Electronic Journal of Statistics
Vol. 6 (2012) 2347–2355
ISSN: 1935-7524
DOI: 10.1214/12-EJS748

Shrinkage estimation with a matrix loss function

Reman Abu-Shanab

Department of Mathematics, University of Bahrain
e-mail: raboshanab@sci.uob.bh

John T. Kent

Department of Statistics, University of Leeds
e-mail: j.t.kent@leeds.ac.uk
url: www.maths.leeds.ac.uk/~john

and

William E. Strawderman

Department of Statistics, Rutgers University
e-mail: straw@stat.rutgers.edu
url: www.stat.rutgers.edu/people/faculty/straw.html

Abstract: Consider estimating an $n \times p$ matrix of means Θ, say, from an $n \times p$ matrix of observations X, where the elements of X are assumed to be independently normally distributed with $E(x_{ij}) = \theta_{ij}$ and constant variance, and where the performance of an estimator is judged using a $p \times p$ matrix quadratic error loss function. A matrix version of the James-Stein estimator is proposed, depending on a tuning constant a. It is shown to dominate the usual maximum likelihood estimator for some choices of a when $n \geq 3$. This result also extends to other shrinkage estimators and settings.

AMS 2000 subject classifications: Primary 62C99; secondary 62H12.
Keywords and phrases: James-Stein estimator, matrix quadratic loss function, risk, Stein's Lemma.

Received August 2012.

1. Introduction

Shrinkage estimators are usually set in the context of *vector* data. In the simplest version the data follow a normal distribution $\mathbf{x} \sim N_n(\boldsymbol{\theta}, I_n)$, where $\boldsymbol{\theta}$ is an n-dimensional column vector of parameters. That is, the x_i are independent normal variates from $N(\theta_i, 1)$. Let $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}$. There are two natural loss functions in this setting: the $n \times n$ matrix loss

$$L_{\text{matrix},n}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T, \tag{1.1}$$

and the scalar loss function

$$L_{\text{scalar}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2. \tag{1.2}$$

The $n \times n$ matrix loss function focuses on the error for each linear combination of the elements of $\boldsymbol{\theta}$; the scalar loss function pools the errors across the elements of $\boldsymbol{\theta}$.

It is well-known that under the matrix loss function, the simple unbiased estimator $\hat{\boldsymbol{\theta}}_0 = \mathbf{x}$ is admissible since domination in the matrix norm would imply domination in one dimension. Hence under the matrix norm there is no estimator with a smaller risk function for all $\boldsymbol{\theta}$. However, for scalar loss when $n \geq 3$, it is possible to improve upon $\hat{\boldsymbol{\theta}}_0$ through shrinkage by "borrowing strength" across the elements of $\mathbf{x}$.

Next consider the matrix case where a matrix $X(n \times p)$ of observations from a matrix normal distribution with mean Θ and identity covariance matrix is available. That is, $x_{ij} \sim N(\theta_{ij}, 1)$ independently for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. There are several loss functions that might be considered in this situation. We shall focus on two choices: the $p \times p$ matrix loss

$$L_{\text{matrix},p}(\hat{\Theta}, \Theta) = (\hat{\Theta} - \Theta)^T(\hat{\Theta} - \Theta), \tag{1.3}$$

and the scalar loss function

$$L_{\text{scalar}}(\hat{\Theta}, \Theta) = \text{tr}\{(\hat{\Theta} - \Theta)^T(\hat{\Theta} - \Theta)\} = \sum_{ij}(\hat{\theta}_{ij} - \theta_{ij})^2. \tag{1.4}$$

There are also $n \times n$ and $np \times np$ versions of the matrix loss which will not concern us here.

The matrix loss function (1.3) accommodates estimators which borrow strength across rows but not columns; the scalar loss function (1.4) accommodates estimators which borrow strength across rows and columns. It seems most natural to use the matrix choice when the rows measure commensurate quantities while the columns are qualitatively different. For example, the data might represent the performance of n different schools on $p = 2$ variables such as academic performance and the socio-economic background of the student body.

The matrix loss function (1.3) for matrix data is the focus of this paper. Some authors have considered the use of shrinkage methods in a matrix setting; see, e.g. Efron and Morris (1972, 1976); Haff (1977); Zheng (1986); Ghosh and Shieh (1991), Tsukuma and Kubokawa (2007); Tsukuma (2009). However, these papers use the scalar loss function (1.4). The current paper seems to give the first results showing that shrinkage can yield improvements over $\hat{\Theta}_0 = X$ with the matrix loss function (1.3).

Section 2 reviews the situation for vector data $\mathbf{x}$ in a setting which generalizes the simple normal case set out here. Section 3 gives our new results for matrix data X and the matrix loss function (1.3). A discussion of the implications of our results is given in Section 4.

2. Review of the vector case

If $\mathbf{x}(n \times 1)$ is a random vector with mean $\boldsymbol{\theta}$, then a shrinkage estimator of $\boldsymbol{\theta}$ takes the form

$$\hat{\boldsymbol{\theta}}_a = \mathbf{x} - a\mathbf{g}(\mathbf{x}; u), \tag{2.1}$$

where $a > 0$ is a tuning parameter, and $\mathbf{g}(\mathbf{x}, u)$ $(n \times 1)$ is a "shrinkage function", depending on the data $\mathbf{x}$, and possibly on extra information in an auxiliary random variable u. If the auxiliary random variable is not present, the notation for the shrinkage function can be simplified to $\mathbf{g}(\mathbf{x})$. In Section 3 for $n \times p$ matrix data, the the auxiliary information will be generalized to a p-dimensional random vector $\mathbf{u}$.

Let $F(\mathbf{x}, u)$ denote the joint distribution of $\mathbf{x}$ and u, depending on $\boldsymbol{\theta}$. The classic James-Stein estimator (Stein, 1956; James and Stein, 1961) is a special case in the setting $\mathbf{x} \sim N_n(\boldsymbol{\theta}, \sigma^2 I_n)$, $n \geq 3$. When σ^2 is known, the shrinkage function is given by

$$\mathbf{g}(\mathbf{x}) = \sigma^2(n-2)\mathbf{x}/||\mathbf{x}||^2; \tag{2.2}$$

note the auxiliary random variable is not present in this case. When σ^2 is unknown, the shrinkage function is given by

$$\mathbf{g}(\mathbf{x}, u) = \{u/(\nu+2)\}(n-2)\mathbf{x}/||\mathbf{x}||^2, \tag{2.3}$$

where $u \sim \sigma^2\chi^2_\nu$ is an auxiliary random variable independent of $\mathbf{x}$ which is used to estimate σ^2.

The objective in shrinkage estimation is to estimate the vector parameter $\boldsymbol{\theta}$, where the performance of an estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$ is judged by the scalar loss function (1.2) and associated risk function $R_{\text{scalar}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = E_F\{L_{\text{scalar}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\}$.

In order to guarantee that the shrinkage estimator dominates the simple unbiased estimator $\hat{\boldsymbol{\theta}}_0 = \mathbf{x}$, the usual strategy is to demonstrate the "cross-product inequality"

$$E_F\{(\mathbf{x}-\boldsymbol{\theta})^T\mathbf{g}\} \geq E_F(\mathbf{g}^T\mathbf{g}) > 0, \tag{2.4}$$

for all $\boldsymbol{\theta}$, where $\mathbf{g} = \mathbf{g}(\mathbf{x}, u)$ is a function of the random vector $\mathbf{x}$ (and of u when present). The last inequality has been included to ensure that $\mathbf{g}$ is nontrivial. Throughout the paper we assume that $\mathbf{x}$ and $\mathbf{g}(\mathbf{x}, u)$ have finite second moments. In particular, this property is true when $\mathbf{x}$ is normally distributed and $n \geq 3$. Then the following well-known result holds.

Theorem 1. *Let $\mathbf{x}(n \times 1)$ be a random vector and u be an auxiliary random variable such that $E_F(\mathbf{x}) = \boldsymbol{\theta}$ under a probability model F depending on $\boldsymbol{\theta}$. Also suppose there exists a shrinkage function $\mathbf{g} = \mathbf{g}(\mathbf{x}, u)$ such that the cross-product inequality (2.4) holds. Then the shrinkage estimator $\hat{\boldsymbol{\theta}}_a$ in (2.1) dominates the simple estimator $\hat{\boldsymbol{\theta}}_0 = \mathbf{x}$ under the scalar loss function (1.2) provided the tuning parameter a satisfies $0 < a < 2$.*

Proof. Write $\delta = E_F\{(\mathbf{x}-\boldsymbol{\theta})^T\mathbf{g}\}$ and $\gamma = E_F(\mathbf{g}^T\mathbf{g})$, so $\delta \geq \gamma > 0$. Then the risk takes the form

$$\begin{aligned} R_{\text{scalar}}(\hat{\boldsymbol{\theta}}_a, \boldsymbol{\theta}) &= E_F\{(\mathbf{x}-\boldsymbol{\theta}-a\mathbf{g})^T(\mathbf{x}-\boldsymbol{\theta}-a\mathbf{g})\} \\ &= E_F\{(\mathbf{x}-\boldsymbol{\theta})^T(\mathbf{x}-\boldsymbol{\theta})\} - 2a\delta + a^2\gamma \\ &\leq E_F\{(\mathbf{x}-\boldsymbol{\theta})^T(\mathbf{x}-\boldsymbol{\theta})\} - 2a\gamma + a^2\gamma \\ &< E_F\{(\mathbf{x}-\boldsymbol{\theta})^T(\mathbf{x}-\boldsymbol{\theta})\} = R_{\text{scalar}}(\hat{\boldsymbol{\theta}}_0, \boldsymbol{\theta}), \end{aligned} \tag{2.5}$$

provided $0 < a < 2$. □

For the James-Stein estimator with $\mathbf{g}(\mathbf{x})$ given by (2.2), Stein's Lemma (Stein, 1981) states that the cross-product inequality (2.4) for known σ^2 holds for $n \geq 3$ and is actually an equality. That is, if $\mathbf{x} \sim N_n(\boldsymbol{\theta}, \sigma^2 I_n),\ n \geq 3$, then

$$\sigma^2 E_F\left[\{\mathbf{x}^T(\mathbf{x}-\boldsymbol{\theta})/||\mathbf{x}||^2\right\} = (n-2)\sigma^4 E_F\left\{1/||\mathbf{x}||^2\right\} = \sigma^2 A, \text{ say,} \tag{2.6}$$

where $A = A(\lambda^2)$ depends on $\lambda^2 = \boldsymbol{\theta}^T\boldsymbol{\theta}/\sigma^2$ and $0 < A < \infty$. Stein's Lemma can be proved using integration by parts (e.g. Efron and Morris (1976) or Stein (1981)). An equality also holds in the analogue of (2.6) for the unknown σ^2 case since $E(u) = \nu\sigma^2,\ E(u^2) = \nu(\nu+2)\sigma^2$ in (2.3). Hence Theorem 1 for the James-Stein estimator, in both the known and unknown σ^2 cases, can be strengthened to conclude that the optimal value of the tuning constant is $a = 1$, uniformly over all $\boldsymbol{\theta}$.

As a simple example where the cross-product inequality is strict, consider a Baranchik-type estimator (Baranchik, 1970) with

$$\mathbf{g}(\mathbf{x}) = \left\{(n-2)r(||\mathbf{x}||^2)/||\mathbf{x}||^2)\right\}\mathbf{x},$$

where $r(||\mathbf{x}||^2)$ is differentiable, bounded between 0 and 1, and strictly increasing. Then under normality, $\mathbf{x} \sim N_n(\boldsymbol{\theta}, I_n),\ n \geq 3$,

$$\begin{aligned} E_F\left\{(\mathbf{x}-\boldsymbol{\theta})^T\mathbf{g}(\mathbf{x})\right\} &= E_F\left\{(n-2)^2\frac{r(||\mathbf{x}||^2)}{||\mathbf{x}||^2} + 2(n-2)r'(||\mathbf{x}||^2)\right\} \\ &> E_F\left\{(n-2)^2\frac{r^2(||\mathbf{x}||^2)}{||\mathbf{x}||^2}\right\} = E_F\left\{||\mathbf{g}(\mathbf{x})||^2\right\}. \end{aligned}$$

3. Matrix data

Suppose the data take the form of an $n \times p$ matrix X, plus auxiliary random variables $\mathbf{u} = (u_1, \ldots, u_p)^T$, when present. Let $\Theta = E_F(X)$ denote the $n \times p$ matrix of means, where F denotes the joint distribution of X and $\mathbf{u}$. The objective is to estimate Θ under the $p \times p$ matrix quadratic loss function (1.3). Let $\mathbf{x}_{(j)}$ denote the jth column of X.

Suppose that for each column $j = 1, \ldots, p$, there is a shrinkage function $\mathbf{g}_{(j)} = \mathbf{g}_{(j)}(\mathbf{x}_{(j)}, u_j)$. A natural estimator is the "diagonal shrinkage estimator", defined by applying the vector shrinkage estimator separately to each column of X. That is, define $\hat{\Theta}_a = \hat{\Theta}_a(X)$ in terms of its columns $\hat{\boldsymbol{\theta}}_{a,(j)}$ by

$$\hat{\boldsymbol{\theta}}_{a,(j)} = \hat{\boldsymbol{\theta}}_a(\mathbf{x}_{(j)}, u_j) \tag{3.1}$$

using (2.1). Note that the shrinkage applied to each column does not depend on the data in other columns. We use the term "diagonal" because in the setting (2.2) the estimator can also be written in matrix form using a diagonal matrix,

$$\hat{\Theta}_a = XD, \quad D = \text{diag}(d_j), \quad d_j = 1 - a\sigma^2(n-2)/||\mathbf{x}_{(j)}||^2, \quad j = 1, \ldots, p.$$

Given two estimators $\hat{\Theta}^{(1)}$ and $\hat{\Theta}^{(2)}$ depending on X, say that $\hat{\Theta}^{(1)}$ strictly dominates $\hat{\Theta}^{(2)}$ if $R_{\text{matrix}}(\hat{\Theta}^{(1)}, \Theta) < R_{\text{matrix}}(\hat{\Theta}^{(2)}, \Theta)$ for all Θ, where "$<$" means

that the difference between the right- and left-hand sides is a positive-definite matrix. The following theorem is the main result of this paper.

Theorem 2. *Let $X(n \times p)$ be a random matrix and $\mathbf{u} = (u_1, \ldots, u_p)^T$ be a vector of auxiliary random variables such that $E_F(X) = \Theta$ under a probability model F depending on Θ, and the data $\{\mathbf{x}_{(j)}, u_j\}$ are independent for different j. Suppose there exist shrinkage functions $\mathbf{g}_{(j)} = \mathbf{g}_{(j)}(\mathbf{x}_{(j)}, u_j)$ such that the cross-product inequality (2.4) holds for each $j = 1, \ldots, p$. Then the shrinkage estimator $\hat{\Theta}_a$ in (3.1) dominates the simple estimator $\hat{\Theta}_0 = X$ under the matrix loss function (1.3) provided the tuning parameter a satisfies $0 < a < 2/p$.*

Proof. The proof makes use of the following inequality, where $\boldsymbol{\alpha}$ is a $p \times 1$ vector and G is an $n \times p$ matrix with columns $\mathbf{g}_{(j)},\ j = 1, \ldots, p$,

$$\begin{aligned} \sum_{j,k=1}^{p} \alpha_j \alpha_k \mathbf{g}_{(j)}^T \mathbf{g}_{(k)} &\le \sum_{j,k=1}^{p} |\alpha_j|\, |\alpha_k|\, ||\mathbf{g}_j||\, ||\mathbf{g}_{(k)}|| \\ &= \left\{ \sum_{j=1}^{p} |\alpha_j|\, ||\mathbf{g}_{(j)}|| \right\}^2 \\ &\le p \sum_{j=1}^{p} \alpha_j^2 ||\mathbf{g}_{(j)}||^2. \end{aligned} \tag{3.2}$$

The two inequalities follow from two versions of the Cauchy-Schwarz inequality.

To show $\hat{\Theta}_a$ dominates $\hat{\Theta}_0 = X$ for a particular choice of a, we need to show that

$$R_{\text{matrix}}(\hat{\Theta}_a, \Theta) < R_{\text{matrix}}(\hat{\Theta}_0, \Theta) \text{ for all } \Theta.$$

Equivalently we need to show that

$$\boldsymbol{\alpha}^T R \boldsymbol{\alpha} < n \boldsymbol{\alpha}^T I_n \boldsymbol{\alpha} = n \text{ for all } \Theta, \tag{3.3}$$

where $R = R_{\text{matrix}}(\hat{\Theta}_a, \Theta)$ and $\boldsymbol{\alpha}$ is an arbitrary standardized p-dimensional vector, $\boldsymbol{\alpha}^T \boldsymbol{\alpha} = 1$.

The left-hand side of (3.3) can be written as

$$\sum_{j,k=1}^{p} \alpha_j \alpha_k E_F \left\{ \left(\hat{\boldsymbol{\theta}}_{a,(j)} - \boldsymbol{\theta}_{(j)} \right)^T \left(\hat{\boldsymbol{\theta}}_{a,(k)} - \boldsymbol{\theta}_{(k)} \right) \right\} \tag{3.4}$$

$$= \sum_{j,k=1}^{p} \alpha_j \alpha_k E_F \left[\left\{ \left(\mathbf{x}_{(j)} - \boldsymbol{\theta}_{(j)} \right) - a\mathbf{g}_{(j)} \right\}^T \left\{ \left(\mathbf{x}_{(k)} - \boldsymbol{\theta}_{(k)} \right) - a\mathbf{g}_{(k)} \right\} \right] \tag{3.5}$$

$$= \sum_{j=1}^{p} \alpha_j^2 \left[E_F \left\{ \left(\mathbf{x}_{(j)} - \boldsymbol{\theta}_{(j)} \right)^T \left(\mathbf{x}_{(j)} - \boldsymbol{\theta}_{(j)} \right) \right\} - 2a\delta_j \right] + a^2 \sum_{j,k=1}^{p} \alpha_j \alpha_k E_F \left(\mathbf{g}_{(j)}^T \mathbf{g}_{(k)} \right) \tag{3.6}$$

$$\le \sum_{j=1}^{p} \alpha_j^2 \left[E_F \left\{ \left(\mathbf{x}_{(j)} - \boldsymbol{\theta}_{(j)} \right)^T \left(\mathbf{x}_{(j)} - \boldsymbol{\theta}_{(j)} \right) \right\} - 2a\delta_j + a^2 p \gamma_j \right] \tag{3.7}$$

$$\leq \sum_{j=1}^{p} \alpha_j^2 \left[E_F \left\{ \left(\mathbf{x}_{(j)} - \boldsymbol{\theta}_{(j)} \right)^T \left(\mathbf{x}_{(j)} - \boldsymbol{\theta}_{(j)} \right) \right\} - 2a\gamma_j + a^2 p \gamma_j \right] \tag{3.8}$$

$$< \boldsymbol{\alpha}^T R_0 \boldsymbol{\alpha} = n, \tag{3.9}$$

for $0 < a < 2/p$, where $\delta_j = E_F\{(\mathbf{x}_{(j)} - \boldsymbol{\theta}_{(j)})^T \mathbf{g}_{(j)}\}$ and $\gamma_j = E_F(\mathbf{g}_{(j)}^T \mathbf{g}_{(j)})$, so $\delta_j \geq \gamma_j > 0$. In going from (3.5) to (3.6) notice that many of the off-diagonal terms vanish because the different columns are independent and $E_F(\mathbf{x}_{(j)} - \boldsymbol{\theta}_{(j)}) = \mathbf{0}$. Equation (3.7) follows from (3.6) by the Cauchy-Schwarz based inequality (3.2). Equation (3.9) follows from (3.8) by simple properties of quadratic functions. □

Comments

(a) The allowable interval for a decreases with p. This property is related to the result that for a matrix loss function, it is harder to dominate the maximum likelihood estimator than for a scalar loss function.

(b) For the James-Stein case, the p-dimensional result is less powerful than the one-dimensional result. In one dimension $a = 1$ is optimal; $\hat{\boldsymbol{\theta}}_1$ dominates $\hat{\boldsymbol{\theta}}_a$ for all other choices of a. In contrast, if $p > 1$ there is no single choice of a for $\hat{\Theta}_a$ which dominates all other choices.

(c) Further, at least for the James-Stein case, the interval $(0, 2/p)$ is the best possible interval for a. If $a < 0$ or $a > 2/p$, it is possible to find values of Θ such that $\hat{\Theta}_a$ does not dominate $\hat{\Theta}_0$.
Here is a simple construction in the case of known variance $\sigma^2 = 1$. Recall $x_{ij} \sim N(\theta_{ij}, 1)$ independently for $i = 1, \ldots, n,\ j = 1, \ldots, p$. Let $\alpha_j = 1/\sqrt{p},\ j = 1, \ldots, p$. Let $\boldsymbol{\theta}^*$ be a n-vector of unit size, $\boldsymbol{\theta}^{*T}\boldsymbol{\theta}^* = 1$, and suppose all of the columns of Θ are equal to the same multiple of $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}_{(j)} = \kappa\boldsymbol{\theta}^*$. For large κ it is straightforward to show that

$$\delta_j = \gamma_j = E_F(\mathbf{g}_{(j)}^T \mathbf{g}_{(j)}) = (m^2/\kappa^2) + O(1/\kappa^4)$$

for all j, where $m = n-2$. Further (3.2) becomes an equality in this setting so that the risk in (3.3) reduces to

$$\boldsymbol{\alpha}^T R \boldsymbol{\alpha} = n - 2a(m^2/\kappa^2) + a^2(m^2/\kappa^2)p + O(1/\kappa^4). \tag{3.10}$$

Ignoring the remainder term, the quadratic function of a in (3.10) is less than n for $0 < a < 2/p$ and exceeds n for $a < 0$ or $a > 2/p$. Hence for any specific choice of $a < 0$ or $a > 2/p$, $\boldsymbol{\alpha}^T R \boldsymbol{\alpha} > n$ for sufficiently large κ.
The same argument works for the case of unknown σ^2.

(d) In the vector case, if the shrinkage function $\mathbf{g}$ is re-scaled to $c\mathbf{g}$ for some constant $c > 0$, then the cross-product inequality needs minor adjustment and the allowable interval for the tuning parameter a changes from $(0, 2)$ to $(0, 2/c)$. The scaling convention for the cross-product inequality chosen in this paper has been made to make the treatment of different columns as consistent as possible in the extension to the matrix case.

(e) Efron and Morris (1972) proposed the "matrix" James-Stein estimator

$$\hat{\boldsymbol{\theta}}^{MJS} = X\{I_p - (n-p-1)S^{-1}\}, \quad S = X^T X,$$

and investigated its properties under the scalar loss function (1.2). However, its properties under the matrix loss function (1.3) are unknown.

4. Discussion

For the classic vector James-Stein estimator there are several ingredients in the formulation of the problem and the estimator such as the following: (a) normality of the data, (b) uncorrelated components, (c) the specific choice (2.2) for the shrinkage function $\mathbf{g}$, and (d) the assumption that the range of possible values for $\boldsymbol{\theta}$ spans all of $\mathbb{R}^n$.

Each of these ingredients can be relaxed, either individually or in combination. Here are some examples.

(a) relax normality to (i) more general spherical distributions (Brandwein and Strawderman, 1991; Cellier and Fourdrinier, 1995) or (ii) independent components (Shinozaki, 1984);
(b) allow correlated normal or more general elliptic distributions (Fourdrinier, Strawderman and Wells, 2003);
(c) use other shrinkage estimators such as (i) subspace shrinkage, or more generally (ii) Bayes or generalized Bayes estimators based on superharmonic prior distributions (Stein, 1981);
(d) relax the range of possible values for $\boldsymbol{\theta}$ from all of $\mathbb{R}^n$ to a specified cone (Fourdrinier, Strawderman and Wells, 2006).

In each case the improved performance of the shrinkage estimator is justified by a version of the cross-product inequality. Hence in each case there is an immediate extension to the matrix case.

Another important direction in which the paper might be extended is to allow dependence between the columns. For simplicity limit attention to the normal case. Thus let $X(n \times p)$ follow an np-dimensional normal distribution with mean $E(X) = \Theta$, with independent rows and with common covariance matrix Σ within each row.

(a) (Known Σ) In this case, it is straightforward to adapt the results of this paper. Let B be a matrix square root of Σ^{-1}, so that $BB^T = \Sigma^{-1}$. Then $Y = XB$ has independent columns. Hence the methodology of Section 3 can be applied to Y to yield an estimator $\hat{\Phi}_a$ of $\Phi = \Theta B$. Back-transforming yields an estimator $\hat{\Theta}_a = \hat{\Phi}_a B^{-1}$ which dominates $\hat{\Theta}_0$ in the matrix sense (1.3), provided $0 < a < 2$.
(b) (Unknown Σ) When Σ needs to be estimated by an auxiliary random matrix, it is an open question whether or not there exist any shrinkage estimators which can be guaranteed to improve on the simple estimator $\hat{\Theta}_0$. Simply replacing Σ by an estimate in (a) seems to lead to intractable calculations.

(c) (Choice of basis) Even when Σ is known there is an interesting side issue. The matrix B in (a) is defined only up to a multiplication on the left by a $p \times p$ orthogonal matrix, with each choice defining a different estimator. Thus the methodology of Section 3 defines a whole family of estimators, each with the same statistical properties. It is not clear whether it might be possible to combine them in some way to yield a superior estimator.

Acknowledgement

This work was partially supported by a grant from the Simons Foundation (#209035 to William Strawderman).

References

Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist.* **41** 642–645. MR0253461 (40 ##6676)

Brandwein, A. C. and Strawderman, W. E. (1991). Generalizations of James-Stein estimators under spherical symmetry. *Ann. Statist.* **19** 1639–1650. MR1126343 (92i:62137)

Cellier, D. and Fourdrinier, D. (1995). Shrinkage estimators under spherical symmetry for the general linear model. *J. Multivariate Anal.* **52** 338–351. MR1323338 (96f:62020)

Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika* **59** 335–347. MR0334386 (48 ##12705)

Efron, B. and Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4** 22–32. MR0394960 (52 ##15759)

Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (2003). Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *J. Multivariate Anal.* **85** 24–39. MR1978175 (2004b:62060)

Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (2006). Estimation of a location parameter with restrictions of "vague information" for spherically symmetric distributions. *Ann. Inst. Statist. Math.* **58** 73–92. MR2281207 (2008e:62050)

Ghosh, M. and Shieh, G. (1991). Empirical Bayes minimax estimators of matrix normal means. *J. Multivariate Anal.* **38** 306–318. MR1131723 (92i:62107)

Haff, L. R. (1977). Minimax estimators for a multinormal precision matrix. *J. Multivariate Anal.* **7** 374–385. MR0451480 (56 ##9762)

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, Calif. MR0133191 (24 ##A3025)

Shinozaki, N. (1984). Simultaneous estimation of location parameters under quadratic loss. *Ann. Statist.* **12** 322–335. MR733517 (85k:62013)

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I* 197–206. University of California Press, Berkeley and Los Angeles. MR0084922 (18,948c)

STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR630098 (83a:62080). http://links.jstor.org/sici?sici=0090-5364(198111)9:6<1135:EOTMOA>2.0.CO;2-5&origin=MSN

TSUKUMA, H. (2009). Generalized Bayes minimax estimation of the normal mean matrix with unknown covariance matrix. *J. Multivariate Anal.* **100** 2296–2304. MR2560370 (2010m:62018)

TSUKUMA, H. and KUBOKAWA, T. (2007). Methods for improvement in estimation of a normal mean matrix. *J. Multivariate Anal.* **98** 1592–1610. MR2370109

ZHENG, Z. (1986). On estimation of matrix of normal mean. *J. Multivariate Anal.* **18** 70–82. MR827168 (87d:62110)