research was partially supported by NSF grant DMS-1208488.

AMS 2000 subject classifications: Primary 62F10; secondary 62C20, 62H12

Keywords and phrases: covariance estimation, James-Stein estimation, invariant quadratic loss, large-p-small-n problems, location parameter, minimax estimation, Moore-Penrose inverse, risk function, singular Wishart distribution

processing, climate modeling, and finance. For instance, in the case of microarray data, the dimensionality is frequently in thousands or beyond, while the sample size is typically in the order of tens. The large-p, small-n scenario poses challenges in most inferential settings. We are considering a canonical setting. For the usual multivariate location-scale estimation problem let $W = (W_1, ..., W_p)$ denote an $N \times p$ matrix of data (N is the number of observations and p the number of features) where W_i are taken from a p-dimensional normal distribution with mean vector θ and covariance matrix Ξ. In this article we let the X and S be the sample mean and covariance of the features, respectively. In the context of this notation, $\Sigma = N^{-1}\Xi$ and $n = N - 1$.

The usual estimator under invariant quadratic loss is $\delta_0(X) = X$. It is minimax and admissible when $p \leq 2$ and $p \leq n$. However, when $p \geq 3$ and $p \leq n$, $\delta_0(X)$ remains minimax but is no longer admissible. Explicit improvements are known in the multivariate normal case [James and Stein (1961); Berger and Bock (1976); Berger et al. (1977); Gleser (1979); Berger and Haff (1983); Gleser (1986)] and in the case of elliptically symmetric distribution [Srivastava and Bilodeau (1989); Fourdrinier, Strawderman and Wells (2003)].

In this article we primarily concentrate on the case $p > n$ and construct a class of estimators, depending on the sufficient statistics (X, S), of the form

$$\delta(X, S) = X + g(X, S) \tag{1.2}$$

which dominate $\delta_0(X)$ under invariant quadratic loss. Note that, although the loss in (1.1) is invariant, the estimate in (1.2) may not be (except for $\delta_0(X)$). This class generalizes several estimators studied previously for the multivariate normal distribution to the $p \leq n$ setting [James and Stein (1961); Berger and Bock (1976); Berger et al. (1977); Gleser (1979); Berger and Haff (1983); Gleser (1986)]. Examples of estimators we study here in this setting extend the class of so-called Baranchik estimators and includes a new high dimensional James-Stein estimator

$$\delta_a^{JS}(X, S) = \left(I - \frac{aSS^+}{X'S^+X}\right) X$$

where $0 \leq a \leq \frac{2(n-2)}{p-n+3}$ and S^+ is the Moore-Penrose inverse of S.

The estimation of the inverse covariance matrix, namely the precision matrix Σ^{-1}, of a multivariate normal distribution has been an important problem in practical situations as well as from a theoretical perspective. But, when $p > n$, the Wishart-distributed sample covariance matrix is singular;

in this case, one is tempted to construct estimators using the Moore-Penrose generalized inverse S^+. Recently there has been an increased interest in the problem of estimating the covariance matrix of large dimension given variables of dimension larger than the number of observations [Bickel and Levina (2008); Aspremont, Banerjee and Ghaoui (2008); Konno (2009); Ledoit and Wolf (2004); Levina, Rothman and Zhu (2008); Rothman et al. (2008)].

Our method of proof relies on an unbiased estimator of risk difference, say $\rho(X, S)$. Specifically, we show that, for $g(X, S)$ of the form $-\frac{r(X'S^+X)SS^+}{X'S^+X}X$, the estimator $\delta(X, S) = X + g(X, S)$ dominates X provided $\rho(X, S) \leq 0$. In the next section we present the main results and their proofs are given in Section 3. We need Stein's integration-by-parts identity [Stein (1981)] and the so-called Stein-Haff identity for the singular Wishart distribution. The Stein-Haff identity was derived by Haff (1979) and Stein (1977) for the full rank Wishart distribution. A similar identity for the elliptically contoured model has been given by Fourdrinier, Strawderman and Wells (2003). We make some concluding comments in Section 4.

For a matrix M, let M' denote its transpose, M^+ its Moore-Penrose pseudo-inverse and $\frac{\partial M}{\partial t}$ its componentwise derivative matrix, i.e. the matrix such that $\left(\frac{\partial M}{\partial t}\right)_{ij} = \frac{\partial M_{ij}}{\partial t}$. Moreover, let δ_{ij} denote the Kronecker delta.

2. Main Results. Let X be a random vector distributed as $N_p(\theta, \Sigma)$ with unknown θ and Σ. Suppose an estimator of Σ is available, say $S \sim \text{Wishart}_p(n, \Sigma)$, with S independent of X. By definition of the Wishart distribution, we can write $S = Y'Y$ for some matrix normal $Y \sim N_{n\times p}(0, I \otimes \Sigma)$. An elementary property of this distribution is that S is (almost surely) invertible if $p \leq n$, and (almost surely) singular if $p > n$ [cf. Srivastava and Khatri (1979)].

An usual estimator of θ is $\delta^0(X, S) = X$; however, it turns out that this estimator is inadmissible under quadratic loss. If some estimator $S \sim \text{Wishart}_p(n, \Sigma)$ is available, with $n \geq p \geq 3$, δ^0 is dominated by the so-called James-Stein estimator

$$\delta^{JS}(X, S) = \left(1 - \frac{(p-2)/(n-p+3)}{X'S^{-1}X}\right) X.$$

The main contribution of this article is to extend this type of result to a more general class of estimators in the $p > n$ setting.

For some positive, bounded, and differentiable function $r : \mathbb{R} \to \mathbb{R}$, define

the Baranchik-type estimator

$$\delta_r(X,S) = \left(I - \frac{r(X'S^+X)SS^+}{X'S^+X}\right)X = X + g(X,S) \tag{2.1}$$

where I is the identity matrix and S^+ denotes the Moore-Penrose inverse of S. This estimator generalizes the usual Baranchik (1970) estimator to the unknown covariance setting for $p > n$.

THEOREM 1. *Let* $\min(p,n) \geq 3$. *Suppose that:*

(i) r *satisfies* $0 \leq r \leq \frac{2\,(\min(n,p)-2)}{n+p-2\,\min(n,p)+3}$;
(ii) r *is nondecreasing; and*
(iii) r' *is bounded.*

Then under invariant quadratic loss, δ_r *dominates* δ^0.

Throughout the article we will use the expression $tr(SS^+)$, which of course equals $\min(n,p)$. This notation allows us to simultaneously handle both the $p > n$ and $n \geq p$ cases. The condition $\min(p,n) \geq 3$ merely guarantees that condition (i) of Theorem 1 holds for some r and is reminiscent of the dimension cut-off in classical Stein estimation.

PROOF. The hypotheses of the theorem imply that r is differentiable almost everywhere. Under invariant quadratic loss, the difference in risk between δ_r and δ^0 is given by

$$\begin{aligned}\Delta_\theta =& E_\theta\left[(X+g(X,S)-\theta)'\Sigma^{-1}(X+g(X,S)-\theta)\right] \\ &- E_\theta\left[(X-\theta)'\Sigma^{-1}(X-\theta)\right] \\ =& 2E_\theta\left[g(X,S)'\Sigma^{-1}(X-\theta)\right] + E_\theta\left[g(X,S)'\Sigma^{-1}g(X,S)\right].\end{aligned} \tag{2.2}$$

In order to show the domination result we need to show that under the sufficient conditions on r, (2.2) is nonpositive for all θ. First, for the left most term of (2.2) it can be shown that

$$2E_\theta\left[g(X,S)'\Sigma^{-1}(X-\theta)\right] = 2E_\theta\left[\mathrm{div}_X g(X,S)\right].$$

Fourdrinier, Strawderman and Wells (2003) give a more general form of this result in their Lemma 1(i); it is essentially an extension of Stein's classical

integration by parts identity. By using Lemma 2 in Section 3, we have that

$$
\begin{aligned}
2E_\theta\left[\mathrm{div}_X g(X,S)\right] &= -2E_\theta\left[\mathrm{div}_X \frac{r(X'S^+X)SS^+X}{X'S^+X}\right] \\
&= -2E_\theta\left[2r'(X'S^+X) + r(X'S^+X)\frac{\mathrm{tr}(SS^+)-2}{X'S^+X}\right]. \qquad (2.3)
\end{aligned}
$$

For the right term of (2.2), we find, through Lemma 3 in Section 3:

$$
\begin{aligned}
&E_\theta\left[g(X,S)'\Sigma^{-1}g(X,S)\right] \\
&= E_\theta\left[\mathrm{tr}\left(\Sigma^{-1}Sr^2(X'S^+X)\frac{S^+XX'S^+S}{(X'S^+X)^2}\right)\right] \\
&= E_\theta\left[n\ \mathrm{tr}\left(r^2(X'S^+X)\frac{S^+XX'S^+S}{(X'S^+X)^2}\right)\right. \\
&\quad \left. + \mathrm{tr}\left(Y'\nabla_Y\left\{r^2(X'S^+X)\frac{SS^+XX'S^+}{(X'S^+X)^2}\right\}\right)\right].
\end{aligned}
$$

The finiteness of the risk of δ_r is guaranteed to hold by Theorem 2 in Section 3 for all p and n.

Now applying Lemma 1 in Section 3, we find

$$
\begin{aligned}
&E_\theta\left[n\ \mathrm{tr}\left(r^2(X'S^+X)\frac{S^+XX'S^+S}{(X'S^+X)^2}\right)\right. \\
&\quad \left. +\mathrm{tr}\left(Y'\nabla_Y\left\{r^2(X'S^+X)\frac{SS^+XX'S^+}{(X'S^+X)^2}\right\}\right)\right] \\
&= E_\theta\left[n\frac{r^2(X'S^+X)}{X'S^+X} - 4r(X'S^+X)r'(X'S^+X)\right. \\
&\quad \left. +r^2(X'S^+X)\frac{p-2\,\mathrm{tr}(SS^+)+3}{X'S^+X}\right] \\
&= E_\theta\left[r^2(X'S^+X)\frac{n+p-2\,\mathrm{tr}(SS^+)+3}{X'S^+X} - 4r(X'S^+X)r'(X'S^+X)\right]. \qquad (2.4)
\end{aligned}
$$

Replacing (2.3) and (2.4) back into (2.2), we obtain

$$
\begin{aligned}
\Delta_\theta = E_\theta\Bigg[&r^2(X'S^+X)\frac{n+p-2\,\mathrm{tr}(SS^+)+3}{X'S^+X} \\
&- 2r(X'S^+X)\frac{\mathrm{tr}(SS^+)-2}{X'S^+X} \\
&- 4r'(X'S^+X)\left\{1+r(X'S^+X)\right\}\Bigg].
\end{aligned}
$$

Since r is nonnegative and nondecreasing it follows that $-4r'(X'S^+X)\{1 + r(X'S^+X)\} \leq 0$. Finally, for the X and S such that $r(X'S^+X) \neq 0$,

$$\begin{aligned} & r^2(X'S^+X)\frac{n+p-2\operatorname{tr}(SS^+)+3}{X'S^+X} - 2r(X'S^+X)\frac{\operatorname{tr}(SS^+)-2}{X'S^+X} \leq 0 \\ \Leftrightarrow\ & r(X'S^+X) \leq \frac{2\left(\operatorname{tr}(SS^+)-2\right)}{n+p-2\operatorname{tr}(SS^+)+3} = \frac{2\left(\min(n,p)-2\right)}{n+p-2\min(n,p)+3}. \end{aligned}$$

Therefore, under the three sufficient conditions on r, it follows that $\Delta_\theta \leq 0$ for any θ, that is, the domination result holds. □

In the $p > n$ setting, we obtain the following two corollaries.

Corollary 1. *For $p > n \geq 3$, δ_r dominates δ^0 under invariant quadratic loss for all r nondecreasing, differentiable and satisfying*

$$0 \leq r \leq \frac{2(n-2)}{p-n+3}. \tag{2.5}$$

Corollary 2 (James-Stein estimator with large p and small n). *For $p > n \geq 3$ and $a \in \mathbb{R}$, the James-Stein-like estimator*

$$\delta_a^{JS}(X,S) = \left(I - \frac{aSS^+}{X'S^+X}\right)X \tag{2.6}$$

dominates δ^0 under invariant quadratic loss for all

$$0 \leq a \leq \frac{2(n-2)}{p-n+3}.$$

Note that if p is only moderately larger than n, Corollary 1 implies that one can construct an estimator with substantial improvement over δ^0. However, in the ultra-high dimensional setting the denominator in (2.5) could be quite large and consequently the amount of improvement over δ^0 could be quite small. The estimator in (2.6) generalizes the classical James-Stein with unknown covariance matrix,

$$\delta_a^{JS}(X,S) = \left(1 - \frac{a}{X'S^{-1}X}\right)X$$

which is, of course, restricted to the case $p \leq n$, for $a \in \mathbb{R}_+$. In this setting, this result is consistent with previously bounds in Fourdrinier, Strawderman and Wells (2003) (where $n-1$ is used instead of our n.)

3. Technical Results and Proofs. It remains to clarify several of the somewhat technical computations used in the proof of Theorem 1. We provide them in this section; these computations are likely to be of independent interest and showcase several technical maneuvers that the reader could find useful in dealing with singular Wishart matrices.

PROPOSITION 1. *Let Y be an $n \times p$ matrix, $S = Y'Y$, X a p vector and $F = X'S^+X$. It then follows that*

(i) $$\left\{\frac{\partial S}{\partial Y_{\alpha\beta}}\right\}_{kl} = \delta_{\beta k} Y_{\alpha l} + \delta_{\beta l} Y_{\alpha k};$$

(ii) $$\frac{\partial F}{\partial Y_{\alpha\beta}} = -2(X'S^+Y')_\alpha (S^+X)_\beta + 2(X'S^+S^+Y')_\alpha ((I - SS^+)X)_\beta;$$

(iii) $$\frac{\partial \{S^+XX'SS^+\}_{kl}}{\partial Y_{\alpha\beta}} =$$

$$\begin{aligned}
&(S^+S^+Y')_{k\alpha}((I - SS^+)XX'SS^+)_{\beta l} \\
&- S^+_{k\beta}(YS^+XX'SS^+)_{\alpha l} - (S^+Y')_{k\alpha}(S^+XX'SS^+)_{\beta l} \\
&+ (I - SS^+)_{k\beta}(YS^+S^+XX'SS^+)_{\alpha l} \\
&+ (S^+XX')_{k\beta}(YS^+)_{\alpha l} + (S^+XX'Y')_{k\alpha}(S^+)_{\beta l} \\
&+ (S^+XX'S^+Y')_{k\alpha}(I - SS^+)_{\beta l} \\
&- (S^+XX'SS^+)_{k\beta}(YS^+)_{\alpha l} - (S^+XX'SS^+Y')_{k\alpha}(S^+)_{\beta l}.
\end{aligned}$$

PROOF. First, notice that from the usual chain-rule that

$$\left\{\frac{\partial S}{\partial Y_{\alpha\beta}}\right\}_{kl} = \frac{\partial}{\partial Y_{\alpha\beta}} S_{kl} = \frac{\partial}{\partial Y_{\alpha\beta}} \sum_q Y_{qk} Y_{ql} = \delta_{\beta k} Y_{\alpha l} + \delta_{\beta l} Y_{\alpha k}.$$

This shows (i).

Let A be a symmetric matrix and $t \in \mathbb{R}$, then

$$\frac{\partial A^+}{\partial t} = -A^+ \frac{\partial A}{\partial t} A^+ + (I - AA^+)\frac{\partial A}{\partial t} A^+A^+ + A^+A^+ \frac{\partial A}{\partial t}(I - AA^+).$$

This result was, it seems, first proved in Golub and Pereyra (1973), as their Theorem 4.3, but can be found in standard textbooks on elementary linear algebra. Also, again for A symmetric, we have $AA^+ = A^+A$ and $A(I - AA^+) = (I - AA^+)A = A^+(I - AA^+) = (I - AA^+)A^+ = 0$. This easily follows from elementary properties of the Moore-Penrose pseudoinverse.

Since $S = Y'Y$, notice through a singular value decomposition argument that $SS^+Y' = Y'$ and thus $(I - SS^+)Y' = 0$. Using (i) we find that

$$\begin{aligned}
\frac{\partial F}{\partial Y_{\alpha\beta}} =& X' \frac{\partial S^+}{\partial Y_{\alpha\beta}} X \\
=& -\sum_{k,l} (X'S^+)_k \left\{\delta_{\beta k} Y_{\alpha l} + \delta_{\beta l} Y_{\alpha k}\right\} (S^+X)_l \\
&+ \sum_{k,l} (X'S^+S^+)_k \left\{\delta_{\beta k} Y_{\alpha l} + \delta_{\beta l} Y_{\alpha k}\right\} ((I - SS^+)X)_l \\
&+ \sum_{k,l} (X'(I - SS^+))_k \left\{\delta_{\beta k} Y_{\alpha l} + \delta_{\beta l} Y_{\alpha k}\right\} (S^+S^+X)_l \\
=& -\sum_{l} (X'S^+)_\beta Y_{\alpha l} (S^+X)_l - \sum_{k} (X'S^+)_k Y_{\alpha k} (S^+X)_\beta \\
&+ \sum_{l} (X'S^+S^+)_\beta Y_{\alpha l} ((I - SS^+)X)_l \\
&+ \sum_{k} (X'S^+S^+)_k Y_{\alpha k} ((I - SS^+)X)_\beta \\
&+ \sum_{l} (X'(I - SS^+))_\beta Y_{\alpha l} (S^+S^+X)_l \\
&+ \sum_{k} (X'(I - SS^+))_k Y_{\alpha k} (S^+S^+X)_\beta \\
=& -2(X'S^+Y')_\alpha (S^+X)_\beta + 2(X'S^+S^+Y')_\alpha ((I - SS^+)X)_\beta
\end{aligned}$$

which gives (ii).

Using (i) we have that for any conformable matrices A and B

$$\begin{aligned}
\left(A \frac{\partial S}{\partial Y_{\alpha\beta}} B\right)_{kl} =& \sum_{i,j} A_{ki} \left\{\frac{\partial S}{\partial Y_{\alpha\beta}}\right\}_{ij} B_{jl} \\
=& \sum_{i,j} A_{ki} \left\{\delta_{\beta i} Y_{\alpha j} + \delta_{\beta j} Y_{\alpha i}\right\} B_{jl} \\
=& \sum_{j} A_{k\beta} Y_{\alpha j} B_{jl} + \sum_{i} A_{ki} Y_{\alpha i} B_{\beta l} \\
=& A_{k\beta} (YB)_{\alpha l} + (AY')_{k\alpha} B_{\beta l}.
\end{aligned}$$

Therefore, using again $(I - SS^+)Y' = 0$:

$$\frac{\partial \{S^+XX'SS^+\}_{kl}}{\partial Y_{\alpha\beta}} = \Bigg\{ S^+S^+ \frac{\partial S}{\partial Y_{\alpha\beta}} (I - SS^+) XX'SS^+$$

$$
\begin{aligned}
&- S^+\frac{\partial S}{\partial Y_{\alpha\beta}}S^+XX'SS^+ + (I-SS^+)\frac{\partial S}{\partial Y_{\alpha\beta}}S^+S^+XX'SS^+ \\
&+ S^+XX'\frac{\partial S}{\partial Y_{\alpha\beta}}S^+ + S^+XX'SS^+S^+\frac{\partial S}{\partial Y_{\alpha\beta}}(I-SS^+) \\
&-S^+XX'SS^+\frac{\partial S}{\partial Y_{\alpha\beta}}S^+ + S^+XX'S(I-SS^+)\frac{\partial S}{\partial Y_{\alpha\beta}}S^+S^+\Bigg\}_{kl} \\
= &(S^+S^+Y')_{k\alpha}((I-SS^+)XX'SS^+)_{\beta l} \\
&- S^+_{k\beta}(YS^+XX'SS^+)_{\alpha l} - (S^+Y')_{k\alpha}(S^+XX'SS^+)_{\beta l} \\
&+ (I-SS^+)_{k\beta}(YS^+S^+XX'SS^+)_{\alpha l} \\
&+ (S^+XX')_{k\beta}(YS^+)_{\alpha l} + (S^+XX'Y')_{k\alpha}(S^+)_{\beta l} \\
&+ (S^+XX'S^+Y')_{k\alpha}(I-SS^+)_{\beta l} \\
&- (S^+XX'SS^+)_{k\beta}(YS^+)_{\alpha l} - (S^+XX'SS^+Y')_{k\alpha}(S^+)_{\beta l}
\end{aligned}
$$

which gives (iii). □

LEMMA 1. *Under the hypotheses of Theorem 1 we have*

$$
\begin{aligned}
&tr\left(Y'\nabla_Y\left\{r^2(X'S^+X)\frac{SS^+XX'S^+}{(X'S^+X)^2}\right\}\right) \\
&\quad = -4r(X'S^+X)r'(X'S^+X) + r^2(X'S^+X)\frac{p-2tr(SS^+)+3}{X'S^+X}
\end{aligned}
$$

where ∇_Y is interpreted as the matrix with components $(\nabla_Y)_{ij} = \frac{\partial}{\partial Y_{ij}}$.

PROOF. To simplify computations in what will follows, we let $F \equiv X'S^+X$. We then have

$$
\begin{aligned}
&\left[Y'\nabla_Y\left\{r^2(F)\frac{SS^+XX'S^+}{F^2}\right\}\right]_{ij} \\
&= \sum_{\alpha,\beta}(Y')_{i\alpha}\frac{\partial}{\partial Y_{\alpha\beta}}\left\{r^2(F)\frac{(SS^+XX'S^+)_{\beta j}}{F^2}\right\} \\
&= 2\sum_{\alpha,\beta}(Y')_{i\alpha}r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\cdot\frac{(SS^+XX'S^+)_{\beta j}}{F^2} && (3.4) \\
&\quad + \sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{\frac{\partial}{\partial Y_{\alpha\beta}}\{(SS^+XX'S^+)_{\beta j}\}}{F^2} && (3.5) \\
&\quad + \sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{-2\frac{\partial F}{\partial Y_{\alpha\beta}}(SS^+XX'S^+)_{\beta j}}{F^3}. && (3.6)
\end{aligned}
$$

To simplify (3.4) and (3.6) we apply Proposition 1 (ii) to get

$$\begin{aligned}
&\sum_{\alpha,\beta}(Y')_{i\alpha}\left\{\frac{\partial F}{\partial Y_{\alpha\beta}}\right\}(SS^+XX'S^+)_{\beta j}\\
=&-2\sum_{\alpha,\beta}(Y')_{i\alpha}(X'S^+Y')_\alpha(S^+X)_\beta(SS^+XX'S^+)_{\beta j}\\
&+2\sum_{\alpha,\beta}(X'S^+S^+Y')_\alpha(Y)_{\alpha i}(S^+XX'SS^+)_{j\beta}((I-SS^+)X)_\beta\\
=&-2X'S^+X(SS^+XX'S^+)_{ij}.
\end{aligned}$$

Using this, we get for (3.4)

$$\begin{aligned}
&2\sum_{\alpha,\beta}(Y')_{i\alpha}r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\cdot\frac{(SS^+XX'S^+)_{\beta j}}{F^2}\\
&\qquad=-4r(F)r'(F)\frac{(SS^+XX'S^+)_{ij}}{F}
\end{aligned}\tag{3.7}$$

and (3.6) becomes

$$\begin{aligned}
&\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{-2\frac{\partial F}{\partial Y_{\alpha\beta}}\cdot(SS^+XX'S^+)_{\beta j}}{F^3}\\
&\qquad=4r^2(F)\frac{(SS^+XX'S^+)_{ij}}{F^2}.
\end{aligned}\tag{3.8}$$

This leaves the term (3.5) to analyze. Using Proposition 1 (iii):

$$\begin{aligned}
&\sum_{\alpha,\beta}(Y')_{i\alpha}\frac{\partial}{\partial Y_{\alpha\beta}}\left\{(SS^+XX'S^+)_{\beta j}\right\}\\
&=\sum_{\alpha,\beta}(Y')_{i\alpha}\frac{\partial\{S^+XX'SS^+\}_{j\beta}}{\partial Y_{\alpha\beta}}\\
&=\sum_{\alpha,\beta}\Big\{(S^+S^+Y')_{j\alpha}Y_{\alpha i}((I-SS^+)XX'SS^+)_{\beta\beta}\\
&\quad-S^+_{j\beta}(Y')_{i\alpha}(YS^+XX'SS^+)_{\alpha\beta}\\
&\quad-(S^+Y')_{j\alpha}Y_{\alpha i}(S^+XX'SS^+)_{\beta\beta}\\
&\quad+(I-SS^+)_{j\beta}(Y')_{i\alpha}(YS^+S^+XX'SS^+)_{\alpha\beta}\\
&\quad+(S^+XX')_{j\beta}(Y')_{i\alpha}(YS^+)_{\alpha\beta}\\
&\quad+(S^+XX'Y')_{j\alpha}Y_{\alpha i}(S^+)_{\beta\beta}
\end{aligned}$$

$$
\begin{aligned}
&+ (S^+XX'S^+Y')_{j\alpha}Y_{\alpha i}(I - SS^+)_{\beta\beta} \\
&- (S^+XX'SS^+)_{j\beta}(Y')_{i\alpha}(YS^+)_{\alpha\beta} \\
&- (S^+XX'SS^+Y')_{j\alpha}Y_{\alpha i}(S^+)_{\beta\beta}\Big\} \\
=& (S^+XX'SS^+(I - SS^+))_{ij} \\
&- (SS^+XX'S^+)_{ij} \\
&- \operatorname{tr}(S^+XX'SS^+)(SS^+)_{ij} \\
&+ \operatorname{tr}((I - SS^+)XX'SS^+)(S^+)_{ij} \\
&+ (SS^+XX'S^+)_{ij} \\
&+ \operatorname{tr}(S^+)(SXX'S^+)_{ij} \\
&+ \operatorname{tr}(I - SS^+)(SS^+XX'S^+)_{ij} \\
&- (SS^+XX'S^+)_{ij} \\
&- \operatorname{tr}(S^+)(SXX'S^+)_{ij} \\
=& (p - \operatorname{tr}(SS^+) - 1)\left\{SS^+XX'S^+\right\}_{ij} - (X'S^+X)\left\{SS^+\right\}_{ij}.
\end{aligned}
$$

Next applying this computation in (3.5), we obtain

$$
\begin{aligned}
\sum_{\alpha,\beta}(Y')_{i\alpha}r^2(F)\frac{\frac{\partial}{\partial Y_{\alpha\beta}}\{(SS^+XX'S^+)_{\beta j}\}}{F^2} \\
= (p - \operatorname{tr}(SS^+) - 1)r^2(F)\frac{(SS^+XX'S^+)_{ij}}{F^2} \\
- r^2(F)\frac{(SS^+)_{ij}}{F}.
\end{aligned} \tag{3.9}
$$

Now we can combine (3.7), (3.9) and (3.8) together to complete the proof. That is, we have

$$
\begin{aligned}
&\operatorname{tr}\left(Y'\nabla_Y\left\{r^2(F)\frac{SS^+XX'S^+}{F^2}\right\}\right) \\
=& \sum_i\Big\{ - 4r(F)r'(F)\frac{(SS^+XX'S^+)_{ii}}{F} \\
&+ 4r^2(F)\frac{(SS^+XX'S^+)_{ii}}{F^2} \\
&+ (p - \operatorname{tr}(SS^+) - 1)r^2(F)\frac{(SS^+XX'S^+)_{ii}}{F^2} \\
&- r^2(F)\frac{(SS^+)_{ii}}{F}\Big\} \\
=& - 4r(F)r'(F) + r^2(F)\frac{p - 2\operatorname{tr}(SS^+) + 3}{F}
\end{aligned}
$$

as desired. □

LEMMA 2. *Under the hypotheses of Theorem 1 we have*

$$div_X \frac{r(X'S^+X)SS^+X}{X'S^+X} = 2r'(X'S^+X) + r(X'S^+X)\frac{tr(SS^+)-2}{X'S^+X}.$$

PROOF. Again, to simplify computations, let us denote $X'S^+X$ by F. We find

$$\begin{aligned}
\mathrm{div}_X \left\{ r(F)\frac{SS^+X}{F} \right\} &= \sum_i \frac{\partial}{\partial X_i} \left\{ r(F)\frac{(SS^+X)_i}{F} \right\} \\
&= \sum_i r'(F)\frac{\partial F}{\partial X_i}\frac{(SS^+X)_i}{F} \\
&\quad + r(F)\frac{\frac{\partial}{\partial X_i}\{(SS^+X)_i\}}{F} - r(F)\frac{\frac{\partial F}{\partial X_i}(SS^+X)_i}{F^2} \\
&= \sum_i r'(F)\left\{ \frac{\partial}{\partial X_i}\sum_{k,l} X_k X_l S^+_{kl} \right\}\frac{(SS^+X)_i}{F} \\
&\quad + r(F)\frac{\frac{\partial}{\partial X_i}\sum_k (SS^+)_{ik}X_k}{F} \\
&\quad - r(F)\frac{\left\{\frac{\partial}{\partial X_i}\sum_{k,l} X_k X_l S^+_{kl}\right\}(SS^+X)_i}{F^2} \\
&= \sum_i r'(F)\left\{(X'S^+)_i + (X'S^+)_i\right\}\frac{(SS^+X)_i}{F} \\
&\quad + r(F)\frac{(SS^+)_{ii}}{F} - r(F)\frac{\{(X'S^+)_i + (X'S^+)_i\}\cdot(SS^+X)_i}{F^2} \\
&= 2r'(F) + r(F)\frac{\mathrm{tr}(SS^+)-2}{F}
\end{aligned}$$

as desired. □

The following result is an extension of a result in Konno (2009). This type of result was first obtained by Kubokawa and Srivastava (2008) and then was extended by Konno (2009). In our generalization we make use of a divergence version of Stein's lemma that comes with somewhat weaker moment conditions, rather than the element-by-element assumptions in Konno (2009). These weaker moment conditions allow us to cover the p equals n and $n+1$ cases.

LEMMA 3. *Let $Y \sim N_{n\times p}(0, I_n \otimes \Sigma)$, let $S = Y'Y$ which has, by definition, a Wishart$_p(n, \Sigma)$ distribution, and let $G(S)$ be a $p \times p$ random matrix that depends on S. Let ∇_Y be interpreted as the matrix with components $(\nabla_Y)_{ij} = \frac{\partial}{\partial Y_{ij}}$, and for A the symmetric positive definite square root of Σ, define $\tilde{Y} = YA^{-1}$ and $H = AGA^{-1}$. Then*

$$E\Big[tr\left(\Sigma^{-1}SG\right)\Big] = E\Big[ntr(G) + tr\left(Y'\nabla_Y G'\right)\Big]$$

under the conditions

$$E\left[\left|div_{vec(\tilde{Y})} \cdot vec\left(\tilde{Y}H\right)\right|\right] < \infty \tag{3.10}$$

where vec(M) denotes the vectorization of a matrix M.

PROOF. Define $\tilde{S} = \tilde{Y}'\tilde{Y} = A^{-1}SA^{-1}$. Notice that, by construction, $\tilde{Y} \sim N_{n\times p}(0, I_n \otimes I_p)$ - this means, by definition of the matrix normal distribution, that $\mathrm{vec}(\tilde{Y}) \sim N_{np}(0, I_{np})$. We can write

$$E\left[\mathrm{tr}\left(\tilde{S}H\right)\right] = E\left[\sum_{\alpha,i,j} \tilde{Y}_{\alpha i}\tilde{Y}_{\alpha j}H_{ji}\right] = E\left[\mathrm{vec}(\tilde{Y}) \cdot \mathrm{vec}\left(\tilde{Y}H\right)\right].$$

Using the divergence form of Stein's lemma, which can be found in Lemma A.1 in Fourdrinier and Strawderman (2003), we obtain, under the moment conditions outlined in (3.10)

$$\begin{aligned}
E\left[\mathrm{vec}(\tilde{Y}) \cdot \mathrm{vec}\left(\tilde{Y}H\right)\right] &= E\left[\mathrm{div}_{\mathrm{vec}(\tilde{Y})}\ \mathrm{vec}\left(\tilde{Y}H\right)\right] \\
&= E\left[\sum_{\alpha,i,j} \frac{\partial}{\partial \tilde{Y}_{\alpha i}} \tilde{Y}_{\alpha j}H_{ji}\right] \\
&= E\left[\sum_{\alpha,i,j} \delta_{ij}H_{ji} + \tilde{Y}_{\alpha j}\frac{\partial H_{ji}}{\partial \tilde{Y}_{\alpha i}}\right] \\
&= E\left[n\sum_i H_{ii} + \sum_{\alpha,i,j} \tilde{Y}_{\alpha j}\frac{\partial}{\partial \tilde{Y}_{\alpha i}} H_{ji}\right].
\end{aligned}$$

This last expression can be expressed in a compact matrix form as

$$E\left[\mathrm{tr}\left(\tilde{S}H\right)\right] = E\left[n\mathrm{tr}(H) + \mathrm{tr}\left((\tilde{Y}'\nabla_{\tilde{Y}})'H\right)\right].$$

Finally, we notice

$$\begin{aligned}
&E\left[\operatorname{tr}(H)\right] = E\left[\operatorname{tr}\left(AGA^{-1}\right)\right]\\
&E\left[\operatorname{tr}\left(\tilde{S}H\right)\right] = E\left[\operatorname{tr}\left(A^{-1}SGA^{-1}\right)\right]\\
&E\left[\operatorname{tr}\left((\tilde{Y}'\nabla_{\tilde{Y}})'H\right)\right] = E\left[\operatorname{tr}\left(A(Y'\nabla_Y)'GA^{-1}\right)\right],
\end{aligned}$$

which concludes the proof. □

THEOREM 2. *Let $Y \sim N_{n\times p}(0, I_n \otimes \Sigma)$ and for A the symmetric positive definite square root of Σ, let $\tilde{Y} = YA^{-1}$. Let r be any bounded differentiable nonnegative function $r : \mathbb{R} \to [0, C_1]$ with bounded derivative $|r'| \leq C_2$. Define*

$$G = r^2(X'S^+X)\frac{S^+XX'S^+S}{(X'S^+X)^2}$$

and $H = AGA^{-1}$. Then for all p and n

$$E\left[\left|div_{vec(\tilde{Y})}\ vec\left(\tilde{Y}H\right)\right|\right] < \infty. \tag{3.11}$$

PROOF. We first compute $\operatorname{div}_{\operatorname{vec}(\tilde{Y})} \operatorname{vec}\left(\tilde{Y}H\right)$. As always, to ease notation we shall write $F = X'S^+X$. We have

$$\begin{aligned}
\operatorname{div}_{\operatorname{vec}(\tilde{Y})} \operatorname{vec}\left(\tilde{Y}H\right) &= \sum_{\alpha,i,j} \frac{\partial}{\partial \tilde{Y}_{\alpha i}}\left\{\tilde{Y}_{\alpha j}H_{ji}\right\}\\
&= n\sum_i H_{ii} + \sum_{\alpha,j} \tilde{Y}_{\alpha j}\frac{\partial H_{ji}}{\partial \tilde{Y}_{\alpha i}}\\
&= n\sum_i H_{ii} + \sum_{\alpha,\beta,i,j} \tilde{Y}_{\alpha j}A_{\beta i}\frac{\partial}{\partial Y_{\alpha\beta}}\left\{r^2(F)\frac{\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}}{F^2}\right\}\\
&= n\sum_i H_{ii} + \sum_{\alpha,\beta,i,j} \tilde{Y}_{\alpha j}A_{\beta i}\cdot
\end{aligned}$$

$$\left\{2r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\frac{\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}}{F^2}\right. \tag{3.12}$$

$$+\frac{r^2(F)}{F^2}\sum_{k,l} A_{jk}\frac{\partial\left\{S^+XX'SS^+\right\}_{kl}}{\partial Y_{\alpha\beta}}A^{-1}_{li} \tag{3.13}$$

$$\left.-r^2(F)\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}\frac{2\frac{\partial F}{\partial Y_{\alpha\beta}}}{F^3}\right\}. \tag{3.14}$$

We simplify each part of the expression. For (3.12), using Proposition 1 (ii), we find

$$
\begin{aligned}
&2\sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}r(F)r'(F)\frac{\partial F}{\partial Y_{\alpha\beta}}\frac{\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}}{F^2}\\
&\quad=4\frac{r(F)r'(F)}{F^2}\sum_{\alpha,\beta,i,j}\Big\{\\
&\qquad-(X'S^+Y')_\alpha\tilde{Y}_{\alpha j}\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}A_{i\beta}(S^+X)_\beta\\
&\qquad+(X'S^+S^+Y')_\alpha\tilde{Y}_{\alpha j}\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}A_{i\beta}((I-SS^+)X)_\beta\Big\}\\
&\quad=-4\frac{r(F)r'(F)}{F^2}\left(X'S^+Y'YA^{-1}AS^+XX'SS^+A^{-1}AS^+X\right)\\
&\qquad+4\frac{r(F)r'(F)}{F^2}\left(X'S^+S^+Y'YA^{-1}AS^+XX'SS^+A^{-1}A(I-SS^+)X\right)\\
&\quad=-4r(F)r'(F).
\end{aligned}
\tag{3.15}
$$

Similarly, for (3.14)

$$
\begin{aligned}
&\sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}r^2(F)\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}\frac{2\frac{\partial F}{\partial Y_{\alpha\beta}}}{F^3}\\
&\quad=4\frac{r^2(F)}{F^3}\sum_{\alpha,\beta,i,j}(X'S^+Y')_\alpha\tilde{Y}_{\alpha j}\left\{AS^+XX'SS^+A^{-1}\right\}_{ji}A_{i\beta}(S^+X)_\beta\\
&\quad=4\frac{r^2(F)}{F^3}\left(X'S^+Y'YA^{-1}AS^+XX'SS^+A^{-1}AS^+X\right)\\
&\quad=4\frac{r^2(F)}{F}.
\end{aligned}
\tag{3.16}
$$

This leaves us with (3.13). Using Proposition 1 (iii) we obtain

$$
\begin{aligned}
&\sum_{\alpha,\beta,i,j}\tilde{Y}_{\alpha j}A_{\beta i}\frac{r^2(F)}{F^2}\sum_{k,l}A_{jk}\frac{\partial\left\{S^+XX'SS^+\right\}_{kl}}{\partial Y_{\alpha\beta}}A^{-1}_{li}\\
&\quad=\frac{r^2(F)}{F^2}\sum_{\alpha,\beta,i,j,k,l}\tilde{Y}_{\alpha j}A_{\beta i}A_{jk}A^{-1}_{li}\\
&\qquad\cdot\Big\{(S^+S^+Y)_{k\alpha}((I-SS^+)XX'SS^+)_{\beta l}\\
&\qquad-S^+_{k\beta}(YS^+XX'SS^+)_{\alpha l}\\
&\qquad-(S^+Y)_{k\alpha}(S^+XX'SS^+)_{\beta l}
\end{aligned}
$$

$$
\begin{aligned}
&+ (I - SS^+)_{k\beta}(YS^+S^+XX'SS^+)_{\alpha l} \\
&+ (S^+XX')_{k\beta}(YS^+)_{\alpha l} \\
&+ (S^+XX'Y')_{k\alpha}(S^+)_{\beta l} \\
&+ (S^+XX'S^+Y')_{k\alpha}(I - SS^+)_{\beta l} \\
&- (S^+XX'SS^+)_{k\beta}(YS^+)_{\alpha l} \\
&-(S^+XX'SS^+Y')_{k\alpha}(S^+)_{\beta l}\big\} \\
=&\frac{r^2(F)}{F^2}\sum_{\alpha,\beta,i,j,k,l}\Big\{ \\
&A_{jk}(S^+S^+Y)_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}((I - SS^+)XX'SS^+)_{\beta l}A_{li}^{-1} \\
&- \tilde{Y}'_{j\alpha}(YS^+XX'SS^+)_{\alpha l}A_{li}^{-1}A_{i\beta}S^+_{\beta k}A_{kj} \\
&- A_{jk}(S^+Y)_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(S^+XX'SS^+)_{\beta l}A_{li}^{-1} \\
&+ \tilde{Y}'_{j\alpha}(YS^+S^+XX'SS^+)_{\alpha l}A_{li}^{-1}A_{i\beta}(I - SS^+)_{\beta k}A_{kj} \\
&+ \tilde{Y}'_{j\alpha}(YS^+)_{\alpha l}A_{li}^{-1}A_{i\beta}(XX'S^+)_{\beta k}A_{kj} \\
&+ A_{jk}(S^+XX'Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(S^+)_{\beta l}A_{li}^{-1} \\
&+ A_{jk}(S^+XX'S^+Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(I - SS^+)_{\beta l}A_{li}^{-1} \\
&- \tilde{Y}'_{j\alpha}(YS^+)_{\alpha l}A_{li}^{-1}A_{i\beta}(SS^+XX'S^+)_{\beta k}A_{kj} \\
&- A_{jk}(S^+XX'SS^+Y')_{k\alpha}\tilde{Y}_{\alpha j}A_{i\beta}(S^+)_{\beta l}A_{li}^{-1}\Big\} \\
=&\frac{r^2(F)}{F^2}\big\{\mathrm{tr}(AS^+S^+Y'YA^{-1})\cdot\mathrm{tr}(A(I - SS^+)XX'SS^+A^{-1}) \\
&- \mathrm{tr}(A^{-1}Y'YS^+XX'SS^+A^{-1}AS^+A) \\
&- \mathrm{tr}(AS^+Y'YA^{-1})\mathrm{tr}(AS^+XX'SS^+A^{-1}) \\
&+ \mathrm{tr}(A^{-1}Y'YS^+S^+XX'SS^+A^{-1}A(I - SS^+)A) \\
&+ \mathrm{tr}(A^{-1}Y'YS^+A^{-1}AXX'S^+A) \\
&+ \mathrm{tr}(AS^+XX'Y'YA^{-1})\cdot\mathrm{tr}(AS^+A^{-1}) \\
&+ \mathrm{tr}(AS^+XX'S^+Y'YA^{-1})\mathrm{tr}(A(I - SS^+)A^{-1}) \\
&- \mathrm{tr}(A^{-1}Y'YS^+A^{-1}ASS^+XX'S^+A) \\
&-\mathrm{tr}(AS^+XX'SS^+Y'YA^{-1})\mathrm{tr}(AS^+A^{-1})\big\} \\
=&\frac{r^2(F)}{F^2}\cdot\big\{-X'S^+X - \mathrm{tr}(SS^+)\cdot X'S^+X \\
&+ X'S^+X + X'SS^+X\cdot\mathrm{tr}(S^+) + X'S^+X\cdot(p - \mathrm{tr}(SS^+)) \\
&-X'S^+X - X'SS^+X\cdot\mathrm{tr}(S^+)\big\}
\end{aligned}
$$

$$= \frac{r^2(F)}{F}\Big(p - \mathrm{tr}(SS^+) - 1\Big). \tag{3.17}$$

Having reexpressed $\mathrm{div}_{\mathrm{vec}(\tilde{Y})}\ \mathrm{vec}\left(\tilde{Y}H\right)$, we now need to bound it above. By virtue of (3.15), (3.16) and (3.17), we have

$$\begin{aligned}
&E\left[\left|\mathrm{div}_{\mathrm{vec}(\tilde{Y})}\ \mathrm{vec}\left(\tilde{Y}H\right)\right|\right] \\
&= E\left[\left| n\mathrm{tr}(H) + 4\frac{r^2(F)}{F}\right.\right. \\
&\qquad \left.\left. + \ \Big(p - \mathrm{tr}(SS^+) - 1\Big)\frac{r^2(F)}{F} - 4r(F)r'(F)\right|\right] \\
&\leq C_1^2\left|3 + p - \mathrm{tr}(SS^+) + n\right| E\left[\frac{1}{F}\right] + 4C_1C_2. \qquad (3.18)
\end{aligned}$$

It only remains to show that $E\left[\frac{1}{F}\right]$ is finite. By definition of the Wishart matrix distribution we can define a $T \sim \mathrm{Wishart}_p(n, I_n)$ such that $S = ATA$. Let $T = H'DH$ be the spectral decomposition of T, with $D = \mathrm{diag}(\lambda_i)$. Write the eigenvalues of T^+ as λ_i^+, so that $D^{-1} = \mathrm{diag}(\lambda_i^+)$, and let $\lambda_{\min}^+$ be the smallest nonzero eigenvalue of T^+. The following two identities follow from Tian and Cheng (2004) [Theorem 1.1, equations (1.2) and (1.4)] and symmetry of T:

$$\begin{aligned}
(ATA)^+ &= (T^+TA)^+T^+(AT^+T)^+ \\
(T^+TA)^+(T^+T) &= (T^+TA)^+.
\end{aligned}$$

Using these identities we have

$$\begin{aligned}
X'S^+X &= X'(ATA)^+X = X'(T^+TA)^+T^+(AT^+T)^+X \\
&= \sum_k \left\{X'(T^+TA)^+H'\right\}_k^2 \lambda_k^+ \\
&\geq \lambda_{\min}^+ \cdot X'(T^+TA)^+H'H(AT^+T)^+X \\
&= \lambda_{\min}^+ \cdot X'(T^+TA)^+(T^+T)(AT^+T)^+X \\
&= \lambda_{\min}^+ \cdot X'(T^+TA)^+(AT^+T)^+X.
\end{aligned}$$

Applying Cauchy-Schwartz provides us the bound

$$X'(T^+TA)^+(T^+TA)X \leq \ X'(T^+TA)^+(AT^+T)^+XX'(AT^+T)(T^+TA)X$$

so that we then have

$$\begin{aligned}
\frac{1}{F} = \frac{1}{X'S^+X} &\leq \frac{1}{\lambda_{\min}^+}\frac{1}{X'(T^+TA)^+(AT^+T)^+X} \\
&\leq \frac{1}{\lambda_{\min}^+}\frac{X'AT^+TAX}{X'(T^+TA)^+(T^+TA)X}.
\end{aligned}$$

To ease notation let us write $Q = AT^+TA$ and $R = (T^+TA)^+(T^+TA)$. Collecting the results together we bound (3.18) by

$$\leq C_1^2 \left|3 + p - 2\operatorname{tr}(SS^+) + n\right| E\left[\frac{1}{\lambda_{\min}^+}\frac{X'QX}{X'RX}\right] + 4C_1C_2. \tag{3.19}$$

We now use some independence results. We can write the singular value decomposition of T as $T = H'DH$, but we can also write it as $T = H_1'D_1H_1$, where H_1 is semi-orthogonal ($H_1H_1' = I$), and D_1 is the matrix of the positive eigenvalues of T. If T has full rank (i.e., $n \geq p$) then this coincide with the singular value decomposition of T. In the full rank case, Srivastava and Khatri (1979) [Section 3.4, equation (3.4.3)] provides the joint density of H and $D = \operatorname{diag}(d_i)$ in the standard Wishart case (which applies to T) as

$$\begin{aligned} &f_{H,D}(H,D) \\ &= C(p,n)|D|^{\frac{1}{2}(n-p-1)}\left[\operatorname{etr}\left(-\frac{1}{2}D\right)\right]\left[\prod_{i<j}(d_i - d_j)\right] g_p(H) \end{aligned} \tag{3.20}$$

for constants $C(p,n)$ and functions g_p. Therefore, H and D are independent. In the rank-deficient case ($p > n$), Srivastava (2003) [Section 3] provides an equivalent expression which, in the singular Wishart case gives

$$\begin{aligned} &f_{H_1,D_1}(H_1,D_1) \\ &= K(p,n)|D_1|^{\frac{1}{2}(p-n-1)}\left[\operatorname{etr}\left(-\frac{1}{2}D_1\right)\right]\left[\prod_{i<j}(d_i - d_j)\right] g_{n,p}(H_1) \end{aligned} \tag{3.21}$$

for constants $K(p,n)$ and functions $g_{n,p}$ so again, we find H_1 and D_1 independent by factorization. Now, $\lambda_{\min}^+$ is a function, in the full rank case (resp. rank-deficient case), of only D^{-1} (resp. D_1^{-1}), and we can write $T^+T = H'H$ (resp. $T^+T = H_1'H_1$), so $\lambda_{\min}^+$ and T^+T are independent. Being functions of S, they are also both independent of X. Now, the nonzero eigenvalues of T^+ are the inverses of the nonzero eigenvalues of T, a general fact about Moore-Penrose pseudo-inverses. Therefore, denoting the largest eigenvalue of T as $\lambda_{\max}$, we can split up the expectations in (3.19) and get the bound

$$\leq C_1^2 \left|3 + p - 2\operatorname{tr}(SS^+) + n\right| E\left[\lambda_{\max}\right] E\left[\frac{X'QX}{X'RX}\right] + 4C_1C_2. \tag{3.22}$$

Now, it follows from positive semi-definiteness of T that $E\left[\lambda_{\max}\right] \leq E\left[\operatorname{tr}(T)\right]$. If $n \geq p$, $\operatorname{tr}(T) \sim \chi^2_{pn}$ (cf. Muirhead (1982), Theorem 3.2.20)

and so $E[\mathrm{tr}(T)] = pn < \infty$. If $p > n$, recall we can write $T = Z'Z$ for $Z \sim N_{n\times p}(0, I_n \otimes I_p)$ by definition of the Wishart distribution; and $ZZ' \sim \mathrm{Wishart}_n(p, I_n)$ so that $\mathrm{tr}(T) = \mathrm{tr}(ZZ') \sim \chi^2_{pn}$; so again, $E[\mathrm{tr}(T)] = pn < \infty$. Therefore, in either case,$E[\lambda_{\max}] \leq pn < \infty$.

We still have to check that the expectation involving X, Q and R in (3.22) is finite. Let $r = \mathrm{rk}(R) = \mathrm{rk}(Q) = \mathrm{rk}(S)$ and write the spectral decomposition of (T^+TA) as $U\Lambda U'$, with $\Lambda = \mathrm{diag}(L, 0_{(p-r)})$ where L is the vector of the r nonzero eigenvalues of (T^+TA). Then $R = (T^+TA)^+(T^+TA) = U\mathrm{diag}(I_r, 0_{(p-r)})U'$; let us define the $p\times(p-r)$ matrix $E = U\,[0_{(p-r)\times r}\ I_{(p-r)}]'$ i.e. so that $RE = 0$ and E has full column rank $p-r$. Notice that $QE = AT^+TAU\,[0_{(p-r)\times r}\ I_{(p-r)}]' = AU\Lambda U'U\,[0_{(p-r)\times r}\ I_{(p-r)}]' = 0$. Since Q and R are symmetric positive semidefinite, we can use results in Magnus (1990) [Theorem 1(i) with $A = Q$ and $B = R$] to conclude that

$$E\left[\frac{X'QX}{X'RX}\right] < \infty.$$

This concludes the proof of the theorem. □

4. Numerical Study. This section provides some numerical results to showcase the improvement in risk of the minimax estimator over the usual estimator. More precisely, we compared the James-Stein estimator in (2.6) given by

$$\delta^{\mathrm{JS}} = \left(I - \frac{(n-2)SS^+}{(p-n+3)X'S^+X}\right)X$$

and the usual estimator $\delta^0 = X$ under invariant loss. (In addition, we considered the positive James-Stein estimator to be discussed in Section 5.) The empirical approximations of the invariant risk of these estimators were plotted for $p = 10, 20, 50$ and $n = \frac{p}{2}, p-1$. Three covariance matrix structures were considered:

Spiked A diagonal matrix with the first $p/2$ diagonal elements equal to 1, and the last $p/2$ equal to 10.

Autoregressive Autoregressive covariance matrices of the form

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \\ \rho & 1 & \rho & \\ \rho^2 & \rho & 1 & \\ & & & \ddots \end{pmatrix}$$

for $\rho = 0.5$.

Block diagonal Block diagonal matrices with $p/2$ blocks of the form $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ for $\rho = 0.5$.

In all cases, the true mean was chosen as $\theta \propto (1, ..., 1)$.

We remind the reader that the risk of the trivial estimator is always p, regardless of θ or Σ. With this in mind, we see from Figure 1 that in all six scenarios the pattern of domination of the new estimator is similar to one of the usual James-Stein estimator. Also note that, as predicted by the theoretical results, the domination decreases as the smaller n tends to p.

5. Comments. An interesting property of the Moore-Penrose inverse is that for any A, AA^+ is the matrix that projects onto the subspace spanned by A (its column space.) It follows that the proposed generalized Baranchik estimator can be expressed as

$$\begin{aligned} \delta_r(X, S) &= (I - SS^+)X + \left(1 - \frac{r(X'S^+X)}{X'S^+X}\right) SS^+X \\ &= P_{S^\perp}X + \left(1 - \frac{r(X'S^+X)}{X'S^+X}\right) P_S X \end{aligned} \tag{5.1}$$

where $P_S = SS^+$ and $P_{S^\perp} = I - SS^+$ are the projection matrices onto the column space of S and its orthogonal complement, respectively. In terms of the kernel and image of the symmetric matrix S, $\mathrm{Ker}(P_{S^\perp}) = \mathrm{Im}(S)$ and $\mathrm{Im}(P_{S^\perp}) = \mathrm{Ker}(S^+)$. When $p > n$, this means we can interpret our estimator as applying shrinkage only on the component of X in the subspace spanned by our covariance matrix estimator S. In particular, note that the estimator $P_S\delta_r(X, S) = \left(1 - \frac{r(X'S^+X)}{X'S^+X}\right) P_S X$ dominates $P_S X$ under invariant loss function (1.1), since $R(P_S\delta_r, \theta) - R(P_S X, \theta) = R(\delta_r, \theta) - R(X, \theta) \geq 0$ if r satisfies the conditions of Theorem 1. This suggests there might be an easier, more abstract proof of Theorem 1, one not relying on brute computations but on the already known full rank S case, although we have not been able to obtain such a result.

A natural extension of the James-Stein estimator, δ_a^{JS} in (2.6), is a positive-part-type James-Stein estimator. The form of the estimator in (5.1) suggests

$$\delta_a^{\mathrm{JS}+} = \left(I - SS^+\right) X + \left(1 - \frac{a}{X'S^+X}\right)_+ SS^+X, \tag{5.2}$$

where $b_+ = \max(b, 0)$. Simulation evidence from Table 1 suggests that for $a = (n-2)/(p-n+3)$, $\delta_a^{\mathrm{JS}+}$ dominates δ_a^{JS} under invariant loss.

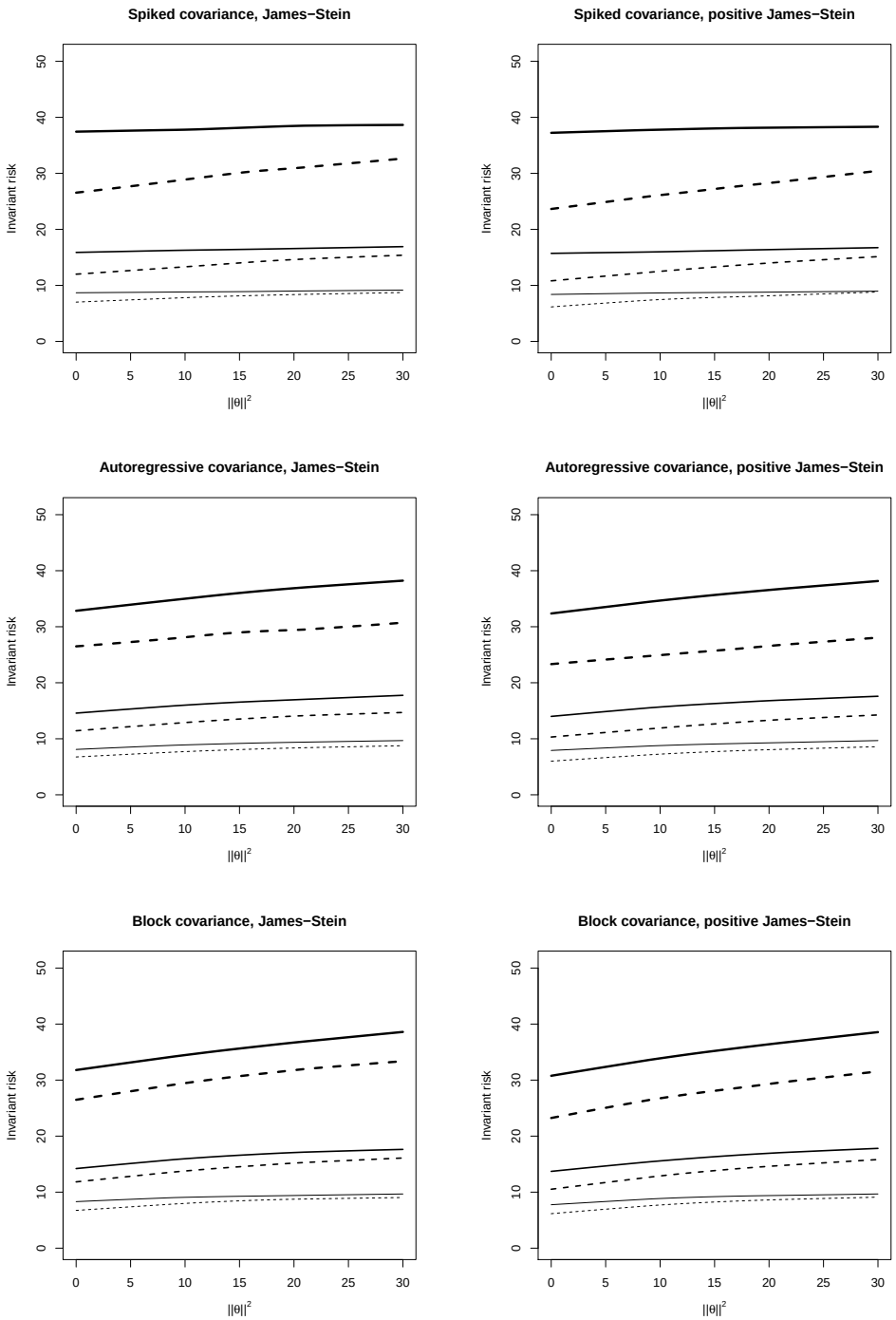

FIG 1. *The risk function plots of* δ_a^{JS} *and* δ_a^{JS+} *for* $a = (n-2)/(p-n+3)$ *are in the left and right columns, respectively. The lines, from thinnest to thickest, are for* $p = 10, 20$ *and* 50. *The solid and dashed lines are respectively for* $n = p/2$ *and* $n = p - 1$.

One of the interesting differences between the $n > p$ and $p > n$ cases is the reversal of the roles of p and n. This is essentially due to the distribution of the singular values of S. Recall that for $S = ATA$, $T \sim W_p(n, I_n)$. We can write the singular value decomposition of T as $T = H'DH$, but we can also write it as $T = H_1' D_1 H_1$, where H_1 is semi-orthogonal ($H_1 H_1' = I$), and D_1 is the matrix of the positive eigenvalues of T. If T has full rank (i.e., $n \geq p$) this coincides with the singular value decomposition of T. In the full rank case the joint density of H and D is given in (3.20), whereas in the rank-deficient case ($p > n$) joint density is given by (3.21), from which stems the reversal of the roles of p and n.

In the heteroscedastic normal mean estimation problem, James and Stein (1961) used the loss function that was weighted by the inverse of the variances, and consequently the problem is essentially transformed to the homoscedastic case under ordinary squared error loss. Similarly in this article, we used the invariant loss function in (1.1), therefore skirting a somewhat subtle issue. In the heteroscedastic setting where there are differing coordinate variances, minimax estimation and Bayes (or empirical Bayes) estimates can be qualitatively different. It turns out that minimax estimators in general shrink most on the coordinates with smaller variances, while Bayes estimators shrink most on large variance coordinates. Brown (1975) shows that the James-Stein shrinkage estimator does not dominate the X when the largest variance is larger than the sum of the rest. Moreover, Casella (1980) points out that the James-Stein shrinkage estimator may not be a desirable shrinkage estimator under heteroscedasticity even when it is minimax. Morris and Lysy (2009) and Brown, Nie and Xie (2012) give an excellent perspective on minimaxity of shrinkage estimator from Bayes and empirical Bayes points of view. Consequently, it would be of interest to examine the shrinkage patterns of the proposed estimates in the case of a non-invariant loss function and assess how well the invariant loss works for $p > n$ applications.

One can imagine an extension of the results of this article beyond the normal distribution setting. Consider a model with the joint density for (X, S) the form

$$f\Big(\operatorname{tr} \Sigma^{-1}[(X-\theta)(X-\theta)' + S]\Big) \tag{5.3}$$

where the $p \times 1$ location vector θ and the $p \times p$ scale matrix Σ are unknown. In the setting of $p \leq n$, Fourdrinier, Strawderman and Wells (2003) and Kubokawa and Srivastava (2001) give some results on improved location estimation for elliptically symmetric distributions. For more on elliptical symmetry and the various choices of $f(\cdot)$ in (5.3), see Fang, Kotz and Ng

(1990); the class in (5.3) contains models such as the multivariate normal, $t-$, and Kotz-type distributions.

Finally, simulation study reveals that, when p is much larger than n, the estimate of Σ and Σ^{-1} are quite poor. This observation agrees with Kubokawa and Srivastava (2008), where Haff (1979)-type improved estimates of Σ are proposed. It would be of interest to use an improved estimator of Σ in $\delta_r(X, S)$ in (2.1). As pointed out in the testing context by Srivastava and Fujikoshi (2006) and Srivastava (2007), a shortcoming of S^+ is that the associated estimator is only orthogonally invariant, while the sample mean vector is invariant.

Acknowledgments. The authors are grateful to the Associate Editor and referees for helpful comments that strengthened the exposition and scope of this paper.

REFERENCES

Aspremont, A., Banerjee, O. and Ghaoui, L. E. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications* **30** 56–66.

Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Annals of Mathematical Statistics* **41** 642–656.

Berger, J. O. and Bock, M. E. (1976). Combining independent normal mean estimation problems with unknown variances. *Annals of Statistics* **4** 642–648.

Berger, J. O. and Haff, L. R. (1983). A class of minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Statistics & Decisions* **1** 105–130.

Berger, J. O., Bock, M. E., Brown, L. D., Casella, G. and Gleser, L. (1977). Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Annals of Statistics* **5** 763–771.

Bickel, P. J. and Levina, E. (2008). Regulatized estimation of large covariance matrices. *Annals of Statistics* **36** 190–227.

Brown, L. (1975). Estimation with incomplete specified loss functions (the case with several location parameters). *Journal of the American Statistical Association* **70** 417-427.

Brown, L. D., Nie, H. and Xie, X. (2012). Ensemble minimax estimation for multivariate normal means. *Annals of Statistics* To appear.

Casella, G. (1980). Minimax ridge regression estimation. *Annals of Statistics* **8** 1036-1056.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall.

Fourdrinier, D. and Strawderman, W. E. (2003). On Bayes and unbiased estimators of loss. *Annals of the Institute of Statistical Mathematics* **55** 803–816.

Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (2003). Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *Journal of Multivariate Analysis* **85** 24–39.

Gleser, L. J. (1979). Minimax estimation of a normal mean vector when the covariance matrix is unknown. *Annals of Statistics* **7** 838–846.

Gleser, L. J. (1986). Minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Annals of Statistics* **14** 1625–1633.

Golub, G. H. and Pereyra, V. (1973). The differentiation of pseudo-inverses and non-linear least squares problems whose variables separate. *SIAM Journal of Numerical Mathematics* **10** 413–432.

Haff, L. R. (1979). An identity for the Wishart distribution with applications. *Journal of Multivariate Analysis* **9** 531–544.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 361–380. Berkeley, University of California Press.

Konno, Y. (2009). Shrinkage estimators for large covariance matrices in multivariate real and complex normal distributions under an invariant quadratic loss. *Journal of Multivariate Analysis* **100** 2237–2253.

Kubokawa, T. and Srivastava, M. S. (2001). Robust improvement in estimation of a mean matrix in an elliptically contoured distribution. *Journal of Multivariate Analysis* **76** 138–152.

Kubokawa, T. and Srivastava, M. S. (2008). Estimation of the precision matrix of a singular Wishart distribution and its application in high dimensional data. *Journal of Multivariate Analysis* **99** 1906–1928.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88** 365–411.

Levina, E., Rothman, A. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Annals of Applied Statistics* **2** 245–263.

Magnus, J. R. (1990). On certain moments relating to ratios of quadratic forms in normal variables: further results. *Sankhyā B* **52** 1–13.

Morris, C. and Lysy, M. (2009). Shrinkage estimation in multi-level normal models. *Statistical Science* **27** 115-134.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

Srivastava, M. S. (2003). Singular Wishart and multivariate beta distributions. *Annals of Statistics* **31** 1537–1560.

Srivastava, M. S. (2007). Multivariate theory for analyzing high dimensional data. *Journal of the Japanese Statistical Society* **37** 53–86.

Srivastava, M. S. and Bilodeau, M. (1989). Stein estimation under elliptical distributions. *Journal of Multivariate Analysis* **28** 247–259.

Srivastava, M. S. and Fujikoshi, Y. (2006). Multivariate analysis of variance with fewer observations than the dimension. *Journal of Multivariate Analysis* **97** 1927–1940.

Srivastava, M. S. and Khatri, C. G. (1979). *An Introduction to Multivariate Statistics.* North-Holland, New York.

Stein, C. (1977). Lectures on the theory of estimation of many parameters. In *Studies in the Statistical Theory of Estimation, Part I.* **74** 4–65. Proc. Scientific Seminars Steklov Institute, Leningrad Division (In Russian).

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

Tian, Y. and Cheng, S. (2004). Some identities for Moore-Penrose inverses of matrix products. *Linear and Multilinear Algebra* **52** 405–420.

Department of Statistical Science
Comstock Hall
Cornell University
Ithaca, New York 14853
USA
E-mail: dc623@cornell.edu
mtw1@cornell.edu