Estimation of a Multivariate Mean under Model Selection Uncertainty

Georges Nguefack-Tsague
University of Yaounde I
Biostatistics Unit
nguefacktsague@yahoo.fr

Abstract

Model selection uncertainty would occur if we selected a model based on one data set and subsequently applied it for statistical inferences, because the "correct" model would not be selected with certainty. When the selection and inference are based on the same dataset, some additional problems arise due to the correlation of the two stages (selection and inference). In this paper model selection uncertainty is considered and model averaging is proposed. The proposal is related to the theory of James and Stein of estimating more than three parameters from independent normal observations. We suggest that a model averaging scheme taking into account the selection procedure could be more appropriate than model selection alone. Some properties of this model averaging estimator are investigated; in particular we show using Stein's results that it is a minimax estimator and can outperform Stein-type estimators.

Keywords: James and Stein estimator, Model selection, Model averaging, Minimax, Normal multivariate mean.

1. Introduction

Stein (1956) considered the problem of estimating several parameters from independent normal observations, and showed that it was possible to uniformly improve on the maximum likelihood estimator under the total squared error risk measure in dimension three and higher. The setting relating to Stein's estimation is as follows: Let $\mathbf{X} = (X_1,\ldots, X_p)'$ have a p-dimensional multivariate normal distribution

$$\mathbf{X} \sim N_p(\mu, \sigma^2 I_p), \tag{1}$$

with unknown mean $\mu = (\mu_1,\ldots,\mu_p)'$, σ known (for simplicity) and I_p the identity covariance matrix; thus $\mathbf{X}$ can be observed from p independent experiments. An estimator $\hat{\mu} = (\hat{\mu}_1,\ldots,\hat{\mu}_p)'$ of μ is evaluated by the risk function

$$R(\hat{\mu}, \mu) = E_\mu \parallel \mu - \mu \parallel^2,$$

where $\parallel \hat{\mu} - \mu \parallel^2 = \sum_{i=1}^{p} (\hat{\mu}_i - \mu_i)^2$ and E_μ stands for averaging over the sample space with respect to the distribution (1).

Under (1), the maximum likelihood estimator (MLE) is

$$\hat{\mu}^{ML} = \mathbf{X}.$$

It is easy to show that the MLE has risk $R^{ML}(\hat{\mu}^{ML}, \mu) = p\sigma^2$.

This MLE was long though to be the "best" estimator for the multivariate mean estimation problem; i.e. it was believed that no estimator existed that achieved a lower risk. Pursuing the work of Stein (1956), James and Stein (1961) showed that if $p \geq 3$ the following estimator of the multivariate mean

$$\hat{\mu}^{JS} = (1 - \frac{(p-2)\sigma^2}{\| \mathbf{X} \|^2})\mathbf{X}$$

achieves uniformly lower risk than the MLE for all parameter values μ; i.e.

$$R^{JS}(\hat{\mu}^{JS}, \mu) < R^{ML}(\hat{\mu}^{ML}, \mu) \qquad \forall \mu.$$

The statistical community was astonished by the proof of James and Stein estimator (JSE) in 1961 (Efron and Morris 1977). Many statisticians were skeptical about JSE (mainly) because it does not share many of the nice properties of the MLE; e.g. it is nonlinear, biased and with probability density function (pdf) which cannot be expressed in a closed form (Efron and Morris 1977, Richards 1998). JSE is now well known and accepted among statisticians and econometricians (Judge and Bock 1978, Greenberg and Webster 1998, Lehman and Casella 1998, Gruber 1998). Several other drawbacks of JSE have been pointed out and many efforts have been made to improve. One major drawback is that the region of the parameter space where the risk of JSE (or some other estimators of a similar type) is significantly smaller than that of the MLE is quite limited (see Akai 1989, Stein 1981, Berger 1982).

Stein (1966) discussed Stein-type estimators for designs admitting a completely orthogonal analysis of variance. He showed that for large sample size, the Stein-type estimator applied separately to each orthogonal subspace is approximatively better than the estimator which shrinks observations towards the general average. Haff (1978) considered the estimator of normal means which are close to each other. He obtained a minimax estimator which is a modification of a Stein-type estimator shrinking observations towards the grand average. Efron and Morris (1973) considered the estimation of normal means divided in two groups with different prior variances and proposed a compromise estimator which improved the risk of Stein-type estimator. Berger and Dey (1983) did the same for k groups, leading also to an improvement of Stein-type estimator. George (1986a,b,c,d) considered situations where only conflicting or vague prior information is available; and proposed minimax estimators called multiple shrinkage Stein estimators. Akai (1989) proposed improvement over MLE when observations are classified into several groups. Since the discovery of JSE, many others shrinkage techniques have evolved. References include Efron and Morris (1975), Fay and Herriot (1979), Rubin (1981), Morris (1983), Jones (1991), Brown (2008), and Brown et al. (2011).

In many applications, several models are plausible a priori, and one of them has to be selected, to be the basis of all subsequent analysis. Overviews, explanations, discussions and examples of model selection procedures can be found in the books by Linhart and Zucchini (1986), McQuarrie and Tsai (1998), Zucchini (2000), Burnham and Anderson (2002), and Claeskens and Hjort (2008). An alternative to selecting a single model for estimation purposes is to use a weighted average of the estimates resulting from each of

the models under consideration. This leads to the class of model averaging estimators. Several options are available for specifying the weights; e.g. they can be based on the Akaike's information criterion, AIC (Akaike 1973) or Bayesian information criterion, BIC (Schwarz 1978). It is not the *construction* of the estimator that causes difficulties; the problem is to determine its *properties*. The same problem arises for estimators obtained after model selection. We refer to these estimators as *post-model selection estimators* (PMSE, Leeb and Pötscher 2005). The inferences are invalid even if different datasets are used for selection and inference because the uncertainty (variation) in the selection is ignored.

Let $M = \{M_1,\ldots,M_K\}$ be a set of K plausible models to estimate $\hat{\mu}$, the quantity of interest. Denote by $\hat{\mu}_k$ the estimator of μ_k obtained when using model M_k. Model averaging involves finding non-negative weights, $w_1,\ldots,w_K$, that sum to one, and then estimating μ by

$$\hat{\mu} = \sum_{k=1}^{K} w_k \hat{\mu}_k. \qquad (2)$$

Clearly, model selection is a special case of model averaging, with one of the weights set to unity, and all the others to zero, i.e the estimator based on a selection procedure is a mixture (0-1 weight) of the candidate estimators $\hat{\mu}_1,\ldots,\hat{\mu}_K$.

Literature on PMSEs includes, *inter alia*, Bancroft (1944) for pretest estimators, Breiman (1992), Hjorth (1994), Chatfield (1995), Draper (1995), Buckland et al. (1997), Zucchini (2000), Candolo et al. (2003), Hjort and Claeskens (2003), Efron (2004), Leeb and Pötscher (2005), Longford (2005), Claeskens and Hjort (2008), Zucchini et al. (2011), Nguefack-Tsague and Zucchini (2011), Nguefack-Tsague et al. (2011), Nguefack-Tsague (2013a,b,c), and Zhang et al. (2014).

Some model averaging weights base the weights on penalized likelihood values. Some classical weights can be seen in Buckland et al. (1997), Hjort and Claeskens (2003). Hansen (2007, 2008, 2009, 2010) and Wan et al. (2010) derived optimal weighting scheme by minimizing a Mallows' Cp criterion (Mallows 1973). Nguefack-Tsague (2014) derived optimal weights with squared error loss and showed that they may exist in theory but once estimated they are no longer optimal. Bayesian model averaging can be found in Hoeting (1999) and Wasserman (2000). Numerous applications of Akaike weights are given in Burnham and Anderson (2002).

The model selection criterion determines which model is assigned the weight one and hence used to estimate μ. The index of the selected model, k, is a random variable. We denote the selected model by M_k, and the PMSE of the quantity of interest by $\hat{\mu}_k$. Let $I()$ denote the indicator function that has the value 1 if the argument is true, and 0 if it is false. Then

$$M_k = \sum_{k=1}^{K} I(\textit{model k is selected}) M_k, \qquad \hat{\mu}_k = \sum_{k=1}^{K} I(\textit{model k is selected}) \hat{\mu}_k.$$

Clearly, the properties of $\hat{\mu}_k$ depend on (among other things) the set of candidate models, $\mathcal{M}$, and on the selection procedure, which we denote by S.

Nguefack-Tsague and Zucchini (2011) proposed a model averaging estimator in which the selection procedure is taken into account. Their proposal depends on estimators $p(M_k|S) = \Pr(M_k \textit{ is selected}|S),\ k = 1,\ldots K$ and the maximized likelihood value L_k for each model M_k. These weights are given by

$$W_k(S) = \frac{p(M_k \mid S)L_k}{\sum_{i=1}^{K} p(M_i \mid S)L_i}, \quad k = 1, 2, \ldots, K; \tag{3}$$

with its associated model averaging estimator given by

$$\hat{\mu}(S) = \sum_{k=1}^{K} W_k(S)\hat{\mu}_k. \tag{4}$$

Nguefack-Tsague and Zucchini (2011) showed that this weighting scheme dominates classical model averaging estimators and PMSEs in a simple linear regression example. The problem that needs to be solved is that of constructing estimators, $p(M_k|S)$, of the model selection probabilities. Hjort and Claeskens (2003) showed that a naive bootstrap estimator of the selection probability of model M_k (namely the proportion of resamples in which M_k is selected) does not work. If the selection probabilities depend on some parameter for which a closed form expression exists, and if one can find an estimator of the parameter, then it is possible to obtain estimators for these probabilities. For the case where there is no closed form, Miller (2002) suggested using a Monte Carlo method based on projection.

We consider the estimation of a multivariate mean when many estimators are plausible for a set of K models. Instead of selecting one of them using a specific selection criterion, we average over all these estimators by taking account this selection criterion (see Nguefack-Tsague and Zucchini 2011). Although each of the competing estimators does not necessary follow a multivariate normal distribution, it is assumed that the true model (i.e. the one that generated the data) does. It is also assumed that the model selection probabilities are computed independently from the data, for example using the Monte Carlo procedure of Miller (2002).

An example of estimators for each model M_k could be those proposed by George (1986a, page 189):

$$\hat{\mu}_{\lambda_k} = \mathbf{X} - (\mathbf{X} - \lambda_k)[\min(1, \frac{p-2}{\| \mathbf{X} - \lambda_k \|^2})] \tag{5}$$

where $\lambda_k \in \mathbf{R}^p$ is fixed, thus $\mathbf{X}$ shrinks towards a target λ_k for each model M_k. In particular if for each model M_k, $\lambda_k = \mathbf{0}$, then $\hat{\mu}_0$ is the original positive part of Stein estimator which shrinks towards $\mathbf{0}$.

This paper considers Stein's estimation problem where many models are a-priori plausible. In this situation one often uses the data to select the ``best'' model; this model is then used to make inferences, ignoring *model selection uncertainty*, i.e. the fact that the selection step and inference were carried out using the same data. We suggest that a model averaging scheme taking into account the selection procedure could be more appropriate than model selection alone. For example, instead of selecting one estimator over those given in Equation (5), we propose to average over K estimators in which the selection procedure is taken into account in the weighting scheme. In particular, this example shows that it is possible to average over Stein-type estimators (which already outperform MLE) and obtain a better estimator. Some properties of this model averaging estimator are investigated; in particular we showed using Stein's results that it is a minimax estimator and can outperformed Stein-type estimators. A Bayesian approach for estimating a multivariate mean under model uncertainty is considered in Nguefack-Tsague (2013c).

This paper is organized as follows. In Section 2 we define some concepts and the properties of the model averaging estimator. We show in Section 3 that it can improve over Stein-type estimators while Section 4 deals with a construction of confidence interval using this estimator. Our article ends with concluding remarks.

2. Definitions and properties of the model averaging estimator

2.1 Definitions

Definition 1. An estimator μ is *minimax* if

$$\sup_{\mu} R(\mu, \mu) \leq \sup_{\mu} R(\mu, \mu) \quad \textit{for any other estimator } \mu;$$

i.e. the largest risk of μ is no greater than that of any other estimator (the best worst-case scenario).

Definition 2. An estimator μ is said to *dominate* estimator μ if
$R(\mu, \mu) \leq R(\mu, \mu)$ *for all* μ *and if there exists some value* μ_* *for which the inequality is strict*.

Definition 3. An estimator μ is said to *admissible* if there is no other estimator that dominates it (otherwise it is *inadmissible*).

Definition 4. A function $h: \mathrm{R}^p \to \mathrm{R}$ is said to be *almost differentiable* if there exists a function $\nabla h: \mathrm{R}^p \to \mathrm{R}$ such that, for all z,

$$h(x+z) - h(x) = \int_0^1 z \nabla h(x+tz) dt$$

for almost all $x \in \mathrm{R}^p$; where ∇ is the vector differential operator of first partial derivatives with ith coordinate $\nabla_i = \frac{\partial}{\partial x_i}$. A function $g: \mathrm{R}^p \to \mathrm{R}^p$ is *almost differentiable* if all its coordinate functions are.

Definition 5. A lower semicontinous function $f:\mathrm{R}^p \to \mathrm{R} \cup +\{\infty\}$ is *superharmonic* at point $x_0 \in \mathrm{R}^p$ if for every $r > 0$, the average of f over the sphere

$$S_r(x_0) = \{x : \| x - x_0 \|^2\} = r^2$$

of radius r centered at x_0 is not greater than $f(x_0)$.

Definition 6. A twice continuously differentiable function $f:\mathrm{R}^p \to \mathrm{R}$ is harmonic at $x_0 \in \mathrm{R}^p$ if $\nabla^2 f(x_0) = 0$; where $\nabla^2 = \sum_{i=1}^{p} \nabla_i^2$ is the Laplacian.

2.2 Properties of the model averaging estimator

Let $H:\mathrm{R}^p \to \mathrm{R}$ be defined by $H(\mathbf{X} \mid S, M) = \sum_{k=1}^{K} p(M_k \mid S) L_k$ (denominator of the model averaging weight defined in Equation (3). In addition, suppose that for each model M_k, the estimator $\hat{\mu}_k$ is of the form $\mathbf{X} + \nabla \log L_k ; \forall M_k \in M$, where $\nabla \log L_k = (\frac{\partial \log L_k}{\partial x_1}, \ldots, \frac{\partial \log L_k}{\partial x_p})'$. The estimator of the form "$\mathbf{X} + \nabla \log L_k$" is motivated by Stein (1981) who explains that small risk may be obtained by such estimator (see also George (1986a, page 190)). We write down in the following the risk of the proposed model averaging estimator given in Equation (4) and prove that it is a minimax estimator for μ.

Assumptions

A.1. $\mathbf{X} \sim N_p(\mu, \sigma^2 I_p)$, σ known.

A.2. H is almost differentiable for which ∇H is also almost differentiable.

A.3. H is superharmonic.

A.4. $E_\mu \mid \frac{\partial^2 H(\mathbf{X} \mid S, M) / \partial X_i^2}{H(\mathbf{X} \mid S, M)} \mid < \infty$, for $i = 1, \ldots, p$.

A.5. $E_\mu \| \nabla \log H(\mathbf{X} \mid S, M) \|^2 < \infty$.

Lemma 1. The model averaging estimator in Equation (4) becomes

$$\hat{\mu}(S) = \mathbf{X} + \nabla \log H(\mathbf{X} \mid S, M). \qquad (6)$$

Proof. Denote $\nabla f = (\frac{\partial f}{\partial x_1},\ldots,\frac{\partial f}{\partial x_p})'$ and $\nabla Logf = \frac{\nabla f}{f}$.

$$\hat{\mu}(S) = \sum_{k=1}^{K} W_k(S)\hat{\mu}_k = \sum_{k=1}^{K}\{\frac{p(M_k \mid S)L_k}{H(\mathbf{X} \mid S,M)}\}\hat{\mu}_k$$

$$= \frac{1}{H(\mathbf{X} \mid S,M)}\sum_{k=1}^{K} p(M_k \mid S)L_k[X + \nabla \log L_k]$$

$$= \mathbf{X} + \frac{1}{H(\mathbf{X} \mid S,M)}\sum_{k=1}^{K} p(M_k \mid S)L_k[\frac{\nabla L_k}{L_k}]$$

$$= \mathbf{X} + \frac{1}{H(\mathbf{X} \mid S,M)}\sum_{k=1}^{K} p(M_k \mid S)\nabla L_k$$

$$= \mathbf{X} + \frac{1}{H(\mathbf{X} \mid S,M)}\nabla(\sum_{k=1}^{K} p(M_k \mid S)L_k)$$

$$= \mathbf{X} + \frac{1}{H(\mathbf{X} \mid S,M)}\nabla H(\mathbf{X} \mid S,M) = \mathbf{X} + \nabla \log H(\mathbf{X} \mid S,M)$$

Lemma 2 (Lemma 1 of Stein (1981, page 1136)). Let Y be a $N(0,1)$ real random variable and let $g : \mathrm{R} \to \mathrm{R}$ be an indefinite integral of the Lebesgue measurable function g', essentially the derivative of g. If $E \mid g'(Y) \mid < \infty$, then $E\{g'(Y)\} = E\{Yg(Y)\}$.

Lemma 3 (Lemma 2 of Stein (1981, page 1137)). If $h : \mathrm{R}^p \to \mathrm{R}$ is an almost differentiable function with $E_\mu \parallel \nabla h(\mathbf{X}) \parallel < \infty$, then $E_\mu \nabla h(\mathbf{X}) = E_\mu\{(\mathbf{X} - \mu)h(\mathbf{X})\}$.

Lemma 4 (Theorem 4.8 of Helms (1969, page 63)). If $f : \mathrm{R}^p \to \mathrm{R}$ is twice continuous differentiable, then f is superharmonic in R^p if and only if, for all $\mathbf{X} \in \mathrm{R}^p$, $\nabla^2 f(\mathbf{X}) \le 0$.

Lemma 5 (Theorem 1 of Stein (1981, page 1138)). Consider an estimator $\mathbf{X} + g(\mathbf{X})$ for μ such that $g : \mathrm{R}^p \to \mathrm{R}^p$ is an almost differentiable function for which

$E_\mu\{\sum_{i=1}^{p} \mid \nabla_i g_i(\mathbf{X}) \mid\} < 0$, then for each $i \in \{1,\ldots,p\}$,

$$E_\mu\{X_i + g_i(\mathbf{X}) - \mu_i\}^2 = 1 + E_\mu\{g_i^2(\mathbf{X}) + 2\nabla_i g_i(\mathbf{X})\},$$

and consequently

$$E_\mu \parallel \mathbf{X} + g(\mathbf{X}) - \mu \parallel^2 = p + E_\mu\{\parallel g(\mathbf{X}) \parallel^2 + 2\nabla . g(\mathbf{X})\}. \qquad (7)$$

Theorem 1. Under the assumptions A.1-A.5, $\hat{\mu}(S)$ has risk

$$R(\hat{\mu}(S),\mu) = E_\mu \parallel \hat{\mu}(S) - \mu \parallel^2 = p + 4E_\mu[\frac{\nabla^2\sqrt{H(\mathbf{X} \mid S,M)}}{\sqrt{H(\mathbf{X} \mid S,M)}}]$$

and is a minimax estimator for μ.

Proof. Let $g : \mathrm{R}^p \to \mathrm{R}^p$ be defined by

$$g = \nabla \log H = \frac{\Delta H}{H}. \tag{8}$$

Then $\nabla.g = \nabla.\nabla \log H = \frac{\nabla^2 H}{H} - \frac{\| \nabla H \|^2}{H^2}$, it follows from Equation (7) that

Note that $\nabla^2 \sqrt{H(\mathbf{X}|S,M)} = \nabla.\nabla \sqrt{H(\mathbf{X}|S,M)} = \nabla. \frac{\nabla H(\mathbf{X}|S,M)}{2\sqrt{H(\mathbf{X}|S,M)}}$

$$= \frac{\nabla^2 H(\mathbf{X}|S,M)}{2\sqrt{H(\mathbf{X}|S,M)}} - \frac{\| \nabla H(\mathbf{X}|S,M) \|^2}{4[\sqrt{H(\mathbf{X}|S,M)}]H(\mathbf{X}|S,M)}. \tag{9}$$

Thus $4\nabla^2 \sqrt{H(\mathbf{X}|S,M)} = \frac{2\nabla^2 H(\mathbf{X}|S,M)}{\sqrt{H(\mathbf{X}|S,M)}} - \frac{\| \nabla H(\mathbf{X}|S,M) \|^2}{[\sqrt{H(\mathbf{X}|S,M)}]H(\mathbf{X}|S,M)}$

$$= \frac{2\sqrt{H(\mathbf{X}|S,M)}\nabla^2 H(\mathbf{X}|S,M)}{H(\mathbf{X}|S,M)} - \frac{\sqrt{H(\mathbf{X}|S,M)}\| \nabla H(\mathbf{X}|S,M) \|^2}{H(\mathbf{X}|S,M)^2}.$$

It follows that $\frac{4\nabla^2 \sqrt{H(\mathbf{X}|S,M)}}{\sqrt{H(\mathbf{X}|S,M)}} = \frac{2\nabla^2 H(\mathbf{X}|S,M)}{H(\mathbf{X}|S,M)} - \frac{\| \nabla H(\mathbf{X}|S,M) \|^2}{H(\mathbf{X}|S,M)^2}$.

Thus $R(\hat{\mu}(S), \mu) = p + 4E_\mu [\frac{\nabla^2 \sqrt{H(\mathbf{X}|S,M)}}{\sqrt{H(\mathbf{X}|S,M)}}]$.

From Equation (9), $\nabla^2 \sqrt{H(\mathbf{X}|S,M)} \le \frac{\nabla^2 H(\mathbf{X}|S,M)}{2\sqrt{H(\mathbf{X}|S,M)}} \le 0$ since H is superharmonic and by Lemma 4, $\nabla^2 H(\mathbf{X}|S,M) \le 0$.

Since $\mathbf{X}$ is minimax for μ with risk p,

it then follows that $R(\hat{\mu}(S), \mu) \le p = \inf_g \sup_\mu E_\mu \| \mathbf{X} + g(\mathbf{X}) - \mu \|$

3. Improvement over James-Stein estimator

Efron and Morris (1971, 1972) propose the following modification of the James-Stein estimator

$$\hat{\mu}^{JSM} = \mathbf{X}(1 - \frac{(p-2)\sigma^2}{\| \mathbf{X} \|^2})_+$$

where $a_+ = aI(a \ge 0)$.

This modification was based on requiring that no coordinate be changed by more than a predetermined quantity d. This resulted in an improvement of $\hat{\mu}^{JSM}$ when the empirical

distribution of $|\hat{\mu}_i|$ is skewed. We now also consider a modification of Efron and Morris based on order statistic.

Let $Y_i = |X_i|$ and the order statistics defined by $Y_{(1)} < \ldots < Y_{(p)}$. Let $j \le p$ be a positive integer. Suppose also that the coordinates of $g(\mathbf{X}|S,M)$ in Equation (8) as $g = \nabla \log H$ are now defined as

$$g_i(\mathbf{X}|S,M) = \begin{cases} -c[\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1} X_i & \text{if } Y_i \le Y_{(j)} \\ -c[\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1} X_i \operatorname{sign} X_i & \text{otherwise} \end{cases} \quad (10)$$

where c is a constant. Let $\tilde{\mu}(S)$ the corresponding model averaging estimator from Equation (6) with g defined as in Equation (10).

Theorem 2. The risk of $\tilde{\mu}(S)$ is given by

$$R(\tilde{\mu}(S),\mu) = E_\mu \| \tilde{\mu}(S) - \mu \|^2 = p + (c^2 - 2(j-2)c) E_\mu [\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1}]. \quad (11)$$

Proof: Using equation (7) of Lemma 5,

$$\begin{aligned} R(\tilde{\mu}(S),\mu) &= p + E_\mu[c^2[\textstyle\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1}] - 2c\sum_{i=1}^{j}([\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1}) \\ &\quad + 4c\sum_{i=1}^{j-1} X_i^2[\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-2} + 4c(p-j+1)Y_{(j)}^2[\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-2}] \\ &= p + (c^2 - 2(j-2)c) E_\mu[\textstyle\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1}] \end{aligned}$$

Corollary 1. The optimum choice of c is $c^* = j-2$ with its associated risk given by

$$R(\tilde{\mu}(S),\mu) = E_\mu \| \tilde{\mu}(S) - \mu \|^2 = p - (j-2)^2 E_\mu [\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1}]. \quad (12)$$

Proof: Taking the derivative of the risk in Equation (11) with respect to c and equating to 0 yields $c^* = j-2$; replacing c by $j-2$ in Equation (11) yields the risk in Equation (12)

Assume that p is large ($p \to \infty$) and $z = \frac{j}{p}$, so that j is close to p; and that $\tilde{\mu}_i(S)$ are independently normally distributed with variance δ^2. The estimated improvement risk for $\tilde{\mu}(S)$ and $\hat{\mu}^{JSM}$ over the MLE $\hat{\mu}^{ML}$ are denoted respectively by $Imp(\tilde{\mu}(S), \hat{\mu}^{ML})$ and $Imp(\hat{\mu}^{JSM}, \hat{\mu}^{ML})$:

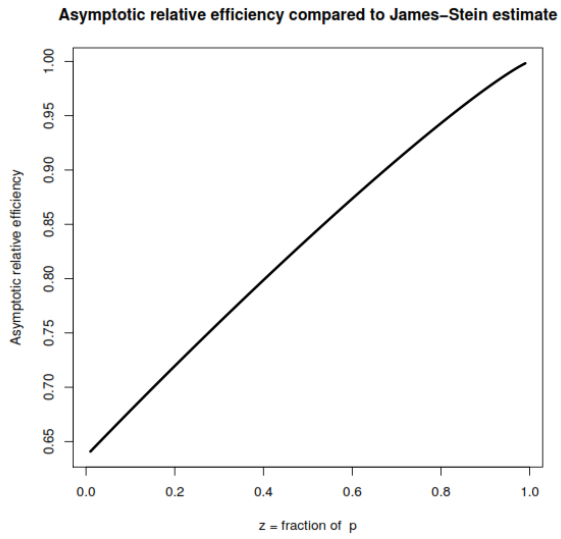
$$\begin{aligned} &Imp(\tilde{\mu}(S), \hat{\mu}^{ML}) = \hat{R}(\hat{\mu}^{ML}, \mu) - \hat{R}(\tilde{\mu}(S), \mu) = (j-2)^2[\textstyle\sum_{l=1}^{p}(\min(X_l^2, Y_{(j)}^2))]^{-1} \, Imp(\hat{\mu}^{JSM}, \hat{\mu}^{ML}) \\ &= \hat{R}(\hat{\mu}^{ML}, \mu) - \hat{R}(\hat{\mu}^{JSM}, \mu) = (p-2)^2[\textstyle\sum_{l=1}^{p} X_l^2]^{-1}. \end{aligned}$$

The asymptotic relative efficiency (Stein, 1981; page 1146) of $\tilde{\mu}(S)$ compared to the modified James-Stein estimate $\hat{\mu}^{JSM}$ is defined by

$$eff(z) = \frac{Imp(\hat{\mu}^{JSM}, \hat{\mu}^{ML})}{Imp(\tilde{\mu}(S), \hat{\mu}^{ML})} = \frac{z^2}{(1-z)w^2 - 2w\phi(w) + z}, \quad (13)$$

where $w = \Phi^{-1}(0.5(1+z))$, with ϕ and Φ respectively as density and distribution functions of standard normal.

Figure (1) shows that the relative efficiency is an increasing function of z, i.e. as the proportion of data increases, the relative efficiency also increases. It also shows that in Equation(13), eff $(z) < 1, \forall z \in (0,1)$; thus $\hat{R}(\tilde{\mu}(S), \mu) < \hat{R}(\hat{\mu}^{JSM}, \mu)$.



Asymptotic relative efficiency of $\tilde{\mu}(S)$ compared to the modified James-Stein estimate $\hat{\mu}^{JSM}$ as a function of the proportion of the dimension of the parameter μ.

This means that the modified version of the weighted estimator (model averaging) given in Equation (10) is better than Jame-Stein estimator, for any proportion z of the data. When $z = 1$, that is $p = j$, eff(1)=1, then both estimators have the same improvement over the maximum likelihood estimator.

4. Confidence sets for the mean

Here we illustrate how to obtain an approximate confidence sets for the parameter μ.

Lemma 6. (Theorem 3 of Stein (1981, page 1149)). Let $\mathbf{X}$ be a random p-dimensional coordinate vector, normally distributed with mean μ and the identity as covariance matrix. Let $g : \mathrm{R}^p \to \mathrm{R}^p$ be a twice continuously differentiable function such that

$$E_\mu\{\| g(\mathbf{X}) \|^2 + \sum_{i,j} g_{ij}^2(\mathbf{X}) + \sum_{i,j} g_{iij}^2(\mathbf{X})\} < \infty, \quad (14)$$

where $g_{ij} = \nabla_j g_i$ and $g_{iij} = \nabla_i \nabla_j g_i$. Then

$$E_\mu[\| \mathbf{X} + g(\mathbf{X}) - \mu \|^2 - U(\mathbf{X})]^2 = V(\mathbf{X}) \tag{15}$$

where

$$U(\mathbf{X}) = p + \| g(\mathbf{X}) \|^2 + 2\nabla' g(\mathbf{X}) \text{ and}$$

$$V(\mathbf{X}) = 2p + 4E_\mu[\| g(\mathbf{X}) \|^2 + 2\nabla' g(\mathbf{X}) + tr\{\nabla g'(\mathbf{X})\}^2],$$

and g' denotes the vector-value function whose value is the transpose of the value of the function g.

Theorem 3. Let $g(\mathbf{X} \mid S, M) = \nabla \log H(\mathbf{X} \mid S, M)$. Under ((14)) with $p \to \infty$,

$$E_\mu[\| \hat{\mu}(S) - U(\mathbf{X} \mid S, M)]^2 = 2p + 4E_\mu[U(\mathbf{X} \mid S, M)], \tag{16}$$

where $U(\mathbf{X} \mid S, M) = p + \| g(\mathbf{X} \mid S, M) \|^2 + 2\nabla' g(\mathbf{X} \mid S, M)$ and

$$V(\mathbf{X} \mid S, M) = 2p + 4E_\mu[\| g(\mathbf{X} \mid S, M) \|^2 + 2\nabla' g(\mathbf{X} \mid S, M) + tr\{\nabla g'(\mathbf{X} \mid S, M)\}^2]$$

Proof. This is a straightforward application of Lemma 6 with $g(\mathbf{X} \mid S, M) = \nabla \log H(\mathbf{X} \mid S, M)$.

As noted by Stein (1981, page 1150),

$E_\mu[\| \hat{\mu}(S) - \mu \|^2 - \{p + \| g(\mathbf{X} \mid S, M) \|^2 + 2\nabla' g(\mathbf{X} \mid S, M)\}]^2$ is approximately normally distributed with mean 0 and that $2p + 4E_\mu[\| g(\mathbf{X} \mid S, M) \|^2 + 2\nabla' g(\mathbf{X} \mid S, M) + tr\{\nabla g'(\mathbf{X} \mid S, M)\}^2]$ is approximately constant.

Thus a confidence sets for μ with approximately $1-\alpha$ of covering μ can be given by $CS = \{\mu : \| \mu - \hat{\mu}(S) \|^2 < U(\mathbf{X} \mid S, M) + Z_{1-\alpha}\sqrt{W(\mathbf{X} \mid S, M)}\}$,

where

$$W(\mathbf{X} \mid S, M) = 2p + 4[\| g(\mathbf{X} \mid S, M) \|^2 + 2\nabla' g(\mathbf{X} \mid S, M) + tr\{\nabla g'(\mathbf{X} \mid S, M)\}^2]$$

5. Concluding Remarks

In this paper, we have considered the estimation of a multivariate mean when many estimators are plausible for a set of models. Instead of selecting one of them using a specific selection criterion, we average over all these models by taking account this selection criterion. We have derived some properties of this estimator and showed that it can outperform Stein-type estimators. This method is particularly useful when prior information suggests that several choices of the estimation of the multivariate mean are feasible.

6. Acknowledgments

We thank the Editor-in-Chief and three anonymous reviewers whose constructive comments and suggestions have led to an improvement on earlier versions of this paper.

References

1. Akai, T. (1989). Simultaneous estimation of means of classified normal observations. *Annals of the Institute of Statistical Mathematics*, 41, 485–502.
2. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory, eds. B. Petrovand F. Csáki, Budapest: *Akadémiai Kiadó*, 267–281.
3. Bancroft, T.A. (1944). On bias in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics*, 15, 190–204.
4. Berger, J. (1982). Selecting a minimax estimator of a multivariate normal mean. *Annals of Statistics*, 10, 81–92.
5. Berger, J. and Dey D. (1983). Combining coordinates in simultaneous estimation of normal means. *Journal of Statistical Planning and Inference*, 8, 143–160.
6. Breiman, L. (1992). The little boot strap and other methods for dimensionality selection in regression: X-Fixed predictor error. *Journal of the American Statistical Association*, 87, 738–754.
7. Brown, L.D. (2008). In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *Annals of Applied Statistics*, 2, 113–152.
8. Brown, L. D., Nie, H. and Xie, X. (2011). Ensemble minimax estimation for multivariate normal mean. Submitted to *Annals of Statistics*.
9. Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997). Model selection: An integral part of inference. *Biometrics*, 53, 603–618.
10. Burnham, P.K. and Anderson, D.R. (2002). *Model Selection and Multi model Inference, a practical Information-Theoretic Approach.* 2nd Edition. Springer-Verlag: New York.
11. Candolo, C., Davison, A.C. and Demétrio, C.G.B. (2003). A note on model uncertainty in linear regression. *The Statistician*, 158, 165–177.
12. Chatfied, C. (1995). Model Uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, series B*, 158, 419–466.
13. Claeskens, G. and Hjort, N.L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98, 900–916.
14. Claeskens, G. and Hjort, N.L. (2008). *Model selection and model averaging*, Cambridge University Press, Cambridge.
15. Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, series B* 57, 45–97.
16. Efron, B. (2004). The estimation of prediction error: covariance penalties and cross- validation. *Journal of the American Statistical Association*, 99, 619–642.

17. Efron, B. and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators- Part I: The Bayes case. *Journal of the American Statistical Association,* 66, 807–815.

18. Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators- Part I: *The empirical Bayes case. Journal of the American Statistical Association,* 67, 130–139.

19. Efron, B. and Morris, C. (1973). Combining possibly related estimation problem. *Journal of the Royal Statistical Society, series B* 35, 379–421.

20. Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association,* 70, 311–319.

21. Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236 (5), 119–127.

22. Fay, R. E. III and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association,* 74, 269–277.

23. George, E.I. (1986a). Minimax multiple shrinkage estimation. *Annals of Statistics*, 14, 188–205.

24. George, E. I. (1986b). Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, 81, 437–445.

25. George, E.I. (1986c). A formal Bayes multiple shrinkage estimator. Communications in Statistics. *A-Theory and Methods*, 15, 2099–2114.

26. George, E.I. (1986d). *Multiple shrinkage generalizations of the James-Stein estimator*. *Contributions to the Theory and Applications of Statistics*, Academic Press, New York.

27. Greenberg, E. and Webster, C.E. (1998). *Advanced econometrics: a bridge to the literature*, John Wiley and Sons, New York.

28. Gruber, M.H.J. (1998). *Improving efficiency by shrinkage: the James-Stein and Ridge regression estimators*, Marcel Dekker: New York.

29. Haff, L.R. (1978). The multivariate normal mean with intra class correlated components: estimation of urban fire alarm probabilities. *Journal of the American Statistical Association*, 60, 806–825.

30. Hansen, B.E. (2007). Least squares model averaging. *Econometrica,* 75, 1175–1189.

31. Hansen, B.E. (2008). Least squares forecast averaging. *Journal of Econometrics*, 146, 342–350.

32. Hansen, B.E. (2009). Averaging estimators for regressions with a possible structural break. *Econometric Theory,* 35, 1498–1514.

33. Hansen, B.E. (2010). Averaging estimators for auto regressions with an ear unit root. *Journal of Econometrics,* 158, 142–155.

34. Helms, L. (1969). *Introduction to potential theory,* John Wiley and Sons, New York.

35. Hjort, N.L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98, 879–899.

36. Hjorth, J. (1994). *Computer intensive statistical methods: Validation, model selection and bootstrap*, Chapman and Hall, London.
37. Hoeting, J., Madigan D., Raftery A. and Volinsky C. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 4, 382–417.
38. Jones, K. (1991). Specifying and estimating multilevel models for geographical research. *Transactions of the institute of British geographers*, 16, 148–159.
39. James, W. and Stein, C. (1961). *Estimation with quadratic loss*. In Proceedings for the Third Berkeley Symposium on Mathematical Statistics and Probability 1, 361–379. University of California Press.
40. Judge, G. and Bock, M. (1978). *The statistical implications of pre-test and Stein rule estimators in Econometrics*, North Holland, Amsterdam.
41. Leeb, H. and Potscher, B.M. (2005). Model selection and inference: Fact and fiction. *Econometric Theory*, 21, 21–59.
42. Lehmann E.L. and Casella G. (1998). *Theory of point estimation*. Springer (Second Edition), New York.
43. Linhart, H. and Zucchini, W. (1986). *Model selection*. John Wiley and Sons, New York.
44. Longford, N.T. (2005). Editorial: Model selection and efficiency-is' which model...?' The right question? *J.R. Statist. Soc. A*, 168, Part3, 469–472.
45. Mallows, C.L. (1973). Some comment son Cp. *Technometrics*, 15, 661–675.
46. Mc Quarrie, A.D.R. and Tsai, C.L. (1998). *Regression and time series model selection*, World Scientific, Singapore.
47. Miller, A.J. (2002). *Subset selection in regression*, 2nd Edition, Chapman and Hall, London.
48. Morris, C. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78, 47–55.
49. Nguefack-Tsague G. (2014). On optimal weighting scheme in model averaging. *American Journal of Applied Mathematics and Statistics*, 2(3), 150–156.
50. Nguefack-Tsague G. (2013a). On boot strap and post-model selection inference. *International Journal of Mathematics and Computation*, 21 (4), 51–64.
51. Nguefack-Tsague G. (2013b). An alternative derivation of some commons distributions functions: A post-model selection approach. *International Journal of Applied Mathematics and Statistics*, 42(12), 138–147.
52. Nguefack-Tsague, G. (2013c). Bayesian estimation of a multivariate mean under model uncertainty. *International Journal of Mathematics and Statistics*, 13, 83–92.
53. Nguefack-Tsague, G. and Zucchini, W. (2011). Post-model selection inference and model averaging. *Pakistan Journal of Statistics and Operation Research*, 7 (2-Sp), 347–361.
54. Nguefack-Tsague, G., Zucchini, W. and Fotso S. (2011). On correcting the effects of model selection on inference in linear regression. *Syllabus Review (Science Series)*, 2(3), 122–140.

55. Richards, J. A. (1999). An Introduction to James-Stein Estimation, M.I.T. EECS Area Exam Report.

56. Rubin, D. (1981). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association*, 75, 801–827.

57. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

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

59. Stein, C.M. (1966). *An approach to the recovery of inter-block information in balanced incomplete block designs*. Festschrift (ed. F. N. David), 351-366, John Wiley and Sons, New York.

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

61. Wan, A.T. K., Zhang X. and Zou G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156(2), 277–283.

62. Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44, 92–107.

63. Zhang X., Zou G. and Liang H. (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika*, 101(1), 205–218.

64. Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, 4, 41–61.

65. Zucchini, W., Claeskens, G. and Nguefack-Tsague, G. (2011). *Model Selection*. In International Encyclopaedia of Statistical Sciences, Part 13, 830–833, DOI: 10. 1007/978-3-642-04898-2373, Editor: M. Lovric, Springer.