ANR ClasSel
Livrable 2.1
Sélection de modèle : Etat de l'art

Dominique Fourdrinier, Martin T. Wells

13 septembre 2010

Résumé

Ce livrable présente un état de l'art sur la sélection de modèle dans une perspective décisionnelle au travers de l'approche "estimation de coût". Dans un premier document, nous enchaînons notre présentation de cet état l'art avec une première mise en oeuvre, à titre d'essai, de l'estimation de coût comme sélecteur de variables d'un modèle de régression linéaire, au travers de l'estimation des coefficients de régression par l'estimateur des moindres carrés et ce, dans un cadre distributionnel qui des lois à symétrie sphérique. Un papier en collaboration avec M.T. Wells en a résulté (1).

Chapitre 1

Risk comparisons of variable selection rules

RISK COMPARISONS OF VARIABLE SELECTION RULES

Dominique Fourdrinier* and Martin T. Wells †

Abstract

A fundamental statistical principle is that of parsimonious modeling, that is, simple models are preferred to complicated ones. A common approach is to formulate the variable selection issue as one of estimation of prediction error. One wishes to choose the submodel which minimize the prediction error sum of squares. The problem with this procedure is that the prediction error sum of squares depends on unknown parameters. Therefore one constructs selection procedures based on estimates of the prediction error sum of squares and select the submodel having minimum estimated prediction error. The best submodel in the sequence of all subsets of models is defined as the one with the minimum value of the estimated prediction error sum of squares. In this article we examine the properties of various families of variable selection rules. With our formulation, we gain insight into some of the classical selection rules, while also proposing a new class of subset selection procedures. In all of our calculations we consider coordinate free linear models with spherically symmetric error terms.

KEYWORDS: Best subset, conditional inference, loss estimation, robustness, spherical symmetry, subset selection, variable selection

AMS 1991 subject classifications: 62C99, 62F10, 62H99, 62J05, 62J04

*Université de Rouen, UPRES-A CNRS 6085, 76821 Mont Saint Aignan Cedex, France.

†Cornell University, Department of Social Statistics, 358 Ives Hall, Ithaca, NY 14851-0952, USA. The support of NSF Grant DMS 9625440 is gratefully acknowledged.

1 Introduction

Consider the linear model $y_i = \theta_i + \varepsilon_i$ $1 \leq i \leq n$ where $\theta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_\Pi X_{i\Pi}$, the $X_{i1}, \ldots, X_{i\Pi}$ are fixed regressors and the ε_i's are stochastic error terms. In linear models with many independent variables, one is often confronted with selecting a subset of the predictors. There are many approaches to this problem, most of which have been implemented in the major statistical packages. Most of the approaches to date are ad hoc and have no basis for their continued use. In this paper, we examine some of the well known dimensionality and subset selection procedures in a decision theoretic framework. We will also propose a new procedure. A decision theoretic examination has practical importance in that one would like to have analytic evidence for the goodness of selection procedure, as opposed to only simulation type evidence ($cf.$ Roecker, 1991). It is surprising that there has been very little work on the theoretical properties of the various selection procedures. The notable exceptions are the article by George and Foster (1994), who examine the risk inflation of variable selection rules, and Efron (1986), who gives a study of the bias properties of various variable selection procedures and finds that some of that classical rules are less biased than others.

A common approach is to formulate the variable selection problem as an estimation prediction error problem. As pointed out by Efron and Tibshirani (1993, §17), "Prediction error is a different quantity that measures how well a model predicts the response value of a future observation. It is often used for model selection, since it is sensible to choose a model that has the lowest prediction error among a set of candidates." Specifically, we wish to choose the subset of covariates which minimizes the prediction error sum of squares *(PES)*. The problem with this procedure is that the *PES* depends on unknown parameters. Therefore one constructs estimates of the *PES* and select the submodel having minimum estimated prediction error. We define the best submodel in the sequence of all subsets of models as the one with the minimum value of the estimated *PES*. This is in the spirit of the approach due to Breiman (1992), Efron and Tibshirani (1993, §17), and Shao and Tu (1995, §7.4); however they all consider the mean *PES (MPES)*. Berger and Pericchi's (1993) intrinsic Bayes factor criteria entails selecting the model that has the largest intrinsic Bayes factor, a rationale quite similar to choosing the subset of covariates which minimize the prediction error sum of squares. As the approach is Bayes there is no direct estimation but just marginalization against a particular prior distribution.

The idea of applying *PES* and *MPES* estimation is routinely applied in nonparametric estimation. The problematic choice of selecting the smoothing parameter is typically formulated as the choosing the degree of smoothness tht minimizes an error estimate. In the context of linear estimators (kernel, spline, and local linear estimators) for nonparametric regression Hurrich et al. (1998) propose an improved version of Akaike's error estimate to select the smoothing parameter. Hurrich et al. (1998) show that the improved estimate

of error yields an improved smoothing parameter via simulation. Donoho and Johnstone (199x, 199x) and Donho et al. (199x) propose unbiased estimators of error and select their smoothing parameter that minimizes this error estimate. In various numerous contexts cross-validation is used to estimate the error of an estimator and the turning parmateric is selected to minimize this error estimate.

Both theoretical and applied statisticians have different ideas what makes a good selection rule. There is one camp of researchers who believe that good rules are made up of *complexity and goodness-of-fit* components. The typical form for these rules are

$$SS_p(\text{Residual}) + \beta(p) \tag{1}$$

where $p \leq \Pi$ is the dimension of the submodel being fit, based on n observations, and where $\beta(p)$ represents a penalty for over-fitting. The well known Mallows' (1973) C_p criterion and Akaike (1970) information criterion *(AIC)* correspond to (1) with $\beta(p) = 2p$. Schwartz's (1978) Bayesian information criterion *(BIC)* corresponds to $\beta(p) = p \log n$. A procedure proposed by Foster and George (1994) takes $\beta(p) = p \log p$ which is essentially the asymptotic form of a squared t-statistic (see (6) below for a finite sample analog). The final prediction error *(FPE)* criterion proposed by Rissanen (1986) and studied by Wei (1992) can also be written in the form of (1) with $\beta(p) = \psi p$ for some $\psi > 0$. The choice $\psi = (2n - d)/(n - d)$ yields a selection procedure asymptotically equivalent to *d-fold* cross validation, see Zhang (1993) for more details. George and Foster (1997) recently proposed a data dependent penalty function (see (6) below) based on an empirical Bayes objective.

It is interesting to note that C_p, *AIC*, and *BIC* were not originally motivated by the goodness-of-fit plus complexity criteria. *AIC* compares the approximate likelihood of a given model to a base model. While *BIC* calculates the posterior probability of model at hand. Mallows' C_p was originally derived as an estimate of the mean squared prediction error. It appears that the complexity and the goodness-of-fit form intuition of the selection rules is only an artifact. Hence one should not base a theory of variable selection solely on the form of the selection rules. The theoretical construction of the notion of goodness of the overall model is clearly the primary concern.

Another class of well known criteria functions are of the form

$$\alpha(p) SS_p(\text{Residual}), \tag{2}$$

for some $\alpha(p) > 0$, which represents a penalty for over and under-fitting. Craven and Wahba's (1979) generalized cross validation *(GCV)* takes $\alpha(p) = n(n - p)^{-2}$. This happens to take a form almost identical to another procedure S_p proposed in Hocking (1976) and Thompson (1978) which sets $\alpha(p) = (n - 1)(n - p)^{-1}(n - p - 1)^{-1}$. S_p was motivated

by treating the response variable and the explanatory variables jointly as a multivariate normal random variable in a prediction problem. Further aspects of S_p are explored in Breiman and Freedman (1983). It can be shown that maximizing the adjusted R^2-statistic is identical to minimizing (2) with $\alpha(p) = n(n-p)^{-1}$. Lastly, Miller (1990) shows that the well known $PRESS_p$ statistic proposed by Allen (1971, 1974) can be approximated by (2) with a $\alpha(p) = (n-2)(n-p)^{-2}$. Furthermore, Berger and Pericchi's (1993) intrinsic Bayes factor criteria is essentially equal to a power of SS_p(Residual) times a multiple that depends on the design matrix.

Many of the measures defined by (1) and (2) have an *ad hoc* motivation and their application to model selection may be somewhat suspect. In this article we try to construct a theory for which these seemingly *ad hoc* measures have some intrinsic meaning. Some comparisons between the different selection procedures via an examination of their associated predictive risk functions are made. George and Foster (1997) recognized the benefits using a risk function analysis, in their simulation study they compare the predictive risks of various selection rules. We will see that the selection criteria of the form of (2) are usually better than these of the form of (1). However, we will construct a class of procedures which dominates the best in the class generated by the form in (2).

Efron and Tibsharani (1993, §17) point out that it is not clear which of the standard variable selection methods is best: "The methods are asymptotically the same, but can behave quite differently in small samples." There have been results on the asymptotic optimality of various selection procedures, *cf.* Li (1987). A problem with the asymptotic results is that there are a variety of optimal procedures, and hence, there is an issue of which "optimal" rule should be used. Furthermore it seems that any selection rule that has an infinite number of information sources to choose a finite number of parameter ought to do quite well.

We find it more intuitive to consider the data at hand and not the unseen possible realizations. This is quite similar to the debate in the smoothing literature on the issue of using the mean integrated squared error *(MISE)* and the integrated squared error *(ISE)*. In nonparameteric function estimation there is also a smoothing parameter that is selected on the basis of minimization of an integrated squared error (ISE) or mean ISE (MISE). The ISE is the infinite dimensional version of the PES. When one considers the *MPES*, one is averaging over experiments that have not been performed. See Jones (1991) for more on this issue.

In this paper, we consider coordinate free linear models with spherically symmetric error terms. The normal distribution has long served as the standard model in the investigation of linear models. One of its main attractive feature is that it depends on a small number of parameters which have a direct interpretation. As an alternative, the normal distribution has been generalized in two important directions, first as a special case of the exponential

family and secondly as a spherically symmetric distribution. We will consider the latter. There are a variety of equivalent definitions and characterizations of the class of spherically symmetric distributions. A comprehensive review is given by Fang, Kotz, and Ng (1990). In order to underline the intrinsic aspect of our results, the approach of multivariate analysis adopted here is coordinate free (*cf.* Stone, 1987). We prefer to follow the coordinate free approach since it is not necessary to choose a basis matrix to describe the subspace Θ. This subspace can be described in two ways. Firstly we can pick a basis matrix and then define Θ as the space spanned by its columns. This is the coordinatized version of the linear model. If $\mathbf{X}$ is a basis matrix for Θ, then $\mathbf{X}$ is a $n \times \Pi$ matrix of rank Π, and there exists a unique $\Pi-$ dimensional parameter β such that $\theta = \mathbf{X}\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\theta$. Secondly, we can define Θ by a set of linear contrasts which the elements must satisfy. This second approach is useful in the analysis of variance problem. Many formulae may be unified using projections and lengths of projections and they are easier to derive using the least squares property of projections rather than the equivalent matrix expressions.

In the next section, we set up the model under study and give a precise formulation of the variable selection problem. In Section 3, we compute the risk functions of some well known cases selection rules and compare their decision theoretic properties. We find conditions for which a class of prediction error estimators of the form (1) are dominated by a class of prediction error estimators of the form (2). The comparison depends on the dimension of the model, the sample size, and the spherically symmetric error distribution. We next propose a new family of prediction error estimators and we show that the new family of rules dominate the best of the rules of the form (2). In Section 4, we compare the outcomes of the various procedures via simulation. The appendix contains some technical results and all of the proofs of the propositions and theorems.

2 The Model and Formulation

Let y be an observation, in an n-dimensional Euclidian space $(\mathbf{E}, <,>)$, from a spherically symmetric distribution Q_θ around a location parameter θ. The main hypothesis about Q_θ is that θ belongs to a linear subspace $\Theta \subset \mathbf{E}$ of dimension Π with $0 < \Pi < n$. That is, $y = \theta + \varepsilon$ where ε is distributed as Q_0 on E and $\theta \in \Theta \subset E$. As mentioned previously, if we choose $\mathbf{X}$ as an $n \times \Pi$ full rank basis matrix in Θ, then we have the usual regression model $\theta = X\beta$. However, since we are taking the coordinate free approach, our results can be applied to a broader class of problems. Suppose we wish to estimate θ, by a decision rule $\varphi(y)$, using the sum of squared error loss $\| \theta - \varphi(y) \|^2$ where $\| \cdot \|$ denotes the norm connected with the inner product $<,>$. This loss is the prediction error sum of squares ($PES(\varphi \mid \theta)$). As $\Pi < n$ the usual estimator of θ is the orthogonal projector φ_0 from $\mathbf{E}$ onto Θ; this is the usual least squares estimator. In the coordinatized linear model with $\theta = X\beta$ the least

squares estimator has the usual form of $\varphi_0(y) = X(X^TX)^{-1}X^Ty$. The analysis here will be based on using the least squares estimator. One could also easily imagine using some other sort of estimator for θ, such as a shrinkage-type estimate. The approach in George and Foster (1997) essentially reduces to using a "positive part" estimator for θ. As one could construct the appropriate estimators of loss (see Fourdrinier and Wells 1995) we could have also considered a fairly wide class of shrinkage estimators for θ and their corresponding loss estimates.

In the multiple regression problem, $\theta = X\beta$, it may turn out that some of the components of β are equal to zero. A more parsimonious linear model might be

$$y = \theta_I + \varepsilon \text{ for } \theta_I = X_I\beta_I, \tag{3}$$

where I is a subset of p_I distinct positive integers less than or equal to Π, β_I is a p_I-vector containing the components of β that are indexed by the integers in I, and X_I is an $n \times p_I$ full rank matrix which contains the columns of X in the subset I.

Let A denote all the non-empty subsets of the integers $1, 2, \ldots, \Pi$. There are $2^\Pi - 1$ possible subsets $I \in A$ which give rise to different models M_I of the form (3). For any such model M_I, the PES can be define as above through the choice of an estimator φ_I of $\theta_I \in \Theta_I$, that is, $PES(\varphi_I \mid \theta_I)$. If one could compute $PES(\varphi_I \mid \theta_I)$, then it would be easy to rank the $2^\Pi - 1$ models, the model with the smallest $PES(\varphi_I \mid \theta_I)$ would be declared best. However we do not have access to $PES(\varphi_I \mid \theta_I)$ since it depends on the unknown components of θ_I. Therefore we propose to estimate the different $PES(\varphi_I \mid \theta_I)$ through a choice of statistics λ_I and to rank the $2^\Pi - 1$ models M_I by their estimated PES. We are then viewing the statistic used in a selection rule as an estimate of the *PES*. That is, the goal of model selection, in our formulation, is to find the subset $I \in \mathbf{A}$ of dimension p_I such that estimated prediction error sum of squares $\lambda_I(y)$ is minimized. This goal is consistent with that of Efron and Tibshirani (1993, S17).

We now study how well such loss estimators, λ_I, of the $PES(\varphi_I \mid \theta_I)$ behaves. Most researchers only insist on unbiasedness (cf. Mallows 1973, Craven and Wahba 1979, Donoho and Johnstone 1994, 199x, Donoho et al. 199x, and Hurrich et al. 1998). An unbiasness criterion does not adequately assess the behavior of an estimate as the variance does not enter the assessment criteria. We will use a risk function criteria. To this aim, a further distance measure is needed. For mathematical simplicity, we use squared error to evaluate λ_I and define the risk function of λ_I by

$$\mathcal{R}(\lambda_I, \theta_I, PES(\varphi_I \mid \theta_I)) = E_{\theta_I}[(\lambda_I - \parallel \varphi_I - \theta_I \parallel^2)^2] = E_{\theta_I}[(\lambda_I - PES(\varphi_I \mid \theta_I))^2] \tag{4}$$

where E_{θ_I} denote the expectation with respect to Q_{θ_I} and $\theta_I \in \Theta_I$ corresponds to one of

the $2^{\Pi} - 1$ possible different models. Note that, as we are considering all the $2^{\Pi} - 1$ possible models M_I, a loss estimator λ is actually a family of estimators, that is, $\lambda = (\lambda_I)_{I \in A}$. Then we will say than a loss estimator $\lambda' = (\lambda'_I)_{I \in A}$ dominates another estimate $\lambda = (\lambda_I)_{I \in A}$ if, for any $I \in A$,

$$\mathcal{R}(\lambda'_I, \theta_I, PES(\varphi_I \mid \theta_I)) \leq \mathcal{R}(\lambda_I, \theta_I, PES(\varphi_I \mid \theta_I)). \tag{5}$$

How does one choose between two selection rules? When do we consider one selection rule better than another? As our selection procedures are based on the minimizers of loss estimates, we will consider that a selection procedure associated to a loss estimator $\lambda' = (\lambda'_I)_{I \in A}$ is better than one associated to a loss estimator $\lambda = (\lambda_I)_{I \in A}$ if the inequality in (5) is satisfied.

We now wish to underline two important points. First we are not at all concerned with the problem of optimally estimating θ, so we just use the least squares estimates $X_I(X_I^T X_I)^{-1} X_I^T y$ for any linear subspace Θ_I. Notice that, as it is an orthogonal projector, its main properties do not depend on the specific linear subspace Θ_I under consideration. Likewise it will be the same for the various loss estimators for which we will compare. Thus the arguments for utilizing one or another decision procedure will be uniform in $I \in A$. So we will now drop the subscript I for sake of presenttion and the results below will be stated in terms of θ, Θ and p rather than θ_I, Θ_I, and p_I, respectively.

The estimated prediction error approach to selecting the subset I, that is to select the parsimonious model, is to formulate the model selection problem as a loss estimation procedure. This is in contrast with the approach taken by Breiman (1992), Efron and Tibshirani (1993, §17), and Shao and Tu (1995, §7.4) who consider a risk estimation procedure. The problem of estimating the post data accuracy was first considered by Lehmann (1950) who estimated the power of a statistical test. In a series of papers Kiefer (1975, 1976, 1977) addressed the problem of developing conditional and estimated confidence theories to provide frequentist estimates of confidence. Berger (1985) compared the Bayesian and frequentist approaches to this problem. Johnstone (1988), Rukhin (1988), Lu and Berger (1989), Casella (1992), and Lele (1993) have discussed this problem in a variety of situations. In this article, we apply these ideas to the problem of variable selection.

We assume throughout this article that the sampling distribution of the errors in the general linear model are spherically symmetric. Recall that if y is an n-dimensional spherical random vector around θ, then y has a stochastic representation $y \stackrel{d}{=} R\mathcal{U}_\theta$ where R is a nonnegative random radius $\mathcal{U}_\theta$ is a uniformly distributed random variable on the unit sphere $S_{1,\theta} = \{y \in E : \| y - \theta \| = 1\}$ and R and $\mathcal{U}_\theta$ are independent. The n-dimensional spherical distributions provide a nice extension of the classical n-dimensional spherical normal centered at the origin. The class of elliptical distribution contains a variety of distributions,

many of which have heavy tails. In addition we can drop the classical independent and identically distributed error term assumption and assume a weaker exchangeability condition. Exchangeability is a much more reasonable and natural assumption for statistical modeling, see Draper, Hodges, Mallows, and Pregibon (1993).

3 Risk comparisons: Winners and Losers

In this section, we can compare various classes of model selection criteria that are based on estimated PES and propose some new ones. We will consider a variable selection criteria to be good if it is a good estimate of $PES(\varphi \mid \theta) = \| \theta - \varphi \|^2$ (see Efron and Tibshirani 1993, S17, Breiman 1992, Shao and Tu 1995, S7.4). Hence the goal of this section is to study estimates of $PES(\varphi \mid \theta)$ which in turn give us a view at the goodness of various selection rules. We first find the best selection procedures in the classes defined in (1) and (2). For (1) we find that Mallows' C_p have some optimal properties. As for (2), we find a criterion related to the well known Wahba's GCV. Next we compare these two optimal procedures. Lastly, we propose a procedure that dominates the previously mentioned selection rules.

We now consider the estimation of the $PES(\varphi_0 \mid \theta)$ of the usual least squares estimator φ_0 of θ (i.e. the orthogonal projector from $\mathbf{E}$ onto Θ). We will first examine estimators of the form of (1), that is, $\lambda_\beta = \| y - \varphi_0 \|^2 + \beta$, where β is constant, possibly depending on p. As previously mentioned the sampling distribution Q_θ is spherically symmetric around θ. Referring to the notations given in the appendix (see especially formula (8)), all the results will be first obtained using the uniform distribution $U_{R,\theta}$ on the sphere $S_{R,\theta} = \{y \in E : \| y - \theta \| = R\}$ of radius R centered at θ. Then they can be expressed through the distribution of the radial distribution (i.e. the distribution of the norm $\| \cdot \|$ under Q_θ and its expectation denoted by E.

First by Lemma A.1 we have the following result, which is proved in the appendix.

Proposition 3.1: Assume that the distribution Q_θ has a finite fourth moment and φ_0 is the least squares orthogonal projection from $\mathbf{E}$ unto Θ. Then

(i) the risk function at θ of λ_β equals

$$\mathcal{R}(\lambda_\beta, \theta, PES(\varphi_0 \mid \theta)) = E_\theta \left[\| y - \varphi_0 \|^2 - \| \varphi_0 - \theta \|^2)^2 \right] + 2\beta(n-2p)E[R^2]/n + \beta^2;$$

(ii) the optimal β is given by $\beta^* = E[R^2](2p-n)/n$;

(iii) the risk of the optimal estimate in the class $\lambda_{\beta *}$ equals

$$E[R^4][\frac{(n-p+2)(n-p)+p(p+2)-2p(n-p)}{n(n+2)}] - [E[R^2]]^2 \frac{(n-2p)^2}{n^2};$$

(iv) the optimal estimate in the class λ_{β^*} is an unbiased estimate of $PES(\varphi_0 \mid \theta)$.

The optimal selection rule in this class can be found for different error distributions. Note that the term $E[R^2]/n$ is exactly the variance of the error distribution of the general linear model. If the error term is normally distributed with scale parameter σ, we have $\lambda_{\beta *} = \| y - \varphi_0 \|^2 - (n - 2p)\sigma^2$, which is the famous Mallows' C_p selection rule. It is customary in practice to estimate σ^2 by the usual estimate of σ^2 under the model with all Π variables included. It is important to note that the true optimal selection rule in this class depends upon knowing the variance of the error distribution; hence, if this variance is estimated, the selection rule is an approximation of the optimal rule. This optimality result is a finite sample results, as opposed to the results of Li (1987) who proved the optimality of C_p in an asymptotic sense. The optimality of λ_{β^*} depends on the fixed distribution at hand; therefore, in this case, we do not have robustness of the optimal rule.

Now we consider selection rules of the form of (2), that is, $\lambda_\alpha = \alpha \| y - \varphi_0 \|^2$, where $\alpha > 0$ is a constant, possibly depending on p. The properties of λ_α, using similar arguments as those used in Proposition 3.1, are given in the following proposition, , which is proved in the appendix.

Proposition 3.2: Assume that the distribution Q_θ has a finite fourth moment and φ_0 is the least squares orthogonal projection from E unto Θ. Then

(i) the risk function at θ of λ_α is given by

$$\mathcal{R}(\lambda_\alpha, \theta, PES(\varphi_0 \mid \theta)) = \left[\alpha^2 - 2a\frac{p}{n-p+2} + \frac{p(p+2)}{(n-p)(n-p+2)}\right]\left[\frac{(n-p+2)(n-p)}{n(n+2)}\right] E(R^4);$$

(ii) the optimal α is given by $\alpha^* = p/(n-p+2)$;

(iii) the risk of λ_{α^*} is given by $\mathcal{R}(\lambda_{\alpha^*}, PES(\varphi_0 \mid \theta)) = \frac{2p}{n(n-p+2)} E[R^4]$;

(iv) the bias of λ_{α^*} equals $\frac{2p}{n(2+p-n)} E[R^2]$.

Note that the optimal λ_α does not depend on the radial distribution as does λ_{β^*}; thus λ_{α^*} has some nice robustness properties. Therefore λ_{α^*} has optimality properties for the entire class of spherically symmetric distributions. The optimal λ_{α^*} is related to, but is not identical to, the GCV and S_p. Again it is important to note that this optimality result holds for finite samples. Li (1987) has shown that GCV is asymptotically optimal. An alternative estimator of $PES(\varphi_0 \mid \theta)$ is the unbiased estimator λ_u which is given by $\lambda_u = p \| y - \varphi_0 \|^2 / (n-p)$. (The unbiasedness of λ_u follows from Lemma A.1 *(i)* by taking $q = 0$ and $\gamma \equiv 1$.) Since λ_{α^*} and λ_u are asymptotically equivalent to GCV, they are also asymptotically optimal.

How do we choose between the two optimal PES estimators λ_{α^*} and λ_{β^*}? Asymptotic results do not give any guidance since both λ_{α^*} and λ_{β^*} are asymptotically optimal. One problem with λ_{β^*} is that it depends on the possibly unknown radial distribution. This dependence on $E(R^2)$ is not a problem in large samples; however, if one wishes to study to

finite sample behavior of estimates, it is a draw back. A risk comparison between the two estimates of *PES* is given in the next result, which is proved in the appendix.

Proposition 3.3: Assume that the distribution Q_θ has a finite fourth moment and φ_0 is the least squares orthogonal projection from E unto Θ. Then the risk of λ_{α^*} is smaller than that of λ_{β^*} when $n > 2p$.

Is there a prediction error estimator better than λ_{α^*}? Since λ_{β^*} depends on the possibly unknown second moment of the radial distribution and it is dominated by λ_{α^*}, in the important case where $n > 2p$, in the light of Proposition 3.3 we think that λ_{β^*} should be dismissed. For the remainder of this section, we will develop a class of procedures that dominate λ_{α^*}. The goal is now to prove the domination of the estimator λ_{α^*} by a competing prediction error loss estimator λ of the form

$$\lambda(y) = \lambda_{\alpha^*}(y) - \parallel y - \varphi_0 \parallel^4 \gamma(\varphi_0) \tag{6}$$

where $\gamma(\cdot)$ is a positive function. It is important to note that in the shrinkage function the residual term $\parallel y - \varphi_0 \parallel$ appears. It turns out that the use of this term needs less assumptions about the distributions than when it does not appear. Specifically, this gives a robustness property to the results since they are valid for the entire class of spherically symmetric distributions (under the required moment conditions). Since, for a given observation y, the residual term $\parallel y - \varphi_0(y) \parallel$ represents the square of the distance between y and its projection on Θ, it is intuitively natural that its consideration strengthens the information we use through the estimator.

Our primary example will be to choose $\gamma(t) = 2(p-4)/[(n-p+4)(n-p+6) \parallel t \parallel^2]$. In this case the prediction error estimated loss $(Peel)$ is

$$\begin{aligned} Peel(y \mid \varphi_0) &= \tfrac{p}{n-p+2} \parallel y - \varphi_0 \parallel^2 - \tfrac{2(p-4)}{(n-p+4)(n-p+6)} \tfrac{\|y-\varphi_0\|^4}{\|\varphi_0\|^2}. \\ &= \left(\tfrac{p}{n-p+2} - \tfrac{2(p-4)}{(n-p+4)(n-p+6)} \tfrac{\|y-\varphi_0\|^4}{\|\varphi_0\|^2}\right) \parallel y - \varphi_0 \parallel^2 \end{aligned} \tag{7}$$

The shrinkage factor turns out to be related to the reciprocal of the *F-statistic* for testing that all of the parameters of a p-dimensional linear model are equal to zero. Hence, if the model does not fit, then the shrinkage factor is larger while, if there is a good fit, the shrinkage factor is smaller. The selection function λ also has the nice intuitive properties of the penalized selectors in (1), however, now the penalty function depends on the data, rather than only on the dimension of the model. The penalty term in λ can also be viewed as the famous *F-to-enter* quantity from stepwise regression. Therefore operationally to compare two models Θ_I and $\Theta_{I'}$, we compute the respective least squares estimators φ_{0_I} and $\varphi_{0_{I'}}$ and declare Θ_I to be better than $\Theta_{I'}$ if $Peel(y \mid \varphi_{0_I}) \leq Peel(y \mid \varphi_{0_{I'}})$.

The general result is given below and its proof is in the appendix.

Theorem 3.1: Assume that $p > 4$, the distribution Q_θ has a finite fourth moment and the function γ is twice weakly differentiable on Θ and there exists a constant $\kappa > 0$ such that, $\gamma(t) \leq \kappa / \parallel t \parallel^2$ for every $t \in \Theta$. A sufficient condition under which the estimator λ given in (6) dominates the estimator $\lambda_{\alpha *}$ is that γ satisfies the differential inequality

$$\gamma^2 + \frac{2}{(n-p+4)(n-p+6)} \triangle \gamma \leq 0.$$

The example in (7) satisfies the condition of the theorem. More precisely, with $\gamma(t) = d/ \parallel t \parallel^2$ for all $t \in \Theta$, it is easy to derive $\triangle\gamma(t) = -2d(p-4)/ \parallel t \parallel^4$ and thus the sufficient condition of the theorem is written as $0 < d \leq 4(p-4)/(n-p+4)(n-p+6)$, which only occurs when $p > 4$. Straightforward calculus shows that the value of d that makes the left hand side of the inequality most negative is given by $2(p-4)/(n-p+4)(n-p+6)$. Notice that this value of d, compared with the proof of Theorem 3.1, does not necessarily lead to a maximum difference in risk between λ and $\lambda_{\alpha *}$. A possible problem with the estimators λ in (6) is that it may be negative, which should not happen since we are estimating a non-negative quantity. A simple remedy to this problem is to use the positive-part estimators.

The comparisons above were done under the hypothesis that there was no model bias, that is, φ_0 is the least squares orthogonal projection onto the correct model space Θ. In the case where the projection onto the wrong space Θ where the true model is defined by Θ^*, there is model bias. In this case one could use the risk function $E_\theta[(\lambda - \parallel \varphi_0 - \theta^* \parallel^2)^2]$, for $\theta^* \in \Theta^*$ to evaluate the prediction error sum of squares estimate. Under this risk, it can be shown that for the both families discussed above the optimal α and β both depend on the non-centrality parameter $\parallel \theta - \theta^* \parallel^2$. Furthermore the bias of λ_{α^*} and λ_{β^*} are $2p/[n(p-n+2)]E(R^2) - \parallel \theta - \theta^* \parallel^2$ and $- \parallel \theta - \theta^* \parallel^2$, respectively. Hence as $\parallel \theta - \theta^* \parallel^2$ increases the bias of both of the procedures becomes more sever. It is difficult to make any statements about risk domination of a particular prediction error estimator in the case of non-zero model bias. The zero model bias assumption is also made in risk comparisons made by Foster and George (1994) in their examination of the risk inflation of variable selection rules. It should be stressed that the assumption is made only in the derivation of the criterion. It is then possible to study the performance of this criterion without regard to the assumptions underlying its derivation.

4 Simulation Study

In this section we consider some simulation studies of the properties of the selection rule proposed in (6). The rules to which we compare *Peel* in (7) are C_p, leave-one-out cross

validation CV, Monte Carlo cross validation, $MCCV(n_v)$, the lasso, the garrotte, and ridge regression. The $MCCV(n_v)$ was proposed by Picard and Cook (1984) and was studied by Shao (1993). The simple idea behind $MCCV(n_v)$ is to randomly split the data b times and average the squared deviation errors over the splits. That is, randomly draw a collection $\mathcal{R}$ of b subsets of $1, ..., n$ that have size n_v, and select a model by minimizing the average of the squared deviation errors over the collection $\mathcal{R}$. Shao (1993) showed, via simulation, that $MCCV(n_v)$ out performs CV.

The *lasso* estimate, due to Tibsharani (1996), $\hat{\beta}$ is defined by

$$\hat{\beta} = \text{ arg } \min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_j \mid \beta_j \mid \leq t.$$

Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$, this computation is a carried out via quadratic programming problem with linear inequality constraints. The parameter $t \geq 0$ controls the amount of shrinkage that is applied to the estimates. Let $\hat{\beta}_j^o$ be the full least squares estimates and let $t_0 = \sum \mid \hat{\beta}_j^o \mid$. Values of $t < t_0$ will cause shrinkage of the solutions towards 0, and some coefficients may be exactly equal to zero. For example, if $t = t_0/2$, the effect will be roughly similar to finding the best subset of size $p/2$.

The motivation for the lasso came from an interesting proposal of Breiman (1995). Breiman's *non-negative garotte* minimizes

$$\sum_{i=1}^{n} \left(y_i - \sum_j c_j \hat{\beta}_j^o x_{ij} \right)^2 \text{ subject to } c_j \geq 0, \sum c_j \leq t.$$

The garotte starts with the least squares estimates and shrinks them by non-negative factors whose sum is constrained. In extensive simulation studies, Breiman showed that the garotte has consistently lower prediction error than subset selection and is competitive with ridge regression except when the true model has many small non-zero coefficients. A drawback of the garotte is that its solution depends on both the sign and the magnitude of the least squares estimates. In overfit or highly correlated settings where the least squares estimates behave poorly, the garotte may suffer as a result. In contrast, the lasso avoids the explicit use of the least squares estimates.

As a first example consider the following model, $y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$ where $i = 1, ..., 40, \varepsilon_i$ are *iid* from $N(0, 1)$, X_{ki} is the ith value of the kth predictor variable $X_k, X_{1k} \equiv 1$, and the values of $X_{ki}, k = 2, ..., 5, i = 1, ..., 40$ are taken from an

example in Gunst and Mason (1980). Some of the β_j's may be equal to zero, hence some prediction variables are selected from five possible variables $X_1, ..., X_5$ and the best model is then chosen via the given selection procedure. There are 31 possible models each denoted by a subset of $1, ..., 5$ that contain the indices of the variables X_k in the model. This example is essentially the one studied in Shao (1993).

In this comparison we considered the usual C_p and CV, and $MCCV(n_v)$ with $n_v = 25$ and $b = 2n$. As for improved estimate, we used positive part of the procedure in (7). Table 1 gives the empirical probabilities (based on 5000 simulation) of selecting each model in several cases. The results of the simulation show that the probability of selecting the correct model for the *Peel* selection procedure in (7) and $MCCV(n_v)$, are roughly equal, while high than CV, and C_p. Although CV and C_p are not very different, however, *Peel* and $MCCV(n_v)$ are much better than CV and C_p. The results are given in Table 1. We also repeated the simulation with student t errors with degrees of freedom equal to 5, 10 and 25. The results of the simulation were basically the same, therefore we do not report the results. Note that even though $p \leq 4$ for this example the nice properties of *Peel* are maintained.

As a second example consider we simulated 1000 datasets of 50 observations from the model $y = x\beta + \varepsilon$ where ε is standard normal. In this example we compare *Peel* in (7) to $MCCV$ (with $n_v = 25$ and $b = 2n$), lasso, garotte, and C_p. All of the turning parameters choosen by five-fold cross-validation. The X's are generated from independent $Normal(1, \cdot 5)$ distributions. In each of these examples the signal to noise ratio is nearly constant. The empirical probabilities of selecting the correct model are given in Table 2. All the methods work roughly the same when the signal is spread equally throughout the parameters outer. However, as one of the coordinater begins to dominate the signal the performance of the lasso and garotte degenerate much quicker than *Peel* and $MCCV$. The lasso and garotte are much less computational intensive than the *Peel* and $MCCV$.

As the next set of examples consider we mimic the simulation study given in Tibsharai (1996). We simulated 50 data sets consisting of 20 observations from the model $y = X\beta + \sigma\epsilon$, where ϵ is standard normal. In the following, we compare the full least squares estimates with the lasso, the non-negative garotte, ridge regression, C_p, and the new *Peel* proposal in (7). We used fivefold cross-validation to estimate the regularization parameter in each case. The mean-squared errors are computed over 200 simulations of this model.

As Example 3 let $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and correlation between x_i and x_j was $\rho^{|i-j|}$ with $\rho = 0.5$. We set $\sigma = 3$, and this gave a signal-to-noise ratio of approximately 5.7. Table 3 shows the mean-squared errors from this model for 200 simulations of this model. The *Peel* performs the best, then followed by the lasso, garotte, and ridge regression. Least squares and C_p both preform poorly.

Example 4 is the same as Example 3, but with $\beta_j = 0.85, = 1, \ldots, 8j$ and $\sigma = 3$; the signal-to-noise ratio was approximately 1.8. The results in Table 3 show that ridge regression

and *Peel* are best. The lasso and garotte seem to over shrink.

For Example 5 we chose a set-up that should be well suited for subset selection. The model is the same as Example 3, but with $\beta = (5, 0, 0, 0, 0, 0, 0, 0)$ and $\sigma = 2$ so that the signal-to-noise ratio was about 7. The results in Table 5 show that the garotte, *Peel* and lasso all work well. The garotte and lasso were designed for this situation.

As example 6 we examine the performance of the lasso in a bigger model. We simulated 50 data sets each having 100 observations and 20. We defined predictors $x_{ij} = z_{ij} + z_i$ where z_{ij} and z_i are independent standard normal variates. This induced a pairwise correlation of 0.5 among the predictors. The coefficient vector was $\beta =$ (0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2). Finally we defined $y = X\beta + 15\epsilon$ where ϵ was standard normal. This produced a signal-to-noise ratio of roughly 9. The results in Table 6 show that the ridge regression and *Peel* perform well in terms of mean squared error while *Peel* is the better of the two in terms of average number of zero coefficients, although the garotte and lasso are not all too much worse. In this example the *Peel* procedure requires much more computation than the other methods.

5 Appendix

An $n \times 1$ random vector ε is said to have a spherically symmetric distribution around zero if for every $\Gamma \in \mathcal{O}(n), \Gamma\varepsilon \stackrel{d}{=} \varepsilon$, where "$\stackrel{d}{=}$" means equal in distribution and $\mathcal{O}(n)$ denotes the group of $n \times n$ orthogonal transformations. A spherically symmetric random vector, in general, does not necessarily possess a density. However, if the density exists it must be at the form $g(\| \varepsilon \|^2)$ for some nonnegative function $g(\cdot)$ of a scalar variable. In this case

$$\int_E g(\| \epsilon \|^2) d\varepsilon = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)} \int_0^\infty z^{n/2-1} g(z) dz = 1.$$

Hence, a nonnegative function $g(\cdot)$ can be used to define a density $g(\| \varepsilon \|^2)$ for some spherical distribution if and only if

$$\int_0^\infty z^{n/2-1} g(z) dz < \infty.$$

The function g is called the density generator of the spherical distribution. A fundamentally important result in spherical distribution theory is the representation of random variables as a random radius times a uniform random vector on the unit sphere. Now if y is an n-dimensional random vector such that $y = \theta + \epsilon$, for some fixed $\theta \in R^P$, then it can

be shown that y has a stochastic representation $y \stackrel{d}{=} R\mathcal{U}_\theta$ where R is a random radius with distribution $\rho, \mathcal{U}_\theta$ is a random variable with uniform distribution on the unit sphere $S_{1,\theta} = \{y \in E :\| y - \theta \| = 1\}$ and R and $\mathcal{U}_\theta$ are independent. Or equivalently, if Q_θ is the distribution of y, then for every bounded function f, we have

$$E_\theta[f] = EE_{R,\theta}[f] = \int_{R_+} E_{R,\theta}[f]\rho(dR) \tag{8}$$

where E and $E_{R,\theta}$ denotes the expectation with respect to the radial distribution ρ (the distribution of the norm $\| \cdot \|$ under Q_0) and the uniform distribution $U_{R,\theta}$ on the sphere $S_{R,\theta} = \{y \in E :\| y - \theta \| = R\}$ of radius R and center θ, respectively.

In order to obtain the risk of any estimator λ of the $PES(\varphi)$, it suffices to calculate it working conditionally to the radius, that is to say to replace Q_θ by $U_{R,\theta}$ in the expression (4). Since the integrand terms in the risk functions depend on the observation only through φ_0, the expressions can be calculated using the fact that the distribution of φ_0, under $U_{R,\theta}$, has a density with respect to the Lebesgue measure on Θ (see Kelker 1970). It is worth noting that we do not assume that Q_θ has a density with respect to the Lebesgue measure on $\mathbf{E}$.

The classical example of a spherical distribution is the n-dimensional normal distribution, that is, ε is distributed $N_n(0, \sigma^2 I_n)$. In this case, the radial distribution is σ times a χ_n-random variable, that is $R \stackrel{d}{=} \sigma\chi_n$. Hence $E(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2 E(R^2) = \sigma^2 E\left(\chi_n^2\right) I_n/n = \sigma^2 I_n$, since χ_n^2 is a Chi Square distribution with n degrees of freedom.

With a multivariate Student distribution, the result differs according to the degrees m of freedom. Indeed consider, for Q_θ, the unscaled density

$$g(\| y - \theta \|^2) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)(\pi m)^{n/2}} \left[1 + \frac{\| y - \theta \|^2}{m}\right]^{-\frac{m+n}{2}}.$$

Here, the density of the radius is equal to

$$f(R) = \frac{2\pi^{n/2}}{\Gamma(n/2)} R^{n-1} g(R^2) = \frac{2\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})m^{n/2}} \left[1 + \frac{R^2}{m}\right]^{-\frac{m+n}{2}} R^{n-1}.$$

After some tedious calculations, we get $E[R^2] = nm/(m-2)$ for $m > 2$ and $E[R^4] = n(n+2)m^2/(m-4)(m-2)$ for $m > 4$. The fact that the Student distribution behaves differently from the normal distribution is well known. Although it gives a good approximation to the normal model, Zellner (1976) has shown that a t-distribution leaves, through the choice of m, more freedom to the experimenter.

Another example, which is not a mixture of normal distributions, is the Kotz distribution whose Q_θ has the density $g(\| x-\theta \|^2)$ with

$$g(s) = \frac{\Gamma(n/2)}{(2\pi)^{n/2}2^\gamma\Gamma(n/2+\gamma)} s^\gamma \exp(-\frac{2}{2}).$$

When $\gamma \neq 0$, the function γ is not completely monotonic (that is, $(-1)^m d^m g/ds^m \geq 0$, does not hold for every m), hence the distributions is not a normal mixture (see Berger, 1976). The density of the radius is given by

$$f(R) = \frac{2\pi^{n/2}}{\Gamma(n/2)} R^{n-1} g(R^2) = \frac{2^{1-(n/2+\gamma)}}{\Gamma(n/2+\gamma)} R^{n+2\gamma-1} \exp\left(-\frac{R^2}{2}\right).$$

A straightforward calculation gives $E[R^2] = n+2\gamma$ and $E[R^4] = (n+2\gamma)/(n+2+2\gamma)$ as long as the inequality $n+2\gamma > 0$ holds. For further examples see Fang, Kotz, and Ng (1990).

The following two results are crucial in all of our calculations. The proofs are given in the appendix of Fourdrinier and Wells (1995). The proof of Lemma A.1 *(i)* follows from two applications of the divergence theorem for weakly differentiable functions (see Ziemer, 1989), while Lemma A.1 *(ii)* follows from straightforward calculation.

Lemma A.1: Let φ_0 be the least squares orthogonal projection from E to Θ. Then, (i) for every twice weakly differentiable function γ and for every integer q,

$$\begin{aligned} E_{R,\theta}\left[\| y-\varphi_0 \|^q \| \varphi_0-\theta \|^2 \gamma(\varphi_0)\right] &= \tfrac{p}{n-p+q} E_{R,\theta}\left[\| y-\varphi_0 \|^{q+2} \gamma(\varphi_0)\right] \\ &+ \tfrac{1}{(n-p+q)(n-p+q+2)} E_{R,\theta}\left[\| y-\varphi_0 \|^{q+4} \triangle\gamma(\varphi_0)\right]; \end{aligned}$$

(ii) then for every integer $j \geq 1$,

$$E_\theta\left[\| y-\varphi_0 \|^{2j}\right] = E[R^{2j}]\Pi_{i=1}^j \frac{\frac{n-k}{2}+j-i}{\frac{n}{2}+j-i}.$$

Proof of Proposition 3.1: *(i)* Evaluating (5) at λ_β we have

$$\begin{aligned} &\mathcal{R}(\lambda_\beta, \theta, PES(\varphi_0 \mid \theta)) \\ &= E_\theta\left[(\| y-\varphi_0 \|^2 + \beta - \| \varphi_0-\theta \|^2)^2\right] \\ &= E\ \left[E_{R,\theta}[(\| y-\varphi_0 \|^2 - \| \varphi_0-\theta \|^2)^2]\right] + 2\beta E\ \left[E_{R,\theta}\left[(\| y-\varphi_0 \|^2 - \| \varphi_0-\theta \|^2\right]\right] + \beta^2. \\ &= E\ \left[E_{R,\theta}\left[(\| y-\varphi_0 \|^2 - \| \varphi_0-\theta \|^2)^2\right]\right] + 2\beta\tfrac{n-2p}{n-p} E\left[E_{R,\theta}[\| y-\varphi_0 \|^2]\right] + \beta^2 \\ &= E_\theta \| y-\varphi_0 \|^2 - \| \varphi_0-\theta \|^2)^2 + 2\beta(n-2p)E(R^2)/n + \beta^2. \end{aligned}$$

(ii) The risk function $\mathcal{R}(\lambda_\beta, \theta, PES(\varphi_0 \mid \theta))$ is quadratic and convex in β. Simple calculus yields the result.

(iii) The first term of $\mathcal{R}(\lambda_\beta, \theta, PES(\varphi_0 \mid \theta))$ may be simplified as

$$\begin{aligned}
&E_\theta \left[\| y - \varphi_0 \|^4 + \| \varphi_0 - \theta \|^4 - 2 \| y - \varphi_0 \|^2 \| \varphi_0 - \theta \|^2 \right] \\
&= E_\theta \left[\| y - \varphi_0 \|^4 \right] \left[1 + \frac{p(p+2)}{(n-p)(n-p+2)} - \frac{2p}{n-p+2} \right] \\
&= E(R^4) \left[\frac{(n-p+2)(n-p) + p(p+2) - 2p(n-p)}{n(n+2)} \right].
\end{aligned}$$

The first equality follows from Lemma A.1 *(i)* applied to the second and third term with $q = 0$ and $\gamma = \| \varphi_0 - \theta \|$ then $q = 2$ and $\gamma \equiv 1$. While the second equality follows from Lemma A.1 *(ii)* with $j = 2$.

(iv) Evaluating the difference in expectation we have

$$\begin{aligned}
&E_\theta[\lambda_{\beta^*} - \| \varphi_0 - \theta \|^2] \\
&= E_\theta \left[\| y - \varphi_0 - \theta \|^2 + \tfrac{(2p-n)}{n} E\,(R^2) - \| \varphi_0 - \theta \|^2 \right] \\
&= \tfrac{n-p}{n} E(R^2) + \tfrac{(2p-n)}{n} E(R^2) - \tfrac{p}{n} E(R^2),
\end{aligned}$$

where the calculation of the first term follows from Lemma A.1 *(ii)* and the third term from Lemma A.1 *(i)* with $q = 0$ and $\gamma \equiv 1$.

Proof of Proposition 3.2: The risk and bias calculations follow from multiple applications of Lemma A.1 as in Proposition 3.1. Statements *(ii)*, *(iii)* and *(iv)* can be deduced by simple calculus.

Proof of Proposition 3.3: The proof follows from a comparison of the risk functions given in Proposition 3.1 *(iii)* and 3.2 *(iii)*. It suffices to show

$$\frac{n}{(n-2p)^2} \left[\frac{(n-p+2)(n-p) + p(p+2) - 2p(n-p)}{n+2} \right] \geq \frac{[E(R^2)]^2}{E(R^4)}.$$

An application of Jensen's inequality yields that the right hand side of the inequality is bounded above by one. Therefore, to prove the result it is sufficient to show that the left hand side of the inequality is greater than one. Extremely tedious algebra shows that this is indeed the case. As an additional check of the algebra, we verified the left hand side of the inequality is greater than one numerically for all $n(> 2p)$ less than ten million. As the sample size tends to infinity it is easy to see that the weak inequality always holds.

Before giving the proof of Theorem 3.1, we consider the problem of the finiteness of the risks of the estimators λ_{a^*} and λ. It is easy to check, using the spherically symmetry of Q_θ and the proportionality of $E_\theta[\| y-\varphi_0 \|^4]$ and $E_\theta[\| \varphi_0-\theta \|^4]$ (this follows from two applications of Lemma *3.1* first with $q=0$ and $\gamma(t)=\| t-\theta \|^2$, then with $q=2$ and $\gamma(t)=1$), that the risk of the optimal estimator λ_* is finite if and only if Q_θ has a finite fourth moment. If the risk of λ_{a*} is finite, straightforward calculation (see the first expression of the risk of λ given at the beginning of the proof of Theorem 3.1) and an application of the Cauchy-Schwarz inequality show that the risk of the shrinkage estimator (6) is finite if and only if $E_\theta[\| y-\varphi_0 \|^8 \gamma^2(\varphi_0)] < \infty$. A straightforward way of showing this expectation is finite is to assume that there exists a constant $\beta > 0$ such that, $\gamma(t) \leq \beta / \| t \|^2$ for every $t \in \Theta$. This condition is often assumed when estimating a location parameter, see Cellier and Fourdrinier (1995) for more details and references. Indeed, working conditionally on the radius R, it implies

$$E_{R,\theta}[\| y-\varphi_0 \|^8 \gamma^2(\varphi_0)] \leq \beta^2 R^4 E_{R,\theta}\left[\left(\frac{\| y-\varphi_0 \|^2}{\| \varphi_0 \|^2}\right)^2\right] \tag{9}$$

On the right hand side of (9), for $\theta=0$, the expectation is independent of R since it is the second moment of a generalized noncentral F distributed random variable with $n-p$ and p degrees of freedom (up to a multiplicative constant). This moment is finite as soon as $p>4$ and remains finite for $\theta \neq 0$ (since the distribution is merely translated by θ) and can be bounded from above by a constant independent of R. Now when we uncondition, with the assumption that Q_θ has a finite fourth moment, the right hand side of (9) is finite. Hence, to ensure risk finiteness, we will assume $p>4$. We can now state the following theorem, whose proof is adapted from Fourdrinier and Wells (1995).

Proof of Theorem 3.1: Since Q_θ is spherically symmetric around θ, it is clear it suffices to obtain the result in working conditionally on the radius. Referring to the notations given above for $R>0$ fixed, we can compute using the uniform distribution $U_{R,\theta}$ on the sphere $S_{R,\theta}$. Hence we have

$$E_{R,\theta}[(\lambda - \| \varphi_0-\theta \|^2)^2] = E_{R,\theta}[(\lambda_{\alpha^*}(y) - \| y-\varphi_0 \|^4 \gamma(\varphi_0) - \| \varphi_0-\theta \|^2)^2].$$

Developing the cross-product term and using the form of $\lambda_{\alpha*}$, we have

$$\begin{aligned} E_{R,\theta}[(\lambda_{\alpha*} - \| \varphi_0-\theta \|^2) \| y-\varphi_0 \|^4 \gamma(\varphi_0)] &= \tfrac{p}{n-p+2} E_{R,\theta}[\| y-\varphi_0 \|^6 \gamma(\varphi_0)] \\ &\quad - E_{R,\theta}[\| \varphi_0-\theta \|^2 \| y-\varphi_0 \|^4 \gamma(\varphi_0)]. \end{aligned}$$

Using Lemma A.1 with $q=4$, the second integral of the right hand side becomes

$$E_{R,\theta}[\| \varphi_0 - \theta \|^2 \| y \quad -\varphi_0 \|^4 \gamma(\varphi_0)] = \tfrac{p}{n-p+4} E_{R,\theta}[\| y - \varphi_0 \|^6 \gamma(\varphi_0)]$$
$$+\tfrac{1}{(n-p+4)(n-p+6)} E_{R,\theta}[\| y - \varphi_0 \|^8 \triangle\gamma(\varphi_0)].$$

Replacing this expression in the cross-product term and combining the terms with $\| y - \varphi_0 \|^6 \gamma(\varphi_0)$, we get

$$E_{R,\theta}[(\lambda - \| \varphi_0 - \theta \|)^2] = E_{R,\theta}[(\lambda_{\alpha*} - \| \varphi_0 - \theta \|)^2] - \frac{4p}{(n-p+2)(n-p+4)} E_{R,\theta}[\| y - \varphi_0 \|^6 \gamma(\varphi_0)]$$

$$+E_{R,\theta}[\| y - \varphi_0 \|^8 \gamma(\varphi_0)] + \frac{2}{(n-p+4)(n-p+6)} E_{R,\theta}[\| y - \varphi_0 \|^8 \triangle\gamma(\varphi_0)].$$

Since, on the right hand side, the second term is negative (γ is positive) and we have the same power 8 for the term $\| y - \varphi_0 \|$ in the two last integrals, it is clear that $\mathcal{R}(\lambda, \theta, PES(\varphi_0)) \leq \mathcal{R}(\lambda_{\alpha*}, \theta, PES(\varphi_0))$ provided that $\gamma^2 + \frac{2}{(n-p+4)(n-p+6)} \triangle \gamma \leq 0$.

The proof of Theorem 3.1 and Lemma A.1 show that the power $q = 4$ chosen for the residual term $\| y - \varphi_0 \|$ in the expression of λ is the only one possible. Indeed for any arbitrary q we would obtain $\| y - \varphi_0 \|^{2q}$ before γ^2 and $\| y - \varphi_0 \|^{q+4}$ before $\triangle\gamma$ and the comparison of these two terms is possible only if $2q = q + 4$, that is to say only if $q = 4$.

6 References

Akaike, H. (1970), "Statistical Predictor Identification," *The Annals of the Institute of Statistical Mathematics*, 22, 203-217.

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions Automatic Control*, 19, 716-723.

Allen, D.M. (1971), "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics*, 16, 125-127.

Allen, D.M. (1974), "Mean Square Error of Prediction as a Criterion for Selecting Variables," *Technometrics*, 13, 469-475.

Berger, J.O. (1976), "Inadmissibility Results for Generalized Bayes Estimators of Coordinates of a Location Parameter," *The Annals of Statistics*, 4, 302-333.

Berger, J.O. (1985), "The Frequentist Viewpoint and Conditioning," Proceedings of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer. (L. LeCam and R. Olshen eds.). Wadsworth, Blemont, 15-44.

Berger, J.O. and Pericchi, L.R. (1993), "The Intrinsic Bayes Factor for Model Selection and Prediction," JASA, xx, xx

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738-754.

Breiman, L. and Freedman, D (1983), "How Many Variables Should be Entered in a Regression Equation," *Journal of the American Statistical Association*, 78, 131-136.

Breiman, L. (1995), " Better Subset Regression Using the Nonnegative Garrotte," *Technometrics* 37, (4) 373-384.

Casella, G. (1988), "Estimating post-data accuracy," In *Statistical Decision Theory and Related Topics IV, vol. 1* (S.S. Gupta and J.O. Berger, eds.). New York: Springer-Verlag.

Cellier D. and Fourdrinier, D. (1992), "Shrinkage Estimators Under Spherically Symmetry for the General Linear Model," *Journal of Multivariate Analysis*, 52, 338-351.

Craven, P. and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Num. Math.* 31, 377-403.

Donoho, D.L. and Johnstone, I.M. (1994), "Ideal Spacial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425-455.

Donoho, D.L. and Johnstone, I.M. (199x), "Adapting to Unknown Smoothness Via Wavelet Shrinkage," *Journal of the American Statistical Association*, xx, xxx.

Donoho, D.L.,Johnstone, I.M., Kerkyacharian, G. and Picard, D. (199x), "Wavelet Shrinkage: Asmyptopia?" *Journal of the Royal Statistical Society, Ser B.*, xx, xxx.

Draper, D., Hodges, J.S., Mallows, C.L. and Pregibon, D. (1993), "Exchangeability and Data Analysis," *Journal of the Royal Statistical Society, Ser.A*, 156, 9-38.

Efron, B. (1986), "How Biased is the Apparent Error Rate of a Prediction Rule," *Journal of the American Statistical Association*, 81, 461-470.

Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

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

Foster, D.P. and George, E.I. (1994), "The Risk Inflation in Multiple Regression," *The Annals of Statistics*, 22, 1947-1975.

Fourdrinier D. and Wells, M.T. (1995), "Estimation of a Loss Function for Spherically Symmetric Distributions in the General Linear Model," *The Annals of Statistics*, 23, 571-592.

George, E.I. and Foster, D.P. (1997), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, in press.

Gunst, R.F. and Mason, R.L. (1980), *Regression Analysis and its Application: A Data Oriented Approach*, New York: Marcel Dekker.

Harvich, C.M., Simonoff, J.S. and Tsai, C.L. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society, Ser. B.*, 60, 271-293.

Hocking, R.R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-49.

Johnstone, I. (1988), "On Inadmissibility of Some Unbiased Estimates of Loss," In *Statistical Decision Theory and Related Topics IV, vol. 1* (S.S. Gupta and J.O. Berger, eds.), 361-379. New York: Springer-Verlag.

Jones, M.C. (1991), "The Roles of ISE and MISE in Density Estimation," *Statist. Probab. Lett.* 12, 51-56.

Kiefer, J. (1975), "Conditional Confidence Approach in Multi-Decision Problems," In *Multivariate Analysis IV* (P.R. Krishnaiah, ed.), New York: Academic Press.

Kiefer, J. (1976), "Admissibility of Conditional Confidence Procedures," *The Annals of Statistics*, 4, 836-865.

Kiefer, J. (1977), "Conditional Confidence Statements and Confidence Estimators," *Journal of the American Statistical Association*, 72, 789-827.

Kelker, D. (1970), "Distribution Theory of Sperhical Distributions and a Location Scale Paramter Generalization," *Sankhyā*, 32, 479-430.

Lehmann, E.L. and Sheffe, H. (1950), Completeness, similar regions, and unbiased estimates, *Sankhya*, 10, 305-340.

Lele, C. (1993), "Admissibility Results in Loss Estimation," *The Annals of Statistics*, 21, 378-390.

Li, K.C. (1987), "Asymptotic Optimality for Cp, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958-975.

Lu, K.L. and Berger J.O. (1989), "Estimation of Normal Means: Frequentist Estimation of Loss," *The Annals of Statistics*, 17, 890-906.

Mallows, C.L. (1973), "Some Comments on Cp," *Technometrics*, 15, 4, 661-6676.

Miller, A.J. (1990), *Subset Selection in Regression*, New York: Chapman and Hall.

Picard, R.R. and Cook, R.D. (1984), "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79, 575-583.

Rissanen, J. (1986), "A Predictive Least Squares Principle," *IMA Journal of Mathematical Control and Information*, 3, 211-222.

Roecker, E.B. (1991), "Prediction Error and its Estimation for Subset-Selected Models," *Technometrics*, 33, 459-468.

Rukhin, A. L. (1988), "Estimated Loss and Admissible Loss Estimators," In *Statistical Decision Theory and Related Topics IV, vol. 1* (S.S. Gupta and J.O. Berger, eds.), 409-418. New York: Springer-Verlag.

Schwarz, G. (1978), Estimating the Dimension of a Model. *The Annals of Statistics*, 6, 461-464.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486-494.

Shao, J. and Tu, S. (1995), *The Jackknife and Bootstrap*, New York, Springer.

Shibata, R. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, 68, 1, 45-54.

Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society, Ser.B*, 39, 44-47.

Stone, M. (1987), *Coordinate-Free Multivariate Statistics*, London/ New York: Oxford Univ. Press (Clarendon).

Thompson, M.L. (1978), "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation," *International Statistical Review*, 46, 1-19.

Thompson, M.L. (1978), "Selection of Variables in Multiple Regression: Part II. Chosen Procedures, Computations and Examples," *International Statistical Review*, 46, 129-146.

Tibsharani, R. (1996) " Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser.B*, 58, 267-288.

Wei, C.Z. (1992), "On Predictive Least Squares Principles," *The Annals of Statistics*, 20, 1-42.

Woodroofe, M. (1982), "On Model Selection and the Arc-Sine Laws," *The Annals of Statistics*, 10, 1182-1194.

Zellner, A. (1976), "Bayesian and Non-Bayesian Analysis of the Regression Model with Multivariate Student - t Error Terms," *Journal of the American Statistical Association*, 71, 400-405.

Zhang, P. (1992), "On the Distributional Properties of Model Selection Criteria," *Journal of the American Statistical Association*, 87, 732-737.

Zhang, P. (1993), "Model Selection Via Multifold Cross Validation," *The Annals of Statistics*, 21, 299-313.

Ziemer, W.P. (1989), "*Weakly Differentiable Functions - Sobolev Spaces and Functions of Bounded Variation*," New York: Springer Verlag.

	Model	C_p	CV	$MCCV$	$Peel$
$\beta = (2,0,0,4,0)^T$	1,4	.524	.578	.820	.831
	1,2,4	.133	.142	.126	.096
	1,3,4	.134	.137	.121	.094
	1,4,5	.138	.139	.104	.099
	1,2,3,4	.064	.072	.021	.023
	1,2,4,5	.050	.046	.011	.014
	1,3,4,5	.021	.031	.009	.009
	1,2,3,4,5	.008	.007	.000	.000
$\beta = (2,0,0,4,8)^T$	1,4,5	.707	.721	.911	9.23
	1,2,4,5	.173	.152	.072	.070
	1,3,4,5	.157	.148	.063	.055
	1,2,3,4,5	.083	.089	.009	.000
$\beta = (2,9,0,4,8)^T$	1,4,5	.021	.013	.019	.007
	1,2,4,5	.821	.836	.937	.955
	1,3,4,5	.020	.019	.014	.011
	1,2,3,4,5	.164	.152	.012	.008
$\beta = (2,9,6,4,8)^T$	1,2,3,5	.001	.001	.001	.000
	1,2,4,5	.002	.001	.003	.001
	1,3,4,5	.026	.030	.024	.019
	1,2,3,4,5	.976	.981	.952	.983

Table 1: Empirical Probabilities of Selecting Each Model.

True Parameter Value	$Peel$	$MCCV$	Lasso	Garrotte	C_p
$(1,1,1,1,1,0,0,0,0,0)^T$	.901	.882	.824	.819	.621
$(2,1,1,.5,.5,0,0,0,0,0)^T$	.863	.841	.793	.774	.542
$(3,.5,.5,.5,.5,0,0,0,0,0)^T$	.852	.794	.682	.673	.535
$(3,1,.5,.4,.1,0,0,0,0,0)^T$	.846	.802	.636	.606	.501
$(4.6,.1,.1,.1,.1,0,0,0,0,0)^T$	.837	.787	.649	.642	.493
$(4.8,.05,.05,.05,.05,0,0,0,0,0)^T$	.724	.703	.617	.586	.462
$(4.9,.075,.05,.05,.025,0,0,0,0,0)^T$	.617	.564	.316	.284	.254

Table 2: Empirical Probabilities of Selecting the Correct Model.

Method	Median mean-squared error	Ave. No. of zero coefficients
Least Squares	3.57 (0.29)	0.0
Lasso	2.41 (0.16)	3.8
Garrotte	2.76 (0.19)	3.3
C_p	3.46 (0.27)	2.6
Ridge Regression	3.04 (0.19)	0.0
Peel	2.40 (0.17)	4.5

Table 3: Results for Example 3 where $\beta = (3, 1.5, 0, 0, 0, 2, 0, 0, 0)^T$

.

Method	Median mean-squared error	Ave. No. of zero coefficients
Least Squares	7.88 (0.72)	0.0
Lasso	6.81 (0.44)	2.8
Garrotte	7.27 (0.44)	3.6
C_p	8.01 (0.52)	1.3
Ridge Regression	2.95 (25)	0.0
Peel	3.30 (1.7)	0.8

Table 4: Results for Example 4 where $\beta = (0.85, \ldots, 0.85)^T$

.

Method	Median mean-squared error	Ave. No. of zero coefficients
Least Squares	2.52 (0.06)	0.0
Lasso	1.01 (0.02)	5.5
Garrotte	0.72 (0.02)	5.9
C_p	1.17 (0.05)	4.4
Ridge Regression	2.93 (0.07)	0.0
Peel	.84 (0.03)	5.8

Table 5: Results for Example 5 where $\beta = (5, 0, 0, 0, 0, 0, 0, 0)^T$)

.

Method	Median mean-squared error	Ave. No. of zero coefficients
Least Squares	83.5 (6.3)	0.0
Lasso	42.7 (5.3)	11.5
Garrotte	47.1 (4.2)	12.1
C_p	58.4 (5.6)	8.2
Ridge Regression	36.5 (2.1)	0.0
$Peel$	39.2 (3.4)	9.6

Table 6: Results for Example 6 where $\beta = (0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2)^T$.

Chapitre 2

On Loss Estimation

On Loss Estimation

Dominique Fourdrinier *
Martin T. Wells †

May 10, 2010

Abstract

Let X be a random vector with distribution P_θ where θ is an unknown parameter. When estimating θ by some estimator $\varphi(X)$ under a loss function $L(\theta, \varphi)$, classical decision theory advocates that such a decision rule should be used if it has suitable properties with respect to the frequentist risk $R(\theta, \varphi)$. However, after having observed $X = x$, instances arise in practice in which φ is to be accompanied by an assessment of its loss $L(\theta, \varphi(x))$, which is, since θ is unknown, unobservable. A common approach to this assessment is to consider estimation of $L(\theta, \varphi(x))$ by an estimator δ, called the loss estimator. To date, there is a sizeable literature dealing with loss estimation. Here, we present an expository development of loss estimation with substantial emphasis on the setting where the distributional context is normal and its extension to the case where the underlying distribution is spherically symmetric. Bayes estimation is also considered and comparisons are made with unbiased estimation.

AMS 2010 subject classifications. Primary 62C15, 62C20, 62F10, 62H12.

Keywords and phrases: conditional inference, linear model, loss estimation, quadratic loss, risk function, robustness, shrinkage estimation, spherical symmetry, SURE, unbiased estimator of loss, uniform distribution on a sphere.

*Université de Rouen, LITIS EA 4108, Avenue de l'Université, BP 12, 76801 Saint-Étienne-du-Rouvray, France. The support of the ANR grant 08-EMER-002 is gratefully acknowledged.

†Cornell University, Department of Statistical Science, 1190 Comstock Hall, Ithaca, NY 14853, USA. The support of NSF Grant 06-12031 and NIH Grant R01-GM083606-01 are gratefully acknowledged.

1 Introduction

Suppose X is an observable from a distribution P_θ parameterized by an unknown parameter θ. In classical decision theory, it is usual, after selecting an estimation procedure $\varphi(X)$ of θ, to evaluate it through a loss criterion, $L(\theta, \varphi(X)$, which represents the cost incurred by the estimate $\varphi(X)$ when the unknown parameter equals θ. In the long run, as it depends on the particular value of X, this loss cannot be appropriate to assess the performance of the estimator φ. Indeed, to be valid (in the frequentist sense), a global evaluation of such a statistical procedure should be based on all the possible observations. Consequently, it is common to report the risk $R(\theta, \varphi) = E_\theta[L(\theta, \varphi(X)]$ as a measure of the efficiency of φ (E_θ denotes expectation with respect to P_θ). Thus we have at our disposal a long run performance of $\varphi(X)$ for each value of θ. However, although this notion of risk can effectively be used in comparing $\varphi(X)$ with other estimators, it is inaccessible since θ is unknown. The usual frequentist risk assessment is the maximum risk $\overline{R}_\varphi = \sup_\theta R(\theta, \varphi)$.

When $X = x$ the loss, $L(\theta, \varphi(x))$, itself could serve as a perfect measure of the accuracy of φ if it were available (which it is not since θ is unknown). It is natural to estimate $L(\theta, \varphi(x))$ by a data-dependent estimator $\delta(X)$, a new estimator called a loss estimator. Such an estimator can serve as a data-dependent assessment (instead of $\overline{R}_\varphi$). This is a conditional approach in the sense that accuracy assessment is made on a data-dependent quantity, the loss, instead of the risk.

To evaluate the extend to which $\delta(X)$ successfully estimates $L(\theta, \varphi(X))$, another loss is required and it has become standard, for simplicity, to use the squared error

$$L^*(\theta, \varphi(X), \delta(X)) = (\delta(X) - L(\theta, \varphi(X))^2 \,. \tag{1.1}$$

In so far as we are thinking in terms of long-run frequencies, we adopt a frequentist approach to evaluating the performance of L^* by averaging over the sampling distribution of X given θ, that is, by using a new notion of risk

$$\mathcal{R}(\theta, \varphi, \delta) = E_\theta[L^*(\theta, \varphi(X), \delta(X))] = E_\theta[(\delta(X) - L(\theta, \varphi(X))^2] \,. \tag{1.2}$$

As $\overline{R}_\varphi$ reports on the worst possible situation (the maximum risk), we may expect that a competitive data-dependent report $\delta(X)$ should improve on $\overline{R}_\varphi$ under the risk (1.2), that is, for all θ, $\delta(X)$ satisfies

$$\mathcal{R}(\theta, \varphi, \delta) \leq \mathcal{R}(\theta, \varphi, \overline{R}_\varphi) \,. \tag{1.3}$$

More generally, a reference loss estimator δ_0 will be dominated by a competitive estimator δ if, for all θ,

$$\mathcal{R}(\theta, \varphi, \delta) \leq \mathcal{R}(\theta, \varphi, \delta_0) \,, \tag{1.4}$$

with strict inequality for some θ.

Unlike the usual estimation setting where the quantity of interest is a function of the parameter θ, loss estimation involves a function of both θ and X (the data). This feature may make the statistical analysis more difficult but it is clear that the usual notions of minimaxity, admissibility, etc, and their methods of proof can be directly adapted to that situation. Also, although frequentist interpretability was evoked above, in case we would be interested in a Bayesian approach, it is easily seen that this approach would consist of the usual Bayes estimator φ_B of θ and the posterior loss $\delta_B(X) = E[L(\theta, \varphi_B)|X]$.

The problem of estimating a loss function has been considered by Sandved [41] who developed a notion of unbiased estimator of $L(\theta, \varphi(X))$ in various settings. However the underlying conditional approach traces back to Lehmann and Sheffé [34] who estimated the power of a statistical test. Kiefer, in a series of papers ([30], [31], [32]), developed conditional and estimated confidence theories through frequentist interpretability. A subjective Bayesian approach was compared by Berger ([3], [4], [5]) with the frequentist paradigm. Jonhstone [29] considered (in)admissibility of unbiased estimators of loss for the maximum likelihood estimator $\varphi_0(X) = X$ and for the James-Stein estimator $\varphi^{JS}(X) = \big(1 - (p-2)/||X||^2\big)\, X$ of a p-variate normal mean θ. For $\varphi_0(X) = X$, the unbiased estimator of the quadratic loss $L(\theta, \varphi_0(X)) = ||\varphi_0(X) - \theta||^2$, that is, the loss estimator δ_0 which satisfies, for all θ,

$$E_\theta[\delta_0] = E_\theta[L(\theta, \varphi_0(X))] = R(\theta, \varphi_0)\,, \tag{1.5}$$

is $\delta_0 = \overline{R}_\varphi = p$. Johnstone proved that (1.3) is satisfied with the competitive estimator $\delta(X) = p - 2(p-4)/||X||^2$ when $p \geq 5$, the risk difference between δ_0 and δ being expressed as $-4(p-4)^2 E_\theta[1/||X||^4]$. For the James-Stein estimator φ^{JS}, the unbiased estimator of loss is itself data-dependent and equal to $\delta_0^{JS}(X) = p - (p-2)^2/||X||^2$. Jonhstone showed that improvement on δ_0^{JS} can be obtained with $\delta^{JS}(X) = p - (p-2)^2/||X||^2 + 2p/||X||^2$ when $p \geq 5$, with strict inequality in (1.4) for all θ since the difference in risk between δ^{JS} and δ_0^{JS} equals $-4p^2 E_\theta[1/||X||^2]$.

In Section 2, we develop the quadratic loss estimation problem for a p-normal mean. After a review of the basic ideas, a new class of loss estimators is constructed in Subsection 2.1. In Subsection 2.2, we turn our focus on some interesting and surprising behavior of Bayesian assessments, this paradoxical result is illustrated in a general inadmissibility theorem. Section 3 is devoted to the case where the variance is unknown. Extensions to the spherical case are given in Section 4. In Subsection 4.1, we consider the general case of a spherically symmetric distribution around a fixed vector $\theta \in \mathbb{R}^p$ and in Subsection 4.2 these ideas are then generalized to the case where a residual vector is available. We conclude by mentioning a number of applied and theoretical developments of loss estimation not covered in this overview. The Appendix gives some necessary background material and technical results.

2 Estimating the quadratic loss of a p-normal mean with known variance

2.1 Dominating unbiased estimators of loss

Let X be a p-variate normal distributed $\mathcal{N}(\theta, I_p)$ random vector with unknown mean θ and identity covariance matrix I_p. To estimate θ, the observable X is itself a reference estimator (it is the maximum likelihood estimator (mle) and it is an unbiased estimator of θ) so that it is convenient to write any estimator of θ through X as $\varphi(X) = X + g(X)$, for a certain function g from $\mathbb{R}^p$ into $\mathbb{R}^p$. Under squared error loss $||\varphi(X) - \theta||^2$, the (quadratic) risk of φ is defined by

$$R(\theta, \varphi) = E_\theta[||\varphi(X) - \theta||^2] \tag{2.1}$$

where E_θ denotes the expectation with respect to $\mathcal{N}(\theta, I_p)$.

Clearly, the risk of the mle X equals p and in general $\varphi(X)$ will be a reasonable estimator only if its risk is finite. It is easy to see (Lemma A.1 in Appendix A.1) through Schwarz's inequality that this is the case as soon as

$$E_\theta[||g(X)||^2] < \infty\,, \tag{2.2}$$

which we will assume in the following (it can be also seen that this condition is in fact necessary to guarantee the risk finitness).

To improve on the mle X when $p \geq 3$ (that is, to have $R(\theta, \varphi) \leq p$), Stein [44] exhibited (under certain differentiability conditions that we recall below) an unbiased estimator of the risk of $\varphi(X)$, that is, a function $\delta_0(X)$ (depending only on X and not on θ) which verifies

$$R(\theta, \varphi) = E_\theta[\delta_0(X)]\,. \tag{2.3}$$

This statistic suggests a natural estimator of the loss $||\varphi(X) - \theta||^2$ since (2.3) implies that

$$E_\theta[||\varphi(X) - \theta||^2] = E_\theta[\delta_0(X)] \tag{2.4}$$

and hence is an unbiased estimator of the loss. Stein [44] proved more precisely that $\delta_0(X) = p + 2\,\mathrm{div}g(X) + ||g(X)||^2$ (where $\mathrm{div}g(X)$ stands for the divergence of $g(X)$, that is, $\mathrm{div}g(X) = \sum_{i=1}^p \partial_i g_i(X)$). One can see that δ_0 may change sign so that, as an estimator of loss (which is non negative), it cannot be completely satisfactory, and hence, is likely to be improved upon.

Any competitive loss estimator $\delta(X)$ can be written as $\delta(X) = \delta_0(X) - \gamma(X)$ for a certain function $\gamma(X)$ which can be interpreted as a correction to $\delta_0(X)$. Note that, for the mle (that is, if $g(X) = 0$), we may expect that an improvement on $\delta_0(X) = p$ would be obtained with a nonnegative function $\gamma(X)$ satisfying the requirement expressed by

Condition (1.3). Note also that, similarly to the finiteness risk condition (2.2), we will require that

$$E_\theta[\gamma^2(X)] < \infty \tag{2.5}$$

to assure that the risk of $\delta(X)$ is finite (see Appendix A.1).

Using straightforward algebra, the risk difference $\mathcal{D}(\theta, \varphi, \delta) = \mathcal{R}(\theta, \varphi, \delta) - \mathcal{R}(\theta, \varphi, \delta_0)$ simplifies in

$$\mathcal{D}(\theta, \varphi, \delta) = E_\theta[\gamma^2(X) - 2\,\gamma(X)\,\delta_0(X)] + 2\,E_\theta[\gamma(X)\,||\varphi(X) - \theta||^2]\,. \tag{2.6}$$

Conditions for which $\mathcal{D}(\theta, \varphi, \delta) \leq 0$ will be formulated after finding an unbiased estimate of the term $\gamma(X)\,||\varphi(X) - \theta||^2$ in the last expectation. We briefly review the flow of ideas of those techniques.

For a function g from $\mathbb{R}^p$ into $\mathbb{R}^p$, the Stein's identity (see Stein [44]) states that

$$E_\theta[(X - \theta)^t g(X)] = E_\theta[\mathrm{div} g(X)] \tag{2.7}$$

provided that these expectations exist. Here Stein specified that g was almost differentiable. Almost differentiability is needed to integrate shrinkage functions $g(X)$, intervening in the James-Stein estimators, of the form $g(X) = -a\,X\,/||X||^2$ which are not differentiable in the usual sense (such $g(X)$ explode at 0). This notion is equivalent (and it is of more common use in analysis) to the statement that g belongs to the Sobolev space $W^{1,1}_{loc}(\mathbb{R}^p)$ of weakly differentiable functions. That equivalence was noticed by Johnstone [29].

Recall that a locally integrable function γ from $\mathbb{R}^p$ into $\mathbb{R}$ is said to be weakly differentiable if, there exist p functions $h_1, \ldots, h_p$ locally integrable on $\mathbb{R}^p$ such that, for any $i = 1, \ldots, p$

$$\int_{\mathbb{R}^p} \gamma(x)\,\frac{\partial \varphi}{\partial x_i}(x)\,dx = -\int_{\mathbb{R}^p} h_i(x)\,\varphi(x)\,dx \tag{2.8}$$

for any infinitely differentiable function φ on $\mathbb{R}^p$ with compact support. The functions h_i are the i-th partial weak derivatives of γ. Their common notation is $\partial\gamma/\partial x_i$ and the vector $\nabla\gamma = (\partial\gamma/\partial x_1, \ldots, \partial\gamma/\partial x_p)^t$ is refereed to the weak gradient of γ.

Note that (2.8) usually holds when γ is continuously differentiable, that is, when $h_i = \partial\gamma/\partial x_i$, the standard partial derivative, is continuous. Thus, via (2.8), the extension to weak differentiability consists in a propriety of integration by parts with vanishing bracketed term. Naturally a function $g = (g_1, \ldots, g_p)$ from $\mathbb{R}^p$ into $\mathbb{R}^p$ is said to be weakly differentiable if each of its components g_j is weakly differentiable. In that case, the function $\mathrm{div} g = \sum_{i=1}^p \partial g_i/\partial x_i$ is refered to as the weak divergence of g; this is the operator intervening in the Stein's identity (2.7).

When dealing with an unbiased estimator of a quantity of the form $||X - \theta||^2\,\gamma(X)$ where γ is a function from $\mathbb{R}^p$ into $\mathbb{R}$, writing

$$||X - \theta||^2\,\gamma(X) = (X - \theta)^t\,(X - \theta)\,\gamma(X) \tag{2.9}$$

naturally leads to an iteration of Stein's identity (2.7) and involves twice weak differentiability of γ. This is of course defined through the weak differentiability of all the weak partial derivatives $\partial\gamma/\partial x_i$; these second weak partial derivatives are denoted by $\partial^2\gamma/\partial x_j\partial x_i$. Thus γ belongs to the Sobolev space $W^{2,1}_{loc}(\mathbb{R}^p)$ and $\Delta\gamma = \sum_{i=1}^p \partial^2\gamma/\partial x_i^2$ is referred to as the weak Laplacian of γ.

By (2.9) and (2.7), we have

$$\begin{aligned} E_\theta[||X-\theta||^2\,\gamma(X)] &= E_\theta[\mathrm{div}((X-\theta)^t\,\gamma(X))] \\ &= E_\theta[p\,\gamma(X) + (X-\theta)^t\,\nabla\gamma(X)] \end{aligned} \tag{2.10}$$

by property of the divergence operator. Then, applying again (2.7) to the last term in (2.10) gives

$$E_\theta[(X-\theta)^t\,\nabla\gamma(X)] = E_\theta[\mathrm{div}(\nabla\gamma(X)] = E_\theta[\Delta\gamma(X)] \tag{2.11}$$

by definition of the Laplacian operator. Finally, gathering (2.10) and (2.11), we obtain that

$$E_\theta[||X-\theta||^2\,\gamma(X)] = E_\theta[p\,\gamma(X) + \Delta\gamma(X)]\,. \tag{2.12}$$

We are now in a position to provide an unbiased estimator of the difference in risk $\mathcal{D}(\theta,\varphi,\delta)$ in (2.6). Its non positivity will be a sufficient condition for $\mathcal{D}(\theta,\varphi,\delta)\le 0$ and hence for δ to improve on δ_0. Indeed we have

$$\begin{aligned} ||\varphi(X)-\theta||^2 &= ||X+g(X)-\theta||^2 \\ &= ||g(X)||^2 + 2\,(X-\theta)^t\,g(X) + ||X-\theta||^2 \end{aligned}$$

so that, according to (2.7) and (2.12),

$$E_\theta[||\varphi(X)-\theta||^2\,\gamma(X)] = E_\theta[\gamma(X)\,||g(X)||^2 + 2\,\mathrm{div}(\gamma(X)\,g(X)) + p\,\gamma(X) + \Delta\gamma(X)]\,.$$

Therefore, as $\mathrm{div}(\gamma(X)\,g(X)) = \gamma(X)\,\mathrm{div}g(X) + \nabla\gamma(X)^t\,g(X)$ and as $\delta_0(X) = p + 2\,\mathrm{div}g(X) + ||g(X)||^2$, the risk difference $\mathcal{D}(\theta,\varphi,\delta)$ in (2.6) reduces to

$$\mathcal{D}(\theta,\varphi,\delta) = E_\theta[\gamma^2(X) + 4\,\nabla\gamma(X)^t\,g(X) + 2\,\Delta\gamma(X)]$$

so that a sufficient condition for $\mathcal{D}(\theta,\varphi,\delta)$ to be nonpositive is

$$\gamma^2(x) + 4\,\nabla\gamma(x)^t\,g(x) + 2\,\Delta\gamma(x) \le 0 \tag{2.13}$$

for any $x\in\mathbb{R}^p$.

How can one determine a "best" correction γ satisfying (2.13)? The following theorem provides a way to associate to the function g a suitable correction γ which satisfies (2.13) in the case where $g(x)$ is of the form $g(x) = \nabla m(x)/m(x)$ for a certain nonnegative function m. This is the case when φ is a Bayes estimator of θ related to a prior π, the function m being the corresponding marginal (see Brown [8]). Bock [7] shows that, through the choice of m, such estimators constitute a wide class of estimators of θ (which are called pseudo-Bayes estimators when the function m does not correspond to a true prior π).

Theorem 2.1 *Let m be a nonnegative function which is also superharmonic (respectively subharmonic on $\mathbb{R}^p$ such that $\nabla m/m \in W^{1,1}_{loc}(\mathbb{R}^p)$. Let ξ be a real valued function, strictly positive and strictly subharmonic (respectively superharmonic) on $\mathbb{R}^p$, and such that*

$$E_\theta\left[\left(\frac{\Delta\xi(X)}{\xi(X)}\right)^2\right] < \infty\,. \tag{2.14}$$

Assume also that there exists a constant $K > 0$ such that, for any $x \in \mathbb{R}^p$,

$$m(x) > K\frac{\xi^2(x)}{|\Delta\xi(x)|} \tag{2.15}$$

and let $K_0 = \inf_{x\in\mathbb{R}^p} m(x)\,\frac{|\Delta\xi(x)|}{\xi^2(x)}$.

Then the unbiased loss estimator δ_0 of the estimator φ of θ defined by $\varphi(X) = X + \nabla m(X)/m(X)$ is dominated by the estimator $\delta = \delta_0 - \gamma$, where the correction term γ is given, for any $x \in \mathbb{R}^p$ such that $m(x) \neq 0$, by

$$\gamma(x) = -\alpha\,\mathrm{sgn}(\Delta\xi(x))\,\frac{\xi(x)}{m(x)}\,, \tag{2.16}$$

as soon as $0 < \alpha < 2\,K_0$.

PROOF The domination condition will be shown by proving that the risk difference is less than zero. We only consider the case where m is superharmonic and ξ is strictly subharmonic, the case where m is subharmonic and ξ is strictly superharmonic being similar.

First note that the finiteness risk condition (2.5) is guaranteed by Condition (2.14) and the fact that (2.15) implies that, for any $x \in \mathbb{R}^p$,

$$\gamma^2(x) = \alpha^2\,\frac{\xi^2(x)}{m^2(x)} \leq \frac{\alpha^2}{K_0^2}\,\left(\frac{\Delta\xi(x)}{\xi(x)}\right)^2\,.$$

Also note that, for a shrinkage function g of the form $g(x) = \nabla m(x)/m(x)$, the left hand side of (2.13) can be expressed as

$$\mathcal{R}\gamma(x) = \gamma^2(x) + 2\left\{2\,\frac{\Delta(m(x)\,\gamma(x))}{m(x)} - \gamma(x)\frac{\Delta m(x)}{m(x)}\right\} \tag{2.17}$$

and hence, for γ in (2.16), as

$$\mathcal{R}\gamma(x) = \alpha^2\,\frac{\xi^2(x)}{m^2(x)} + 2\,\alpha\left\{-\frac{\Delta\xi(x)}{m(x)} + \frac{\xi(x)\,\Delta m(x)}{m^2(x)}\right\}\,. \tag{2.18}$$

Now, since m is superharmonic and ξ is positive, it follows from (2.18) that

$$\mathcal{R}\gamma(x) \leq \frac{\alpha}{m(x)} \left\{ \frac{\alpha\, \xi^2(x)}{m(x)} - 2\, \Delta\xi(x) \right\}$$

and hence, by subharmonicity of ξ, inequality (2.15) and definition of K_0, that

$$\mathcal{R}\gamma(x) < \frac{\alpha}{m(x)} \{\alpha - 2\, K_0\} \frac{\xi^2(x)}{m(x)}\,. \tag{2.19}$$

Finally, since $0 < \alpha < 2\, K_0$, Inequality (2.19) gives $\mathcal{R}\gamma(x) < 0$, which is the desired result. □

As an example, consider $m(x) = 1/||x||^{p-2}$, that is, the fundamental harmonic function which is superharmonic on the entire space $\mathbb{R}^p$ (see Du Plessis [39]). Then we have $\nabla m(x)/m(x) = -(p-2)/||x||^2$ and $\varphi(X)$ is the James-Stein estimator whose unbiased estimator of loss is $\delta_0(X) = p - (p-2)^2/||X||^2$. First note that $\nabla m/m \in W^{1,1}_{loc}(\mathbb{R}^p)$ for $p \geq 3$. Now choosing, for any $x \neq 0$, the function $\xi(x) = 1/||x||^p$ gives rise to $\Delta\xi(x) = 2\, p/||x||^{p+2} > 0$ and hence to

$$\frac{\xi^2(x)}{|\Delta\xi(x)|} = \frac{1}{2\, p} \frac{1}{||x||^{p-2}}\,,$$

which means that Condition (2.15) is satisfied with $K < 2\, p$. Also we have

$$\left(\frac{\Delta\xi(x)}{\xi(x)} \right)^2 = \frac{4\, p^2}{||x||^4}$$

which implies that Condition (2.14) is satisfied for $p \geq 5$. Now it is clear that the constant K_0 is equal to $2\, p$ and that the correction term γ in (2.16) equals, for any $x \neq 0$, $\gamma(x) = -\alpha/||x||^2$. Finally, Theorem 2.1 guarantees that an improved loss estimator over the unbiased estimator of loss $\delta_0(X)$ is $\delta(X) = \delta_0(X) + \alpha/||x||^2$ for $0 < \alpha < 4\, p$, which is Johnstone's result [29] for the James-Stein estimator.

Similarly Johnstone's result for $\varphi(X) = X$ can be constructed with $m(x) = 1$ (which is both subharmonic and superharmonic) and with the choice of the superharmonic function $\xi(x) = 1/||x||^2$, for which $K_0 = 2\,(p-4)$, so that $\delta(x) = p - \alpha/||x||^2$ dominates p for $0 < \alpha < 4\,(p-4)$.

We have shown that the unbiased estimator of loss can be dominated. Often one may wish to add a frequentist-validity constraint to a loss estimation problem. Specifically in our problem, the frequentist-validity constraint for some estimator δ would be $E_\theta[\delta(X)] \geq E_\theta[\delta_0(X)]$ for all θ. Kiefer [32] suggested that conditional and estimated confidence assessments should be conservatively biased, that is, the average reported loss should be greater than the average actual loss. Under such a frequentist-validity condition Lu and Berger [37] give improved loss estimators for several of the most important Stein-type estimators. One of their estimators is a generalized Bayes estimator, suggesting that Bayesians and frequentists can potentially agree on a conditional assessment of loss.

2.2 Dominating the posterior risk

In the previous sections, we have seen that the unbiased estimator of loss should be often dismissed since it can be dominated. When a (generalized) Bayes estimator of θ is available, incorporating the same prior information for estimating the loss of this Bayesian estimator is coherent, and we may expect that the corresponding Bayes estimator is a good candidate to improve on the unbiased estimator of loss. However, somewhat surprisingly, Fourdrinier and Strawderman [19] found that, in the normal setting considered in Section 2, the unbiased estimator often dominates the corresponding generalized Bayes estimator of loss for priors which give minimax estimators in the original point estimation problem. They also give a general inadmissibility result for a generalized Bayes estimator of loss. While much of their focus is on pseudo-Bayes estimators, in this section, we essentially present their results on generalized Bayes estimators.

For a given generalized prior π, we denote the generalized marginal by m and the generalized Bayes estimator of θ by

$$\varphi_m(X) = X + \frac{\nabla m(X)}{m(X)}\,. \tag{2.20}$$

Then (see Stein [44]) the unbiased estimator of risk of $\varphi_m(X)$ is

$$\delta_0(X) = p + 2\frac{\Delta m(x)}{m(X)} - \frac{\|\nabla m(X)\|^2}{m^2(X)} \tag{2.21}$$

while the posterior risk of $\varphi_m(X)$ is

$$\delta_m(X) = p + \frac{\Delta m(X)}{m(X)} - \frac{\|\nabla m(X)\|^2}{m^2(X)}\,. \tag{2.22}$$

Domination of $\delta_0(X)$ over $\delta_m(X)$ is obtained thanks to the fact that their risk admits $(\Delta m(X)/m(X))^2 - 2\,\Delta^{(2)}m(X)/m(X)$ as an unbiased estimator of their risk difference, that is,

$$\mathcal{R}(\theta, \varphi_m, \delta_0) - \mathcal{R}(\theta, \varphi_m, \delta_m) = E_\theta\left[\left(\frac{\Delta m(X)}{m(X)}\right)^2 - 2\,\frac{\Delta^{(2)}m(X)}{m(X)}\right] \tag{2.23}$$

where $\Delta^{(2)}m = \Delta(\Delta m)$ is the bi-Laplacian of m (see [19]). Thus the above domination will occur as soon as

$$\left(\frac{\Delta m(X)}{m(X)}\right)^2 - 2\,\frac{\Delta^{(2)}m(X)}{m(X)} \leq 0\,. \tag{2.24}$$

Applicability of that last condition is underlined by the remarkable fact that if the prior π satisfies (2.24), that is, if

$$\left(\frac{\Delta \pi(\theta)}{\pi(\theta)}\right)^2 - 2\,\frac{\Delta^{(2)}\pi(\theta)}{\pi(\theta)} \leq 0\,, \tag{2.25}$$

then (2.24) is satisfied for the marginal m.

As an example, [19] considers $\pi(\theta) = (\|\theta\|^2/2+a)^{-b}$ (where $a \geq 0$ and $b \geq 0$) and show that, if $p \geq 2(b+3)$ then (2.25) holds and hence δ_u dominates δ_m. Since π is integrable if and only if $b > \frac{p}{2}$ (for $a > 0$), the prior π is improper whenever this condition for domination of δ_u over δ_m holds. Of course, whenever π is proper, the Bayes estimator δ_m is admissible provided its Bayes risk is finite.

Inadmissibility of the generalized Bayes loss estimator is not exceptional. Thus, in [19], the following general inadmissibility result is given; its proof is parallel to the proof of Theorem 2.1.

Theorem 2.2 *Let m be a nonnegative function such that $\nabla m/m \in W^{1,1}_{loc}(\mathbb{R}^p)$. Let ξ be a real valued function satisfying the conditions of Theorem 2.1. Then δ_m is inadmissible and a class of dominating estimators is given by*

$$\delta_m(X) + \alpha \text{ sgn } (\Delta\xi(X))\frac{\xi(X)}{m(X)} \quad \textit{for } \; 0 < \alpha < 2\, K_0\,.$$

Note that, unlike Theorem 2.1, neither the superharmonicity condition nor the subharmonicity condition on m are needed. Note also that Theorem 2.2 gives conditions of improvement on δ_m while Theorem 2.1 looks for improvements on δ_0. As we saw that, often, δ_0 dominates δ_m. So it is not surprising that the proof of the two theorems are parallel; more precisely, it suffices to suppress, in the proof of Theorem 2.1), the superharmonicity (or subharmonicity) condition on m to obtain the proof of Theorem 2.2.

In [19], it is suggested that the inadmissibility of the generalized Bayes (or pseudo-Bayes) estimator is due to the fact that the loss function $(\delta(x) - \|\varphi(x) - \theta\|^2)^2$ may be inappropriate. The possible deficiency of this loss is illustrated by the following simple result concerning estimation of the square of a location parameter in $\mathbb{R}^1$.

Suppose $X \in \mathbb{R}^1 \sim f\left((X-\theta)^2\right)$ such that $E_\theta[X^4] < \infty$. Consider estimation of θ^2 under loss $(\delta - \theta^2)^2$. The generalized Bayes estimator δ_π of θ^2 with respect to the uniform prior $\pi(\theta) \equiv 1$ is given by

$$\delta_\pi(X) = \frac{\int \theta^2 f\left((X-\theta)^2\right) d\theta}{\int f\left((X-\theta)^2\right) d\theta} = X^2 + E_0\left[X^2\right].$$

Since this estimator has constant bias $2\,E_0[X^2]$, it is dominated by the unbiased estimator $X^2 - E_0[X^2]$ (the risk difference is $4\,(E_0[X^2])^2$). Hence δ_π is inadmissible for any $f(\cdot)$ such that $E_\theta[X^4] < \infty$.

2.3 Examples of improved estimators

In this subsection, we give some examples of Theorems 2.1 and 2.2. The only example up to this point of an improved estimator over the unbiased estimator of loss $\delta_0(X)$ is $\delta(X) = \delta_0(X) + \alpha/||x||^2$ for $0 < \alpha < 4\,p$, which is Johnstone's result [29]. Although the shrinkage factor in Theorems 2.1 and 2.2 are the same, in the examples below we will only focus on improvements of posterior risk.

As an application of Theorem 2.2, let $\xi_b(x) = \left(\|x\|^2 + a\right)^{-b}$ (with $a \geq 0$ and $b \geq 0$). It can be shown that we have $\Delta\xi_b(x) < 0$ for $a \geq 0$ and $0 < 2\,(b+1) < p$. Also $\Delta\xi_b(x) > 0$ if $a = 0$ and $2(b+1) > p$. Furthermore

$$\frac{\xi_b^2(x)}{|\Delta\xi_b(x)|} = \frac{1}{2\,b\,\left|p - 2(b+1)\frac{\|x\|^2}{\|x\|^2+a}\right|}\,\frac{1}{(\|x\|^2+a)^{b-1}}\,.$$

a) Suppose that $0 < 2\,(b+1) < p$ and $a \geq 0$. Then

$$\frac{\xi_b^2(x)}{|\Delta\xi_b(x)|} \leq \frac{1}{2\,b(p-2(b+1))}\,\frac{1}{(\|x\|^2+a)^{b-1}}$$

and $E_\theta\left[\left(\Delta\xi_b(X)/\xi_b(X)\right)^2\right] < \infty$ since it is bonded from above by a quantity proportional to $E_\theta\left[(\|X\|^2+a)^{-2}\right]$, which is finite for $a > 0$ or for $a = 0$ and $p > 4$.

Suppose that $m(x)$ is greater than or equal to some multiple of $\left(\|x\|^2+a\right)^{1-b}$ or equivalently

$$m(x) \geq \frac{k}{2\,b\,(p-2\,(b+1))}\,\frac{1}{\left(\|x\|^2+a\right)^{b-1}} \tag{2.26}$$

for some $k > 0$. Theorem 2.2 implies that $\delta_m(X)$ is inadmissible and is dominated by

$$\delta_m(X) - \frac{\alpha}{m(X)(\|X\|^2+a)^b}$$

for $0 < \alpha < 4\ b(p-2(b+1))\inf_{x\in\mathbb{R}^p}(m(x)(\|x\|^2+a)^{b-1})$. Note that the improved estimators shrink towards 0.

Suppose, for example, that $m(x) \equiv 1$. Then (2.26) is satisfied for $b \geq 1$. Here $\varphi_m(X) = X$ and $\delta_m(X) = p$. Choosing $b = 1$, an improved class of estimators is given by $p - \frac{\alpha}{\|X\|^2+a}$ for $0 < \alpha < 4\,(p-4)$. The case $a = 0$ is equivalent to Johnstone's result for this marginal.

b) Suppose that $2\,(b+1) > p > 4$ and $a = 0$. Then

$$\frac{\xi_b^2(x)}{|\Delta\xi_b(x)|} = \frac{1}{2\ b(2(b+1)-p)}\ \frac{1}{\|x\|^{2(b-1)}}.$$

A development similar to the above implies that, when $m(x)$ is greater than or equal to some multiple of $\|x\|^{2(1-b)}$, an improved estimator is

$$\delta_m(X) + \frac{\alpha}{m(X)\|X\|^{2b}}$$

for $0 < \alpha < 4\ b(2\ (b+1)-p) \inf_{x\in\mathbb{R}^p}\left(m(x)\|x\|^{2(b-1)}\right)$.

Note that, in this case, the correction term is positive and hence the estimators expands away from 0. Note also that this result only works for $a = 0$ and hence applies to pseudo-marginals which are unbounded in a neighborhood of 0. Since all marginals corresponding to a generalized prior π are bounded, this result can never apply to generalized Bayes procedures but only to pseudo-Bayes procedures.

Suppose, for example, that $m(x) = \|x\|^{2-p}$. Here $\varphi_m(X) = \left(1 - \frac{p-2}{\|x\|^2}\right) X$ is the James-Stein estimator and $\delta_m(X) = p - \frac{(p-2)^2}{\|X\|^2}$. In particular, the above applies for $b - 1 = \frac{p-2}{2}$, that is, for $b = \frac{p}{2} > \frac{p-2}{2}$. An improved estimator is given by $\delta_m(X) + \frac{\gamma}{\|X\|^2}$ for $0 < \gamma < 4\ p$. This again agrees with Johnstone's result for James-Stein estimators.

3 Estimating the quadratic loss of a p-normal mean with unknown variance

In Section 2 is was tacitly assumed that the covariance matrix was known and equal to the identity matrix I_p. Typically, this covariance is unknown and should be estimated. In the case where it is of the form $\sigma^2 I_p$ with σ^2 unknown, Wan and Zou [48] show that, for the invariant loss $||\varphi(X) - \theta||^2/\sigma^2$, Johnstone's result [29] can be extended when estimating the loss of the James-Stein estimator. In fact, the general framework considered in Section 2 can be extended to the case where σ^2 is unknown, and we show that a condition parallel to Condition (2.13) can be found.

Before stating the main result for the unknown variance case, we need an extension of Stein's identity involving their sample variance.

Lemma 3.1 *Let $X \sim \mathcal{N}(\theta, \sigma^2 I_p)$ and let S be a nonnegative random variable independent of X such that $S \sim \sigma^2\chi_k^2$. Denoting by E_{θ,σ^2} the expectation with respect to the joint distribution of (X, S), we have, provided the corresponding expectations exist, the following two results:*

(i) if $g(x,s)$ is a function from $\mathbb{R}^p \times \mathbb{R}_+$ into $\mathbb{R}^p$ such that, for any $s \in \mathbb{R}_+$, $g(\cdot,s)$ is weakly differentiable then

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

where $\mathrm{div}_x g(x,s)$ is the divergence of $g(x,s)$ with respect to x;

(ii) if $h(x,s)$ is a function from $\mathbb{R}^p \times \mathbb{R}_+$ into $\mathbb{R}$ such that, for any $s \in \mathbb{R}_+$, $h(\cdot,s)$ is weakly differentiable then

$$E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}\,h(X,S)\right] = E_{\theta,\sigma^2}\left[2\,\frac{\partial}{\partial S}h(X,S) + (k-2)S^{-1}\,h(X,S)\right].$$

PROOF Part (i) is Stein's lemma 1981 (cf. [44]). Part (ii) can be seen as a particular case of Lemma 1 (ii) (established for elliptically symmetric distributions) of Fourdrinier et al. [20], although we will present a direct proof. The joint distribution of (X,S) can be viewed as resulting, in the setting of the canonical form of the general linear model, from the distribution of $(X,U) \sim \mathcal{N}((\theta,0),\sigma^2 I_{p+k})$ with $S = ||U||^2$. Then we can write

$$\begin{aligned} E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}\,h(X,S)\right] &= E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}\,U^t\,\frac{U}{||U||^2}\,h(X,||U||^2)\right] \\ &= E_{\theta,\sigma^2}\left[\mathrm{div}_U\left(\frac{U}{||U||^2}\,h(X,||U||^2)\right)\right] \end{aligned}$$

according to part (i). Hence, expending the divergence term, we have

$$\begin{aligned} E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}\,h(X,S)\right] &= E_{\theta,\sigma^2}\left[\frac{k-2}{||U||^2}\,h(X,||U||^2) + \frac{U^t}{||U||^2}\,\nabla_U h(X,||U||^2)\right] \\ &= E_{\theta,\sigma^2}\left[\frac{k-2}{S}\,h(X,S) + 2\,\frac{\partial}{\partial S}h(X,S)\right] \end{aligned}$$

since

$$\nabla_U h(X,||U||^2) = 2\,\frac{\partial}{\partial S}h(X,S)\Big|_{S=||U||^2} U\,.$$

□

The following theorem provides an extension of results in Section 2 to the setting of an unknown variance. The necessary conditions to insure the finitness of the risks are given in Appendix A.1.

Theorem 3.1 *Let $X \sim \mathcal{N}(\theta,\sigma^2 I_p)$ where θ and σ^2 are unknown and $p \geq 5$ and let S be a nonnegative random variable independent of X and such that $S \sim \sigma^2\chi^2_k$. Consider an estimator of θ of the form $\varphi(X,S) = X + S\,g(X,S)$ with $E_{\theta,\sigma^2}[S^2\,||g(X,S)||^2] < \infty$, where E_{θ,σ^2} denotes the expectation with respect to the joint distribution of (X,S).*

Then an unbiased estimator of the loss $||\varphi(X,S)-\theta||^2/\sigma^2$ *is*

$$\delta_0(X,S) = p + S\left\{(k+2)\,||g(X,S)||^2 + 2\,\mathrm{div}_X g(X,S) + 2\,S\frac{\partial}{\partial S}||g(X,S)||^2\right\}. \quad (3.1)$$

Its risk $\mathcal{R}(\theta,\sigma^2,\varphi,\delta_0) = E_{\theta,\sigma^2}[(\delta_0(X,S) - ||\varphi(X,S)-\theta||^2/\sigma^2))^2]$ *is finite as soon as* $E_{\theta,\sigma^2}[S^2\,||g(X,S)||^4] < \infty$, $E_{\theta,\sigma^2}[(S\,\mathrm{div}_X g(X,S))^2] < \infty$ *and* $E_{\theta,\sigma^2}\left[\left(S^2\,\frac{\partial}{\partial S}||g(X,S)||\right)^2\right] < \infty$.

Furthermore, for any function $\gamma(X)$ *such that* $E_{\theta,\sigma^2}[\gamma^2(X)] < \infty$, *the risk difference* $\mathcal{D}(\theta,\sigma^2,\varphi,\delta) = \mathcal{R}(\theta,\sigma^2,\varphi,\delta) - \mathcal{R}(\theta,\sigma^2,\varphi,\delta_0)$ *between the estimators* $\delta(X,S) = \delta_0(X,S) - S\,\gamma(X)$ *and* $\delta_0(X,S)$ *is given by*

$$E_{\theta,\sigma^2}\left[S^2\left\{\gamma^2(X) + \frac{2}{k+2}\,\Delta\gamma(X) + 4\,g^t(X,S)\nabla\,\gamma(X) + 4\,\gamma(X)\,||g(X,S)||^2\right\}\right], \quad (3.2)$$

so that a sufficient condition for $\mathcal{D}(\theta,\sigma^2,\varphi,\delta)$ *to be non positive, and hence for* $\delta(X,S)$ *to improve on* $\delta_0(X,S)$, *is*

$$\gamma^2(x) + \frac{2}{k+2}\,\Delta\gamma(x) + 4\,g^t(x,s)\,\nabla\gamma(x) + 4\,\gamma(x)\,||g(x,s)||^2 \le 0 \quad (3.3)$$

for any $x \in \mathbb{R}^p$ *and any* $s \in \mathbb{R}_+$.

PROOF According to the expression of $\varphi(X,S)$, its risk $R(\theta,\varphi)$ is the expectation of

$$\frac{1}{\sigma^2}\,||X-\theta||^2 + 2\,\frac{S}{\sigma^2}\,(X-\theta)^t g(X,S) + \frac{S^2}{\sigma^2}\,||g(X,S||^2\,. \quad (3.4)$$

Clearly $E_{\theta,\sigma^2}\left[\sigma^{-2}\,||X-\theta||^2\right] = p$ and Lemma 3.1 (i) and (ii) express respectively that

$$E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}(X-\theta)^t g(X,S)\right] = E_{\theta,\sigma^2}[\mathrm{div_X} g(X,S)]$$

and, with $h(x,s) = s^2\,||g(x,s)||^2$, that

$$E_{\theta,\sigma^2}\left[\frac{S^2}{\sigma^2}\,||g(X,S)||^2\right] = E_{\theta,\sigma^2}\left[S\left\{(k+2)\,||g(X,S)||^2 + 2\,S\,\frac{\partial}{\partial S}||g(X,S)||^2\right\}\right].$$

Therefore $R(\theta,\varphi) = E_{\theta,\sigma^2}[\delta_0(X,S)]$ with $\delta_0(X,S)$ given in (3.1), which means that $\delta_0(X,S)$ is an unbiased estimator of the loss $||\varphi(X,S)-\theta||^2/\sigma^2$. The fact that the risk $\mathcal{R}(\theta,\sigma^2,\varphi,\delta_0)$ of $\delta_0(X)$ is finite is shown in Lemma A.1.

Consider now the finiteness of the risk of the alternative loss estimator $\delta(X,S) = \delta_0(X,S) - S\,\gamma(X)$. It is easily seen that its difference in loss $d(\theta,\sigma^2,X,S)$ with $\delta_0(X,S)$ can be written as

$$\begin{aligned} d(\theta,\sigma^2,X,S) &= \Big(\delta_0(X,S) - \frac{1}{\sigma^2}||\varphi(X)-\theta||^2 - S\,\gamma(X)\Big)^2 - \Big(\delta_0(X,S) - \frac{1}{\sigma^2}||\varphi(X)-\theta||^2\Big)^2 \\ &= S^2\,\gamma^2(X) - 2\,S\,\gamma(X)\,(\delta_0(X,S) - \frac{1}{\sigma^2}||\varphi(X)-\theta||^2)\,. \end{aligned} \tag{3.5}$$

Hence, since $E_{\theta,\sigma^2}[||\varphi(X,S)-\theta||^2/\sigma^2] < \infty$ as the risk of the estimator $\varphi(X,S)$, the condition $E_{\theta,\sigma^2}[\gamma^2(X)] < \infty$ insures that the expectation of the loss in (3.5), that is, the risk difference $\mathcal{D}(\theta,\sigma^2,\varphi,\delta)$ is finite. Then $\mathcal{R}(\theta,\sigma^2,\varphi,\delta) < \infty$ since $\mathcal{R}(\theta,\sigma^2,\varphi,\delta_0) < \infty$.

We now express the risk difference $\mathcal{D}(\theta,\sigma^2,\varphi,\delta) = E_{\theta,\sigma^2}[d(\theta,\sigma^2,X,S)]$. Using (3.1) and expanding $||\varphi(X,S)-\theta||^2/\sigma^2$ give that $d(\theta,\sigma^2,X,S)$ in (3.5) can be written as $d(\theta,\sigma^2,X,S) = A(X,S) + B(\theta,\sigma^2,X,S)$ where

$$\begin{aligned} A(X,S) = S^2\,\gamma^2(X) &- 2\,p\,S\,\gamma(X) - 2\,(k+2)\,S^2\,\gamma(X)\,||g(X,S)||^2 \\ &- 4\,S^2\,\gamma(X)\,\mathrm{div}_X g(X,S) - 4\,S^3\,\gamma(X)\,\frac{\partial}{\partial S}||g(X,S)||^2 \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} B(\theta,\sigma^2,X,S) = 2\,\frac{S^3}{\sigma^2}\,\gamma(X)\,||g(X,S)||^2 &+ 2\,\frac{S}{\sigma^2}\,\gamma(X)\,||X-\theta||^2 \\ &+ 4\,\frac{S^2}{\sigma^2}\,\gamma(X)\,(X-\theta)^t g(X,S)\,. \end{aligned} \tag{3.7}$$

Through Lemma 3.1 (ii) with $h(x,s) = 2\,\frac{s^3}{\sigma^2}\,\gamma(x)\,||g(x,s)||^2$, the expectation of the first term in the right hand side of (3.7) equals

$$\begin{aligned} E_{\theta,\sigma^2}\Big[2\,\frac{S^3}{\sigma^2}\,\gamma(X)\,||g(X,S)||^2\Big] = E_{\theta,\sigma^2}\big[&2\,(k+4)\,S^2\,\gamma(X)\,||g(X,S)||^2 + \\ &4\,S^3\,\gamma(X)\,\frac{\partial}{\partial S}||g(X,S)||^2\Big]\,. \end{aligned} \tag{3.8}$$

Also a reiterated application of Lemma 3.1 (i) to the expectation of the second term in the right hand side of (3.7) allows to write

$$\begin{aligned} E_{\theta,\sigma^2}\Big[2\,\frac{S}{\sigma^2}\,\gamma(X)\,||X-\theta||^2\Big] &= E_{\theta,\sigma^2}[2\,\frac{1}{\sigma^2}\,(X-\theta)^t\,S\,\gamma(X)\,(X-\theta)] \\ &= E_{\theta,\sigma^2}[2\,\mathrm{div}_{\mathrm{X}}\{S\,\gamma(X)\,(X-\theta)\}] \\ &= E_{\theta,\sigma^2}[2\,p\,S\,\gamma(X) + 2\,S\,(X-\theta)^t\nabla\gamma(X)] \\ &= E_{\theta,\sigma^2}[2\,p\,S\,\gamma(X) + 2\,\sigma^2\,S\,\Delta\gamma(X)] \end{aligned}$$

which, as $S \sim \sigma^2 \chi^2_k$ entails that $E[S^2/(k+2)] = E[\sigma^2 S]$ and as S is independent of X, gives

$$E_{\theta,\sigma^2}\left[2\,\frac{S}{\sigma^2}\,\gamma(X)\,||X-\theta||^2\right] = E_{\theta,\sigma^2}\left[2\,p\,S\,\gamma(X) + 2\,\frac{S^2}{k+2}\,\Delta\gamma(X)\right]. \tag{3.9}$$

As for the third term in the right hand side of (3.7), its expectation can also also be expressed using Lemma 3.1 (i) as

$$\begin{aligned} E_{\theta,\sigma^2}\left[4\,\frac{S^2}{\sigma^2}\,\gamma(X)\,(X-\theta)^t g(X,S)\right] &= E_{\theta,\sigma^2}[4\,S^2\,\mathrm{div_X}\{\gamma(X)\,g(X,S)\}] \\ &= E_{\theta,\sigma^2}[4\,S^2\,\gamma(X)\,\mathrm{div_X}\{g(X,S)\} + 4\,S^2\,g(X,S)^t\,\nabla\gamma(X)] \end{aligned} \tag{3.10}$$

by propriety of the divergence. Finally, gathering (3.8), (3.9) and (3.10) yields an expression of (3.7) which, with (3.6), gives the integrand term of (3.2), which is the desired result. □

As an example, consider the James-Stein estimator

$$\varphi^{JS}(X,S) = X - \frac{p-2}{k+2}\frac{S}{||X||^2}X\,.$$

Here the shrinkage factor is the product of a function of S by a function of X so that, through routine calculation, the unbiased estimator of loss is

$$\delta_0(X,S) = p - \frac{(p-2)^2}{k+2}\frac{S}{||X||^2}\,.$$

For a correction of the form $\gamma(x) = -d/||x||^2$ with $d \geq 0$, it is easy to check that the expression in (3.3) equals

$$d^2 + 4\,\frac{p-4}{k+2}\,d - 8\,\frac{p-2}{k+2}\,d - 4\,\left(\frac{p-2}{k+2}\right)^2\,d = d\left(d - \frac{4}{k+2}\left[p + \frac{(p-2)^2}{k+2}\right]\right)$$

which is negative for $0 < d < \frac{4}{k+2}\left[p + \frac{(p-2)^2}{k+2}\right]$ and gives domination of $p - \frac{(p-2)^2}{k+2}\frac{S}{||X||^2} + \frac{d}{||x||^2}$ over $p - \frac{(p-2)^2}{k+2}\frac{S}{||X||^2}$. This condition recovers the result of Wan and Zou [48] who considered the case $d = \frac{2}{k+2}\left[p + \frac{(p-2)^2}{k+2}\right]$.

4 Extensions to the spherical case

4.1 Estimating the quadratic loss of the mean of a spherical distribution

In the previous sections the loss estimation problem was considered for the normal distribution setting. The normal distribution has been generalized in two important directions, first as a special case of the exponential family and secondly as a spherically symmetric distribution. In this section we will consider the latter. There are a variety of equivalent definitions and characterizations of the class of spherically symmetric distributions, a comprehensive review is given by [17]. We will use the representation of a random variable from a spherically symmetric distribution, $X = (X_1, \ldots, X_p)^t$, as $X \stackrel{d}{=} RU^{(p)} + \theta$, where $R = ||X - \theta||$ is a random radius, $U^{(p)}$ is a uniform random variable on the p-dimensional unit sphere, where R and $U^{(p)}$ are independent. In such situation, the distribution of X is said spherically symmetric around θ and we write $X \sim SS_p(\theta)$. We also extend, in Subsection 4.2, these results to the case where the distribution of X is spherically symmetric and when a residual vector U is available (which allows an estimation of the variance factor σ^2).

Assume $X \sim SS_p(\theta)$ and suppose we wish to estimate $\theta \in \mathbb{R}^p$ by a decision rule $\delta(X)$ using quadratic loss. Suppose that we also use quadratic loss to assess the accuracy of loss estimate $\delta(X)$, then the risk of this loss estimate is given by (1.2). In [23], the problem of estimating the loss when $\varphi(X) = X$ is the estimate of the location parameter θ is considered. The estimate φ is the least squares estimator and is minimax among the class of spherically symmetric distributions with bounded second moment. Furthermore if one assumes the density of X exists and is unimodal, then φ is also the maximum likelihood estimator.

The unbiased constant estimate of the loss $||X - \theta||^2$ is $\delta_0 = E_\theta[R^2]$. Note that δ_0 is independent of θ, since $E_\theta[||X - \theta||^2] = E_0[||X||^2]$. Fourdrinier and Wells [23] show that the unbiased estimator δ_0 can be dominated by $\delta_0 - \gamma$, where γ is a particular superharmonic function for the case where the sampling distribution is a scale mixture of normals and in a more general spherical case.

The development of the results depends on some interesting extensions of the classical Stein identities in (2.7) and (2.12) to the general spherical setting. Since the distribution of X, say P_θ, is spherically symmetric around θ, for every bounded function f, we have $E_\theta[f] = E\, E_{R,\theta}[f] = \int_{\mathbb{R}_+} E_{R,\theta}[f]\rho(dR)$, where ρ is the distribution of the radius, namely the distribution of the norm $||X - \theta||$ under P_θ and where E and $E_{R,\theta}$ denotes respectively the expectation with respect to the radial distribution and uniform distribution $U_{R,\theta}$ on the sphere $S_{R,\theta} = \{x \in \mathbb{R}^p / \ ||x - \theta|| = R\}$ of radius R and center θ. To deduce the various risk domination results it suffices to work conditionally on the radius, that is to say to replace P_θ by $U_{R,\theta}$ in the risk expressions.

Denote $\sigma_{R,\theta}$ as the area measure on $S_{R,\theta}$. Therefore, for every Borel measurable set A, $U_{R,\theta}(A) = \sigma_{R,\theta}(A)/\sigma(S_{R,\theta}) = \Gamma(p/2)\sigma_{R,\theta}(A)/2\pi^{p/2}R^{p-1}$. Define the volume measure $\tau_{R,\theta}$ on the ball $B_{R,\theta} = \{x \in E/\, ||x-\theta|| \leq R\}$ of radius R and center θ and denote the uniform distribution on $B_{R,\theta}$ as $V_{R,\theta}$. Hence, for every Borel measurable set A, $V_{R,\theta}(A) = \tau_{R,\theta}(A)/\tau_{R,\theta}(B_{R,\theta}) = p\Gamma(p/2)\tau_{R,\theta}(A)/2\pi^{p/2}R^{p}$. Suppose γ is a weakly differentiable vector valued function, then by applying the Divergence Theorem for weakly differentiable functions to the definition of the expectation we have

$$\begin{aligned} E_\theta[(X-\theta)^t\gamma(X) \mid ||X-\theta|| = R] &= \int_{S_{R,\theta}} (x-\theta)^t\,\gamma(x)\,U_{R,\theta}(dx) \\ &= \frac{R}{\sigma_{R,\theta}(S_{R,\theta})}\int_{B_{R,\theta}} \operatorname{div}\gamma(x)\,dx. \end{aligned} \tag{4.1}$$

If γ is a real valued function then it follows from (4.1) and the product rule applied to the vector valued function $(x-\theta)\,\gamma(x)$ that

$$\begin{aligned} E_\theta[||X-\theta||^2\gamma(X) \mid ||X-\theta|| = R] &= \int_{S_{R,\theta}} (x-\theta)^t\,(x-\theta)\,\gamma(x)\,U_{R,\theta}(dx) \\ &= \frac{R}{\sigma_{R,\theta}(S_{R,\theta})}\int_{B_{R,\theta}} \left[p\,\gamma(x) + (x-\theta)^t\,\nabla\gamma(x)\right]dx\,. \end{aligned} \tag{4.2}$$

Our first extension of Theorem 2.1 is to the class of spherically symmetric distributions that are scale mixtures of normal distributions. Well known examples in the class of densities include the double exponential, multivariate t-distribution (hence, the multivariate Cauchy distribution). Let $\phi(x;\theta,I)$ be the probability density function of a random vector X with a normal distribution with mean vector θ and identity covariance matrix. Suppose that there is a probability measure on $\mathbb{R}_+$ such that the probability density function p_θ may be expressed as

$$p_\theta(x|\theta) = \int_0^\infty \phi(x;\theta,I/t)G(dt). \tag{4.3}$$

One can think of T being a random variable with distribution G, the conditional distribution of X given $T = t, X \mid T = t$, is $N_p(\theta, I/t)$. This class contains heavy tailed distributions, possibly with no moments. It is well known (see [17]) that, if a spherical distribution has a density p_θ, it is of the form $p_\theta(x) = g(||x-\theta||^2)$ for a measurable positive function g (called the generating function).

In the scale mixture of normals setting the unbiased estimate, δ_0, of risk equals

$$E[R^2] = E_\theta[||X-\theta||^2] = p\int_0^\infty t^{-1}G(dt).$$

It is easy to see that the risk of the unbiased estimator δ_0 is finite if and only if $E_\theta[||X-\theta||^4] < \infty$, which holds if

$$\int_0^\infty t^{-2} G(dt) < \infty. \tag{4.4}$$

The main theorem in [23] is the following domination result of an improved estimator of loss over the unbiased loss estimator.

Theorem 4.1 *Assume the distribution of X is a scale mixture of a normal random variables as in (4.3) such that (4.4) is satisfied and such that*

$$\int_{\mathbb{R}_+} t^{p/2}\, G(dt) < \infty\,. \tag{4.5}$$

Also, assume that the shrinkage function γ is twice weakly differentiable on $\mathbb{R}^p$ and satisfies $E_\theta[\gamma^2] < \infty$, for every $\theta \in \mathbb{R}^p$. Then a sufficient condition for $\delta_0 - \gamma$ to dominate δ_0 is that γ satisfies the differential inequality

$$k\,\Delta\gamma + \gamma^2 < 0 \quad \text{with} \quad k = 2\,\frac{\int_{\mathbb{R}_+} t^{p/2}\, G(dt)}{\int_{\mathbb{R}_+} t^{p/2-2}\, G(dt)}\,. \tag{4.6}$$

As an example let $\gamma(x) = c/||x||^2$ where c is a positive constant. Note that γ is only twice weakly differentiable (but not twice differentiable in the usual sense) only when $p > 4$ (thus its Laplacian exists as a locally integrable function). Then it may be shown that $\Delta\gamma(x) = -2c(p-4)/||x||^4$. Hence $k\Delta(x) + \gamma^2(x) = -2kc(p-4)/||x||^4 + c^2/||x||^4 < 0$ if $-2kc(p-4) + c^2 < 0$, that is, $0 < c < 2k(p-4)$. It is easy to see that the optimal value of c for which this inequality is the most negative equals $k(p-4)$, so an interesting estimate in this class of γ's is $\delta = \delta_0 - k(p-4)/||x||^2 (p > 4)$. This is precisely the estimate proposed by [29] in the normal distribution case $N_p(\theta, I)$ where $k = 2$; recall, in that case $\delta_0 = p$.

In this example, we have assumed that the dimension p is greater than four. In general we can have domination as long as the assumptions of the theorem are valid. Actually, Blanchard and Fourdrinier [6] show explicitly that, when $p \leq 4$, the only solution γ in $L^2_{\text{loc}}(\mathbb{R}^p)$ of the inequality $k\Delta\gamma + \gamma^2 \leq 0$ is $\gamma \equiv 0$, almost everywhere with respect to the Lebesgue measure λ. Now, in the normal case $N_p(\theta, I/t)$, an unbiased estimator of the risk difference between an estimator $\delta = \delta_0 - \gamma$ and δ_0 is $2t^{-2}\Delta\gamma + \gamma^2$. Hence, for dimensions four or less, it is impossible to find an estimator $\delta = \delta_0 - \gamma$ whose unbiased estimate of risk is always less that of δ_0. Indeed we cannot have $E_\theta[2t^{-2}\Delta\gamma + \gamma^2] < 0$, for some θ, without having $\lambda[t^{-2}\Delta\gamma(x) + \gamma^2(x) < 0] > 0$, which entails that $\lambda[\gamma(x) \neq 0] > 0$.

In the case of scale mixture of normal distributions, the conjecture of admissibility of $\delta_0 - \gamma$ for lower dimensions, although it is probably true, remains open. Indeed, under

conditions of Theorem 4.1, $k\Delta\gamma + \gamma^2$ is no longer an unbiased estimator of the risk difference and $E_\theta[k\Delta\gamma + \gamma^2]$ is only its upper bound. The use of Blyth's method would need to specify the distribution of X (that is, the mixture distribution G). It is worth noting that dimension-cutoff also arises through the finiteness of $E_\theta[\gamma^2]$ when using the classical shrinkage function $c/||x||^2$.

In order to prove Theorem 4.1 we need some additional technical results. The first lemma gives some important properties of superharmonic functions and is found in Du Plessis [39] and the second lemma links the integral of the gradient on a ball with the integral of the Laplacian.

Lemma 4.1 *If γ is a real valued superharmonic function then*

(i) $\int_{S_{R,\theta}} \gamma(x) U_{R,\theta}(dx) \leq \int_{B_{R,\theta}} \gamma(x) V_{R,\theta}(dx)$.

(ii) Both of the integrals in (i) are decreasing in R.

PROOF See Sections 1.3 and 2.5 in [39]. □

Lemma 4.2 *Suppose γ is a twice weakly differentiable function. Then*

$$\int_{B_{R,\theta}} (x-\theta)^t \nabla\gamma(x) V_{R,\theta}(dx) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} \frac{1}{R^p} \int_0^R r \int_{B_{r,\theta}} \triangle\gamma(x) dx dr.$$

PROOF Since the density of the distribution of the radius under $V_{R,\theta}$ is $(p/R^p) r^{p-1} \mathbb{1}_{[0,R]}(r)$, we have

$$\int_{B_{R,\theta}} (x-\theta)^t \nabla\gamma(x) V_{R,\theta}(dx) = \int_0^R \int_{S_{r,\theta}} (x-\theta)^t \nabla\gamma(x) U_{r,\theta}(dx) \frac{p}{R^p} r^{p-1} dr.$$

The result follows from applying (4.1) to the inner most integral of the right hand side of this equality and by recalling the fact that $\sigma_{r,\theta}(S_{r,\theta}) = (2\pi^{p/2}/\Gamma(p/2)) r^{p-1}$. □

PROOF OF THEOREM 4.1 Denoting by ρ the distribution of the radius $||X-\theta||$, the risk difference between δ_0 and $\delta_0 - \gamma$ equals $\alpha(\theta) + \beta(\theta)$ where

$$\alpha(\theta) = \int_{\mathbb{R}_+} \alpha_R(\theta)\, \rho_(dR) \quad \text{and} \quad \beta(\theta) = \int_{\mathbb{R}_+} \beta_R(\theta)\, \rho_(dR) \tag{4.7}$$

with

$$\alpha_R(\theta) = 2\, R^2 \int_{B_{R,\theta}} \gamma(x)\, V_{R,\theta}(dx) - 2\, \lambda_0 \int_{S_{R,\theta}} \gamma(x)\, U_{R,\theta}(dx) \tag{4.8}$$

and

$$\beta_R(\theta) = 2\,\frac{R^2}{p}\int_{B_{R,\theta}} (x-\theta)^t\,\nabla\gamma(x)\,V_{R,\theta}(dx) + \int_{S_{R,\theta}} \gamma^2(x)\,U_{R,\theta}(dx)\,. \tag{4.9}$$

Indeed, the risk difference conditional on the radius R equals

$$\int_{S_{R,\theta}} \left[2\,||x-\theta||^2\,\gamma(x) - 2\,\lambda_0\,\gamma(x) + \gamma^2(x)\right] U_{R,\theta}(dx)$$

and the result follows from (4.2) applied to the first term between brackets.

Let us first deal with $\alpha(\theta)$ considering the first term in (4.8). We have from the definition of $V_{R,\theta}$ and an application of Fubini's theorem

$$\begin{aligned}\int_{\mathbb{R}_+} R^2 \int_{B_{R,\theta}} \gamma(x)\,V_{R,\theta}(dx)\,\rho(dR) &= p\,\frac{\Gamma(p/2)}{2\pi^{p/2}}\int_{\mathbb{R}_+} R^{2-p}\int_{B_{R,\theta}}\gamma(x)\,dx\,\rho(dR)\\ &= p\,\frac{\Gamma(p/2)}{2\pi^{p/2}}\int_{\mathbb{R}^p}\gamma(x)\int_{||x-\theta||}^{+\infty} R^{2-p}\,\rho(dR)\,dx \end{aligned}\tag{4.10}$$

Now, for fixed $t \geq 0$, in the normal case $N_p(\theta, I/t)$ the distribution ρ_t of the radius has the density f_t of the form $f_t(R) = \frac{t^{p/2}}{2^{p/2-1}\Gamma(p/2)} R^{p-1}\exp\{-\frac{tR^2}{2}\}$ and $\delta_0 = \frac{p}{t}$. Thus the expression (4.10) becomes

$$\begin{aligned}\int_{\mathbb{R}_+} R^2 \int_{B_{R,\theta}} \gamma(x)\,V_{R,\theta}(dx)\,\rho(dR) &= \frac{p\,t^{p/2}}{(2\pi)^{p/2}}\int_{\mathbb{R}^p}\gamma(x)\int_{||x-\theta||}^{+\infty} R\exp\left\{-\frac{tR^2}{2}\right\}dR dx\\ &= \frac{p\,t^{p/2-1}}{(2\pi)^{p/2}}\int_{\mathbb{R}^p}\gamma(x)\exp\left\{-\frac{t}{2}||x-\theta||^2\right\}dx\\ &= \frac{p}{t}\int_{\mathbb{R}_+}\int_{S_{R,\theta}}\gamma(x)\,U_{R,\theta}(dx)\,\rho_t(dR),\end{aligned}$$

the last equality holding since $X \stackrel{D}{=} RU^{(p)}$. Turning back to (4.7) and (4.8) and using the mixture representation with mixing distribution G, the expression of $\alpha(\theta)$ is written as

$$\alpha(\theta) = 2p\int_{\mathbb{R}_+}\left(\frac{1}{t}-\frac{\delta_0}{p}\right)\int_{\mathbb{R}^p}\gamma(x)\left(\frac{t}{2\pi}\right)^{p/2}\exp\left(-\frac{t}{2}||x-\theta||^2\right)dx\,G(dt). \tag{4.11}$$

It can be easily seen that the inner most integral in (4.11) is proportional to

$$\int_0^\infty\int_{S_{(u/t)^{1/2},\theta}}\gamma(x)\,dU_{S_{(u/t)^{1/2},\theta}}\,u^{p/2-1}\,\exp\left(-\frac{u}{2}\right)du$$

and hence is non decreasing in t by superharmonicity of γ induced by Inequality (4.6) and by Lemma 4.1 (ii). Thus, since $\delta_0 = p/t$ for fixed t, the expression for $\alpha(\theta)$ in (4.11) is a non positive covariance with respect to G.

We can now treat the integral of the expression $\beta(\theta)$ in the same manner. The function $x \to (x-\theta)^t \nabla\gamma(x)$ and the function $x \to \nabla\gamma(x)$ taking successively the role of the function γ, we obtain

$$\begin{aligned}
\int_{R_+} \frac{R^2}{p} \int_{B_{R,\theta}} (x-\theta)^t \nabla\gamma(x) V_{R,\theta}(dx)\rho_t(dR) &= \frac{1}{t}\int_{\mathbb{R}_+}\int_{S_{R,\theta}} (x-\theta)^t \nabla\gamma(x) U_{R,\theta}(dx)\rho_t(dR) \\
&= \frac{1}{t}\int_{R_+} \frac{R^2}{p}\int_{B_{R,\theta}} \nabla\gamma(x)dx\rho_t(dR) \\
&= \frac{t^{p/2-2}}{(2\pi)^{p/2}}\int_{\mathbb{R}^p} \nabla\gamma(x)\exp\left\{-\frac{t}{2}||x-\theta||^2\right\}dx
\end{aligned}$$

applying (4.1) for the second equality and remembering that $\triangle\gamma = \mathrm{div}(\nabla\gamma)$. Therefore by Fubini Theorem $\beta(\theta)$ can be reexpressed as

$$\begin{aligned}
\beta(\theta) &= \int_{\mathbb{R}^p} \left(2\triangle\gamma(x) \frac{\int_{\mathbb{R}_+} t^{p/2-2}\exp(-t||x-\theta||^2/2)G(dt)}{\int_{\mathbb{R}_+} t^{p/2}\exp(-t||x-\theta||^2/2)G(dt)} + \gamma^2(x)\right) \\
&\times \int_{\mathbb{R}_+} \left(\frac{t}{2\pi}\right)^{p/2} \exp\left(-\frac{t}{2}||x-\theta||^2\right) G(dt)\,dx. \qquad (4.12)
\end{aligned}$$

Now, through a monotone likelihood ratio argument, the ratio of integrals in (4.12) can be seen bounded from below by the constant k in (4.6). Hence Inequality (4.6) gives

$$\beta(\theta) \le \int_{\mathbb{R}^p} (k\triangle\gamma(x) + \gamma^2(x)) \int_{\mathbb{R}_+} \left(\frac{t}{2\pi}\right)^{p/2} \exp\left(-\frac{t}{2}||x-\theta||^2\right) G(dt)\,dx \;<\; 0\,.$$

Finally, remembering that $\alpha(\theta)$ is non positive, it follows that the risk difference $\alpha(\theta)+\beta(\theta)$ between δ_0 and $\delta_0 - \gamma$ is negative, which proves the theorem. □

The improved loss estimator result in Theorem 4.1 for scale mixture of normal distributions family was extended to the more general family of spherically symmetric distributions in [23]. In this setting the conditions for improvement rest on the generating function g of the spherical density p_θ. A sufficient condition for domination of δ_0 has the usual form $k\Delta\gamma + \gamma^2 \le 0$.

Theorem 4.2 *Assume the spherical distribution of X with generating function g has finite fourth moment. Assume the function γ is nonnegative and twice weakly differentiable on $\mathbb{R}^p$ and satisfies $E_\theta[\gamma^2] < \infty$. If, for every $s \ge 0$,*

$$\frac{\int_s^\infty g(z)\,dz}{2\,g(s)} \le \frac{\delta_0}{p} \qquad (4.13)$$

and if there exists a constant k such that, for any $s \ge 0$,

$$0 < k < \frac{\int_s^\infty z g(z) dz - s \int_s^\infty g(z) dz}{2g(s)}. \tag{4.14}$$

Then a sufficient condition for $\delta_0 - \gamma$ to dominate δ_0 is that γ satisfies the differential inequality

$$k \, \Delta\gamma + \gamma^2 < 0 \, .$$

We have shown that one can dominate the unbiased constant estimator of loss by a shrinkage-type estimator. As in the normal case one may wish to add a frequentist-validity constraint to the loss estimation problem. It is easy to show that the only frequentist valid estimator of the form δ_0 would be the only frequentist valid loss estimator. The proof of this result follows from a randomization of the origin technique as in Hsieh and Hwang [27].

4.2 Estimating the quadratic loss of the mean of a spherical distribution with a residual vector

In this subsection, we extend the ideas of the previous sections to a spherically symmetric distribution with a residual vector. We first develop an unbiased estimator of the loss and then construct a dominating shrinkage-type estimator. An important feature of our results is that the proposed loss estimates dominate the unbiased estimates for the entire class of spherically symmetric distributions. That is, the domination results are robust with respect to spherical symmetry.

Let $(X, U) \sim SS(\theta, 0)$ where $\dim X = \dim \theta = p$ and $\dim U = \dim 0 = k$ $(p + k = n)$. For convenience representation, here (X, U) and $(\theta, 0)$ represent $n \times 1$ vectors (see Appendix A.2 for more details on this model). Unlike Subsection 4.1, the dimension of the observable (X, U) is greater than the dimension of the estimand θ. This model arises as the canonical form of the following seemingly more general model, the general linear model. Let V be an $n \times p$ matrix (of full rank p) which is often referred to as the design matrix. Suppose an $n \times 1$ vector Y is observed such that $Y = V\beta + \varepsilon$ where β is a $p \times 1$ vector of (unknown) regression coefficients and ε is an $n \times 1$ vector with a spherically symmetric distribution about 0. A common alternative representation of this model is $Y = \eta + \varepsilon$ where ε is as above and η is in the column space of V.

To understand this representation in terms of the general linear model, let $G = (G_1^t, G_2^t)^t$ be an $n \times n$ orthogonal matrix partitioned such that the first p rows of G (*i.e.* the rows of G_1 considered as column vectors) span the column space of V. Now let

$$\begin{pmatrix} X \\ U \end{pmatrix} = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} Y = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} V\beta + G\varepsilon = \begin{pmatrix} \theta \\ 0 \end{pmatrix} + G\varepsilon$$

with $\theta = G_1 V\beta$ and $G_2 V\beta = 0$ since the rows of G_2 are orthogonal to the columns of V. It follows from the definition that (X, U) has a spherically symmetric distribution about $(\theta, 0)$. In this sense, the model given above is the canonical form of the general linear model.

The usual estimator of θ is the orthogonal projector X. A class of competing point estimators which are also considered are of the form $\varphi = X - ||U||^2 g(X)$, g is a measurable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. This class of estimators is closely related to a Stein-like estimators (when estimating the mean of a normal distribution, the square of the residual term $||u||$ is used as an estimate of the unknown variance). Their domination properties are robust with respect to spherical symmetry (*cf.* [9] and [10]). We will first consider estimation of the loss of the usual least squares estimator X then estimation of the loss of the more general shrinkage estimator φ. In order to assure the finiteness of their risk of the usual estimator X and the risk of the shrinkage estimator φ, we need two hypotheses (H1) and (H2) given in [9].

In the spherical case in Section 3, the risk of X was constant with respect to θ. Thus this risk provides an unbiased estimator of the loss, that is, $\frac{p}{n} E[R^2]$, which is subject to the knowledge of $E[R^2]$. Its properties, as the properties of any improved estimator, may depend on the specific underlying distribution. An important feature of the results in this subsection is that we propose an unbiased estimator δ_0 of the loss of X which is available for every spherically symmetric distribution (with finite fourth moment), that is, $\delta_0(X, U) = p\,||U||^2/k$. Thus we do not need to know the specific distribution, and we get robustness with an estimator which is no longer constant. Notice δ_0 makes sense because $p < n$ (i.e. $k \geq 1$).

In this subsection, we consider estimation of θ by X so that, as in Fourdrinier and Wells [22], we deal with estimating the loss $||X - \theta||^2$. An unbiased estimator of that loss is given by $\delta_0(X, U) = p\,||U||^2/k$, that we write $\delta_0(U)$ since it depends only on U. The unbiasedness of δ_0 follows from Corollary A.1 by taking $q = 0$ and $\gamma \equiv 1$. The goal of this subsection is to prove the domination of the unbiased estimator δ_0 by a competing estimator δ of the form

$$\delta(X, U) = \delta_0(U) - ||U||^4\,\gamma(X), \tag{4.15}$$

where γ is a non negative function. It is important to notice that the "residual term" $||U||$ appears explicitly in the shrinkage function. It has been noted in [9] that the use of this term allows fewer assumptions about the distributions than when it does not appear. Specifically, this including gives a robustness property to the results, since they are valid for the entire class of spherically symmetric distributions.

We require the real-valued function γ to be twice weakly differentiable, in order to include basic examples, which are not twice differentiable. The following domination result is given in [22]. We will see below that it appears as a consequence of a more general result when shrinkage estimator of θ are involved.

Theorem 4.3 *Assume that $p \geq 5$, the distribution of (X, U) has a finite fourth moment and the function γ is twice weakly differentiable on $\mathbb{R}^p$ and there exists a constant β such that $\gamma(t) \leq \beta/||t||^2$. A sufficient condition under which the estimator δ in (4.15) dominates the unbiased estimator δ_0 is that γ satisfies the differential inequality*

$$\gamma^2 + \frac{2}{(k+4)(k+6)} \triangle \gamma \leq 0\,. \tag{4.16}$$

The standard example where $\gamma(t) = d/||t||^2$ for all $t \neq 0$ with $d > 0$ satisfies the conditions of the theorem. More precisely it is easy to deduce that $\triangle\gamma(t) = -2\,d\,(p-4)/|||t||^4$ and thus the sufficient condition of the theorem is written as $0 < d \leq 4\,(p-4)/(k+4)(k+6)$, which only occurs when $p \geq 5$. Straightforward calculus shows that the optimal value of d is given by $2\,(p-4)/(k+4)(k+6)$. The optimal constant in [9] is equal to $2\,(p-4)$. The extra terms in the denominator compensate for the $||U||^4$ term in our estimator.

We will now consider the estimation of the loss of a class of shrinkage estimators considered in [9] (with a slight modification of their form in order to have notations coherent with these of the previous sections), that is, location estimators of the form

$$\varphi_g = X + ||U||^2\, g(X), \tag{4.17}$$

where g is a weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. In [9] it is shown that, if $||g||^2 \leq -2\,\mathrm{div} g/(k+2)$, then φ_g dominates X, under quadratic loss for all spherically symmetric distributions with a finite second moment. A general example of a member of this class of estimators is with $g(X) = -r(||X||^2)\frac{A(X)}{b(X)}$, where r is a positive differentiable and nondecreasing function, A is a positive definite symmetric matrix and b is a positive definite quadratic form of $\mathbb{R}^p$. When r is equal to some constant a, A is the identity on $\mathbb{R}^p$ and the quadratic form b is the usual norm, g reduces to $a/||X||^2$. It can be shown that the optimal choice of a equals $(p-2)/(k+2)$. A member of the class is $\varphi_r = X - (p-2)\frac{||U||^2}{k+2}\frac{X}{||X||^2}$, the James-Stein estimator used when the variance is unknown as in Section 3.

In Proposition 2.3.1 of Section 2.3 of [9], it is shown that an unbiased estimator of the loss of the shrinkage estimator φ_g is given by

$$\delta_0^g(X,U) = \frac{p}{k}||U||^2 + \left(||g(X)||^2 + \frac{2}{k+2}\,\mathrm{div} g(X)\right)||U||^4. \tag{4.18}$$

As in Theorem 4.3 above, the unbiased estimator of the loss can be improved by a shrinkage estimator of the loss. Thus the competing estimator we consider is

$$\delta_\gamma^g(X,U) = \delta_0^g(X,U) - ||U||^4\,\gamma(X), \tag{4.19}$$

where γ is a non negative function. Note that (4.19) is a true shrinkage estimator, while Johnstone's [29] optimal loss estimate for the normal case is an expanding estimator.

This is not contradictory since we are using a different estimator than Johnstone and he is only dealing with the normal case. If $g \equiv 0$ the following result reduces to Therem 4.3.

Theorem 4.4 *Assume that $p \geq 5$, the distribution of (X, U) has a finite fourth moment and the function γ is twice weakly differentiable on $\mathbb{R}^p$ and there exists a constant β such that $\gamma(t) \leq \beta/||t||^2$. A sufficient condition under which the estimator δ_γ^g given in (4.19) dominates the unbiased estimator δ_0^s is that γ satisfies the differential inequality*

$$\gamma^2 - \frac{4}{k+2}\,\gamma\,\mathrm{div} g + \frac{4}{k+6}\,\mathrm{div}(\gamma\, g) + \frac{2}{(k+4)(k+6)}\,\triangle\,\gamma \leq 0\,. \tag{4.20}$$

PROOF Since the distribution of (X, U) is spherically symmetric around θ, it suffices to obtain the result working conditionally on the radius. For $R > 0$ fixed, we can compute using the uniform distribution $U_{R,\theta}$ on the sphere $S_{R,\theta}$. Thus the conditional risk difference between δ_γ^g and δ_0^g, according to (4.19), equals

$$E_{R,\theta}\left[(\delta_\gamma^g(X,U) - ||\varphi(X,U) - \theta||^2)^2\right] - E_{R,\theta}\left[(\delta_0^g(X,U) - ||\varphi(X,U) - \theta||^2)^2\right] =$$
$$E_{R,\theta}\left[||U||^8\,\gamma^2(X)\right] - E_{R,\theta}\left[2\,||U||^4\gamma(X)(\delta_0^g(X,U) - ||\varphi(X,U) - \theta||^2)\right]$$

that is, expanding and separating the integrand terms depending on θ,

$$E_{R,\theta}\left[||U||^8\,\gamma^2(X) - 2\,\frac{p}{k}\,||U||^6\,\gamma(X) - \frac{4}{k+2}\,||U||^8\,\mathrm{div} g(X)\right] +$$
$$E_{R,\theta}\left[4\,||U||^6\,(X-\theta)^t\gamma(X)\,g(X)\right] + E_{R,\theta}\left[2\,||U||^4\,||X-\theta||^2\gamma(X)\right],$$

according to (4.18) (note that the two terms involving $||g(X)||^2$ cancel). Now we have

$$E_{R,\theta}\left[4\,||U||^6\,(X-\theta)^t\gamma(X)\,g(X)\right] = \frac{4}{k+6}E_{R,\theta}\left[||U||^8\,\mathrm{div}\,(\gamma(X)\,g(X))\right]$$

according to Lemma A.2 and

$$E_{R,\theta}\big[2\,||U||^4\,||X-\theta||^2\gamma(X)\big] = E_{R,\theta}\left[\frac{2\,p}{k+4}||U||^6\,\gamma(X) + \frac{2}{(k+4)(k+6)}\,||U||^8\,\triangle\gamma(X)\right]$$

according to Corollary A.1. Therefore the above conditional risk difference is equal to

$$E_{R,\theta}\left[||U||^8\left(\gamma^2(X) - \frac{4}{k+2}\,\mathrm{div} g(X) + \frac{4}{k+6}\,\mathrm{div}\,(\gamma(X)\,g(X)) + \frac{2}{(k+4)(k+6)}\,\triangle\gamma(X)\right)\right]$$
$$+E_{R,\theta}\left[2\,p\left(\frac{1}{k-4} - \frac{1}{k}\right)||U||^6\,\gamma(X)\right]$$

which is bounded above by the first expectation since the function γ is non negative. Hence, the sufficient condition for domination is (4.20) in order that the inequality $R(\delta^g, \theta, \varphi) \leq R(\delta_0^g, \theta, \varphi)$ holds. □

5 Discussion

There are several areas of the theory of loss estimation that we have not discussed. Our primary focus has been on location parameters for the multivariate normal and spherical distributions. Loss estimation for exponential families is addressed in Lele [35] [36] and Rukhin [40]. In [35] and [36] Lele develops improved loss estimators for point estimators in the general setup of Hudson's [28] subclass of continuous exponential family. Hudson's family essentially includes distributions for which the Stein-like identities hold; explicit calculations and loss estimators are given for the gamma distribution, as well as for improved scaled quadratic loss estimators in the Poisson setting for the Clevenson-Zidak [11] estimator. Rukhin [40] studies the posterior loss estimator for a Bayes estimate (under quadratic loss) for the canonical parameter of a linear exponential family.

As point out in the introduction, in the known variance normal setting Johnstone [29] used a version of Blyth's lemma to show that the constant loss estimate p is admissible if $p \leq 4$. Lele [36] give some additional sufficient conditions for admissibility in the general exponential family and works out the precise details for the Poisson model. Rukhin [40] considers loss functions for the simultaneous estimate of θ and $L(\theta, \varphi(X))$ and deduced some interesting admissibility results.

A number of researchers have investigated improved estimators of a covariance matrix, Σ, under the Stein loss, $L_S(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1}) - \log|\hat{\Sigma}\Sigma^{-1}| - p$, using an unbiased estimation of risk technique. In the normal case, [13], [24], [43], [45], and [46] propose improved estimators that dominate the sample covariance under $L_S(\hat{\Sigma}, \Sigma)$. In [33], it is shown that the domination of these improved estimators over the sample covariance estimator are robust with respect to the family of elliptical distributions. To date, there has not been any work on improving the unbiased estimate of $L_S(\hat{\Sigma}, \Sigma)$.

In addition to the theoretical ideas there are very practical applications of loss estimation. The primary application of loss estimation ideas is to model selection. It is shown in Fourdrinier and Wells [21] that improved loss estimators gives more accurate model selection procedures. In linear models the notion of degrees of freedom plays the important role as a model complexity measure in various model selection criteria, such as Akaike information criterion (AIC) [1] , Mallow's C_p [38], and Bayesian information criterion (BIC) [42], and generalized cross-validation (GCV) [12]. In regression the degrees of freedom are the trace of the so-called "hat" matrix. Efron ([15]) pointed out that the theory of Stein's unbiased risk estimation is central to the ideas underlying the calculation of the degrees of freedom of certain regression estimators.

Specifically, let Y be a random vector having a n-variate normal distribution $\mathcal{N}(\theta,\sigma^2 I_n)$ with unknown p-dimensional mean θ and identity covariance matrix $\sigma^2 I_n$. Let $\hat{\theta} = \varphi(Y)$ be an estimate of θ. In regression one focuses is how accurate φ can be in predicting using a new response vector y^{new}. Under the quadratic loss, the prediction risk is $E\{||Y^{new} -$

$\theta||^2\}/n$. Efron [15] notes that

$$E\{||\varphi - \theta||^2\} = E\{||Y - \varphi(Y)||^2 - n\sigma^2\} + 2\sum_{i=1}^{n} \text{Cov}(\varphi_i, Y_i). \tag{5.1}$$

This expression suggests a natural definition of the degrees of freedom for an estimator φ as $df(\varphi) = \sum_{i=1}^{n} \text{Cov}(\varphi_i, Y_i)/\sigma^2 = E_\theta[(Y-\theta)^t\varphi(Y)]/\sigma^2$. Thus one can define a C_p-type quantity

$$C_p(\varphi) = \frac{||Y - \varphi||^2}{n} + \frac{2df(\varphi)}{n}\sigma^2 \tag{5.2}$$

which has the same expectations as the true prediction error but may not be an estimate if $df(\varphi)$ and σ^2 are unknown. However if φ is weak differentiable and $\hat{\sigma}^2$ is an unbiased estimate of σ^2, the integration by parts formula in Lemma 3.1 implies that $df(\varphi)\ \sigma^2 = E_\theta[\text{div}\varphi(Y)\ \hat{\sigma}^2]$, hence $\text{div}\varphi\ \hat{\sigma}^2$ is unbiased estimate for the complexity parameter term, $df(\varphi)\ \sigma^2$, in (5.2). Therefore an unbiased estimate for the prediction error is

$$C_p^*(\varphi) = \frac{||Y - \varphi||^2}{n} + \frac{2\text{div}\varphi}{n}\hat{\sigma}^2. \tag{5.3}$$

Note that if φ is a linear estimator ($\varphi = \mathbf{S}y$ for some matrix $\mathbf{S}$ independent of Y) then it is easy to show that this definition coincides with the definition of generalized degrees of freedom given by Hastie and Tibshirani [25] since $\text{div}\varphi = \text{tr}(\mathbf{S})$. Note that if φ also depends on $\hat{\sigma}^2$ then (5.1) needs to be augmented by additional derivative terms with respect to $\hat{\sigma}^2$ as in the proof of Theorem 3.1.

Other approaches for estimating the complexity term penalty involve the use of resampling methods ([15] [49]) to directly estimate the prediction error. A K-fold cross-validation randomly divides the original sample into K part, and rotates through each part as a test sample and uses the remainder as a training sample. Cross-validation provides an approximately unbiased estimate of the prediction error, although the its variance can be large. Other commonly used resampling techniques are the nonparametric and parametric bootstrap methods.

A number of new regularized regression methods have been recently been developed, starting with Ridge regression [26], followed by the Lasso [47], the Elastic Net [50], and LARS [16]. Each of these estimates are weakly differentiable and have the form of a general shrinkage estimate, thus the prediction error estimate in (5.3) may be applied to construct a model selection procedure. Zou, Hastie and Tibshirani [51] use this idea to develop a model selection method for the Lasso. In some situations verifying the weak differentiability of φ may be complicated.

Loss estimates have been used to derive nonparametric penalized empirical loss estimates in the context of function estimation, which adapt to the unknown smoothness of

the function of interest. See Barron *et al.* [2] and Donoho and Johnstone [14] for more details.

In the previous sections, the usual quadratic loss $L(\theta, \varphi(x)) = ||\varphi(x) - \theta||^2$ was considered to evaluate various estimators $\varphi(X)$ of θ. The squared norm $||x - \theta||^2$ was crucial in the derivation of the properties of the loss estimators in conjonction with its role in the normal density or, more generally, in a spherical density. Other losses are thinkable but, to deal with tractable calculations, it matters to keep the Euclidean norm as a component of the loss in use. Hence a natural extension is to consider losses which are function of $||x-\theta||^2$, that is, of the form $c(||x-\theta||^2)$ for a non-negative function c defined on $\mathbf{R}_+$. The problem of estimating a function c of $||x - \theta||^2$ was tackled by Fourdrinier and Lepelletier [18] whose refer to for more details. In particular, they focus on the fact that estimating $c(||x - \theta||^2)$ can be view as an evaluation of a quantity which is not necessarily a loss. Indeed it includes the problem of estimating the confidence statement of the usual confidence set $\{\theta \in \mathbf{R}^p \,/\, \|x - \theta\|^2 \leq c_\alpha\}$ with confidence coefficient $1 - \alpha$: c is the indicator function $\mathbb{1}_{[0,c_\alpha]}$.

Appendix

A.1 Risk finiteness conditions

Lemma A.1 *1. Let $X \sim \mathcal{N}(\theta, I_p)$, where θ is unknown, and denote by E_θ the expectation with respect to the distribution of X. Consider an estimator of θ of the form $\varphi(X) = X + g(X)$ where g is a function from $\mathbf{R}^p$ into $\mathbf{R}^p$.*

a. If g is such that $E_\theta[||g(X)||^2] < \infty$, then the quadratic risk of $\varphi(X)$, that is, $R(\theta, \varphi) = E_\theta[||\varphi(X) - \theta||^2]$, is finite.

b. If, in addition, the function g is weakly differentiable so that $\delta_0(X) = p + 2\,\mathrm{div} g(X) + ||g(X)||^2$ is an unbiased estimator of the loss $||\varphi(X) - \theta||^2$, then the risk of $\delta_0(X)$ defined by $\mathcal{R}(\theta, \varphi, \delta_0) = E_\theta[(\delta_0(X) - ||\varphi(X) - \theta||^2))^2]$ is finite as soon as $E_\theta[||g(X)||^4] < \infty$ and $E_\theta[(\mathrm{div} g(X))^2] < \infty$.

2. Let $X \sim \mathcal{N}(\theta, \sigma^2 I_p)$, where θ and σ^2 are unknown, let S be a nonnegative random variable independent of X and such that $S \sim \sigma^2 \chi^2_n$ and denote by E_{θ,σ^2} the expectation with respect to the joint distribution of (X, S). Consider an estimator of θ of the form $\varphi(X, S) = X + S\, g(X, S)$ where g is a function from $\mathbf{R}^p \times \mathbf{R}_+$ into $\mathbf{R}^p$.

a. If g is such that $E_{\theta,\sigma^2}[S^2\, ||g(X, S)||^2] < \infty$, then the quadratic risk of $\varphi(X)$, that is, $R(\theta, \sigma^2, \varphi) = E_{\theta,\sigma^2}[||\varphi(X, S) - \theta||^2/\sigma^2]$, is finite.

b. If, in addition, the function g is weakly differentiable so that

$$\delta_0(X,S) = p + S\left\{(n+2)\,||g(X,S)||^2 + 2\,\mathrm{div}_X g(X,S) + 2\,S\frac{\partial}{\partial S}||g(X,S)||^2\right\}.$$

is an unbiased estimator of the loss $||\varphi(X,S)-\theta||^2/\sigma^2$*, then the risk of* $\delta_0(X,S)$ *defined by* $\mathcal{R}(\theta,\sigma^2,\varphi,\delta_0) = E_{\theta,\sigma^2}[(\delta_0(X,S) - ||\varphi(X,S)-\theta||^2/\sigma^2))^2]$ *is finite as soon as* $E_{\theta,\sigma^2}[S^2\,||g(X,S)||^4] < \infty$*,* $E_{\theta,\sigma^2}[(S\,\mathrm{div}g(X,S))^2] < \infty$ *and* $E_{\theta,\sigma^2}\left[\left(S^2\,\frac{\partial}{\partial S}||g(X,S)||\right)^2\right]$.

PROOF 1.a. The loss of $\varphi(X)$ can be expanded as

$$||\varphi(X)-\theta||^2 = ||X-\theta||^2 + ||g(X)||^2 + 2(X-\theta)^t g(X)\,. \tag{A.4}$$

Now we have $E_\theta[||X-\theta||^2] = p < \infty$. Hence, by Schwarz's inequality, it follows from (A.4) that $|E_\theta[(X-\theta)^t g(X)]| \le (E_\theta[||X-\theta||^2])^{1/2}\,(E_\theta[||g(X)||^2])^{1/2}$. Therefore, as soon as $E_\theta[||g(X)||^2] < \infty$, we will have $|E_\theta[||\varphi(X)-\theta||^2] < \infty$. This is the desired result.

b. Note that, under the usual domination condition, that is, $2\,\mathrm{div}g(x) + ||g(x)||^2 \le 0$ for any $x \in R^p$, of $\delta_0(X)$ over X, the condition $E_\theta[(\mathrm{div}g(X))^2] < \infty$ implies that $E_\theta[||g(X)||^4] < \infty$). We will have $\mathcal{R}(\theta,\varphi,\delta_0) = E_\theta[(\delta_0(X) - ||\varphi(X)-\theta||^2))^2] < \infty$ as soon as $E_\theta[\delta_0^2(X)] < \infty$ and $E_\theta[||\varphi(X)-\theta||^4] < \infty$. Now $E_\theta[\delta_0^2(X)] = E_\theta[(p + 2\,\mathrm{div}g(X) + ||g(X)||)^2] < \infty$ since $E_\theta[(\mathrm{div}g(X))^2] < \infty$ and $E_\theta[||g(X)||^4] < \infty$. Also according to (A.4)

$$E_\theta[||\varphi(X)-\theta||^4] = E_\theta[(||X-\theta||^2 + ||g(X)||^2 + 2(X-\theta)^t g(X))^2] < \infty$$

since $E_\theta[||X-\theta||^4] < \infty$ and $E_\theta[||g(X)||^4] < \infty$ and, consequently, since $|(X-\theta)^t g(X)| \le ||X-\theta||\,||g(X)||$ implies that

$$\begin{aligned} E_\theta[|(X-\theta)^t g(X)|^2] &\le E_\theta[||X-\theta||^2\,||g(X)||^2] \\ &\le \left(E_\theta[||X-\theta||^4]\right)^{1/2}\left(E_\theta[||g(X)||^4]\right)^{1/2} \end{aligned}$$

by the Schwarz's inequality.

2.a. Parallel to the case where the variance σ^2 is known, it should be noticed that the corresponding domination condition of $\delta(X,S)$ over $\delta_0(X,S)$, that is, for any $x \in R^p$ and any $s \in \mathbf{R}_+$, $(n+2)\,||g(x,s)||^2 + 2\,\mathrm{div}_x g(x,s) + 2\,s\frac{\partial}{\partial s}||g(x,s)||^2 \le 0$, entails that the two conditions $E_{\theta,\sigma^2}[(S\,\mathrm{div}g(X,S))^2] < \infty$ and $E_{\theta,\sigma^2}\left[\left(S^2\,\frac{\partial}{\partial S}||g(X,S)||\right)^2\right]$ imply the condition $E_{\theta,\sigma^2}[S^2\,||g(X,S)||^4] < \infty$. Also the derivation of the finitness of $R(\theta,\sigma^2,\varphi)$ follows a similar way than in 1.a.

b. We will have $\mathcal{R}(\theta,\sigma^2,\varphi,\delta_0) = E_{\theta,\sigma^2}[(\delta_0(X,S) - ||\varphi(X)-\theta||^2/\sigma^2))^2] < \infty$ as soon as $E_{\theta,\sigma^2}[(\delta_0(X,S))^2 < \infty$ and $E_{\theta,\sigma^2}[||\varphi(X)-\theta||^4] < \infty$. Now $E_{\theta,\sigma^2}[(\delta_0(X,S))^2 = E_{\theta,\sigma^2}[p +$

$S\left\{(n+2)\,||g(X,S)||^2 + 2\,\mathrm{div}_X g(X,S) + 2\,S\frac{\partial}{\partial S}||g(X,S)||^2\right\}] < \infty$ since we assume that $E_{\theta,\sigma^2}[(S\,\mathrm{div}_X g(X,S))^2] < \infty$ and $E_{\theta,\sigma^2}[S^2\,||g(X,S)||^4] < \infty$. Also $E_{\theta,\sigma^2}[||\varphi(X,S)-\theta||^4] = E_{\theta,\sigma^2}[(||X-\theta||^2 + S^2||g(X,S)||^2 + 2S(X-\theta)^t g(X,S))^2])^2] < \infty$ since $E_\theta[||X-\theta||^4] < \infty$ and $E_{\theta,\sigma^2}[S^2\,||g(X,S)||^4] < \infty$ (note that $|(X-\theta)^t g(X,S)| \leq ||X-\theta||\,||g(X,S)||$ implies that

$$\begin{aligned} E_{\theta,\sigma^2}[|(X-\theta)^t S\,g(X,S)|^2] &\leq E_{\theta,\sigma^2}[||X-\theta||^2\,S^2\,||g(X,S)||^2] \\ &\leq \left(E_{\theta,\sigma^2}[||X-\theta||^4]\right)^{1/2}\left(E_{\theta,\sigma^2}[S^2\,||g(X,S)||^4]\right)^{1/2} \end{aligned}$$

by the Schwarz's inequality). □

A.2 Additional Technical Lemmas

This Appendix gives some technical results used in Subsection 4.2. The first two results deal with expectations conditioned on the radius of a spherically symmetric distribution in $\mathbf{R}^p \times \mathbf{R}^k$ centered at $(\theta, 0)$ where $\theta \in \mathbf{R}^p$. These expectations reduce to integrals with respect to the uniform distribution $U_{R,\theta}$ on the sphere

$$S_{R,\theta} = \left\{y = (x,u) \in \mathbf{R}^p \times \mathbf{R}^k / \left(\|x-\theta\|^2 + \|u\|^2\right)^{1/2} = R\right\}.$$

If $E_{R,\theta}[\psi]$ is the expectation of some function ψ with respect to $U_{R,\theta}$, the expectation with respect to the entire distribution is given by $E_\theta[\psi] = E\left[E_{R,\theta}[\psi]\right]$ where E is the expectation with respect to the distribution of the radius.

When the spherical distribution has a density with respect to the Lebesgue measure, it is necessarily of the form $f\left(\|x-\theta\|^2 + \|u\|^2\right)$ for some function f. Then the radius has density $R \to \sigma_{p+k} f(R^2) R^{p+k-1}$ where $\sigma_{p+k} = \frac{2\pi^{p+k}}{\Gamma\left(\frac{p+k}{2}\right)}$. Therefore the expectation of any function ψ can be written as

$$E_\theta[\psi] = \int_0^\infty \left[\int_{S_{R,\theta}} \psi(y) U_{R,\theta}(dy)\right] f(R) dR\,.$$

Note that for a vector $y = (x,u) \in S_{R,\theta}$, we have $x = \pi(y)$ and $\|u\|^2 = R^2 - \|\pi(y)-\theta\|^2$ where π is the orthogonal projector from $\mathbf{R}^p \times \mathbf{R}^k$ onto $\mathbf{R}^p$. Under $U_{R,\theta}$, the distribution $\pi(U_{R,\theta})$ of this projector has a density with respect to the Lebesgue measure on $\mathbf{R}^p$ given by $x \to C_R^{p,k}\left(R^2 - \|x-\theta\|^2\right)^{\frac{k}{2}-1} \mathbf{1}_{B_{R,\theta}}(x)$ where $C_R^{p,k} = \frac{\Gamma\left(\frac{p+k}{2}\right) R^{2-p-k}}{\Gamma\left(\frac{k}{2}\right)\pi^{p/2}}$ and $\mathbf{1}_{B_{R,\theta}}$ is the indicator function of the ball $B_{R,\theta} = \left\{x \in \mathbf{R}^p / \|x-\theta\| \leq R\right\}$ of radius R centered at θ in $\mathbf{R}^p$.

According to the above, as a spherically symmetric distribution on $\mathbf{R}^p$ around θ, the radius of $\pi(U_{R,\theta})$ has density

$$r \to \sigma_p C_R^{p,k} \left(R^2 - r^2\right)^{\frac{k}{2}-1} \mathbf{1}_{]0,R[}(r) r^{p-1} = \frac{2R^{2-p-k}}{B\left(\frac{p}{2}, \frac{k}{2}\right)} r^{p-1} \left(R^2 - r^2\right)^{\frac{k}{2}-1} \mathbf{1}_{]0,R[}(r).$$

We use repeatedly the fact that any such projection onto a space of dimension greater than 0 and less than $p + k$ is spherically symmetric with a density. Then we also often make use of its radial density.

Lemma A.2 *For every twice weakly differentiable function $g(\mathbf{R}^p \to \mathbf{R}^p)$ and for every function $h(\mathbf{R}+ \to \mathbf{R})$,*

$$E_{R,\theta}\left[h(||U||^2)(X-\theta)^t g(X)\right] = E_{R,\theta}\left[\frac{H(||U||^2)}{(||U||^2)^{\frac{k}{2}-1}} \mathrm{div} g(X)\right]. \tag{A.5}$$

where H is the indefinite integral, vanishing at 0, of the function $t \to \frac{1}{2} h(t) t^{\frac{k}{2}-1}$.

PROOF We have

$$\begin{aligned} E_{R,\theta}\left[h(||U||^2)(X-\theta)^t g(X)\right] &= C_R^{p,k} \int_{B_{R,\theta}} h(R^2 - \|x-\theta\|^2)(x-\theta)^t g(x) \left(R^2 - \|x-\theta\|^2\right)^{\frac{k}{2}-1} dx \\ &= C_R^{p,k} \int_{B_{R,\theta}} (\nabla H(R^2 - \|x-\theta\|^2))^t g(x) dx \end{aligned}$$

since

$$\begin{aligned} \nabla H(R^2 - \|x-\theta\|^2) &= -2H'(R^2 - \|x-\theta\|^2)(x-\theta) \\ &= h(R^2 - \|x-\theta\|^2)\left(R^2 - \|x-\theta\|^2\right)^{\frac{k}{2}-1}(x-\theta). \end{aligned}$$

Then, by divergence formula,

$$\begin{aligned} E_{R,\theta}\left[h(||U||^2)(X-\theta)^t g(X)\right] = C_R^{p,k} &\int_{B_{R,\theta}} \mathrm{div}\left(H(R^2 - \|x-\theta\|^2) g(x)\right) dx \\ - \; C_R^{p,k} &\int_{B_{R,\theta}} H(R^2 - \|x-\theta\|^2) \mathrm{div} g(x) dx \end{aligned}$$

Now, if $\sigma_{R,\theta}$ denotes the area measure on the sphere $S_{R,\theta}$, the divergence theorem insures that the first integral equals

$$C_R^{p,k} \int_{S_{R,\theta}} (H(R^2 - \|x-\theta\|^2) g(x))^t \frac{x-\theta}{\|x-\theta\|} \sigma_{R,\theta}(dx)$$

and is null since, for $x \in S_{R,\theta}$, $R^2 - \|x-\theta\|^2 = 0$ and $H(0) = 0$. Hence, in terms of expectation, we have

$$E_{R,\theta}\left[h(||U||^2)(X-\theta)^t g(X)\right]$$

$$= C_R^{p,k} \int_{B_{R,\theta}} \frac{H(R^2 - \|x-\theta\|^2)}{\left(R^2 - \|x-\theta\|^2\right)^{\frac{k}{2}} - 1} \mathrm{div} g(x) \left(R^2 - \|x-\theta\|^2\right)^{\frac{k}{2}-1} dx$$

$$= E_{R,\theta}\left[\frac{H(||U||^2)}{(||U||^2)^{\frac{k}{2}-1}} \mathrm{div} g(X)\right]$$

which is the desired result. □

Corollary A.1 *For every twice weakly differentiable function $\gamma(\mathbf{R}^p \to \mathbf{R}_+)$ and for every integer q,*

$$\begin{aligned} E_{R,\theta}\left[||U||^q \, ||X-\theta||^2 \, \gamma(X)\right] &= \frac{p}{k+q} E_{R,\theta}\left[||U||^{q+2} \, \gamma(X)\right] \\ &+ \frac{1}{(k+q)(k+q+2)} E_{R,\theta}\left[||U||^{q+4} \triangle \gamma(X)\right] . \end{aligned}$$

PROOF Take $h(t) = t^{q/2}$ and $g(x) = \gamma(x)\,(x-\theta)$ and apply Lemma A.2 twice. □

Acknowledgments We are grateful to Bill and Rob Strawderman as well as Rajendran Narayanan for their helpful suggestions and comments that greatly aided the revision of the manuscript.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, 25:267–281, 1973.

[2] A. R. Barron, L. Birgé, and P. Massart. Risk bound for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[3] J. O. Berger. *Decision Theory and Bayesian Analysis*. Springer, New York, 1985.

[4] J. O. Berger. The frequentist viewpoint and conditioning. In L. Le Cam and R. Olshen, editors, *Proceedings of the Berkeley Conference in Honor of Jersy Neyman and Jack Kiefer*, volume 1, pages 15–44. Wadsworth, Monterey, California, 1985.

[5] J. O. Berger. In the defense of likelihood principle: Axiomatics and coherency. In D. V. Lindley J. M. Bernardo, M. H. DeGroot and A. F. M. Smith, editors, *Bayesian Statistics II*. Norh Holland, Amsterdam, 1985.

[6] D. Blanchard and D. Fourdrinier. Non trivial solutions of non-linear partial differential inequations and order cut-off. *Rendiconti di Matematica*, 19:137–154, 1999.

[7] M. E. Bock. Shrinkage estimator: Pseudo-bayes estimators for normal mean vectors. In S. S. Gupta and J. 0. Berger, editors, *Statistical Decision Theory and Related Topics 4*, volume 1, pages 281–298. Springer-Verlag, New York, 1988.

[8] L. D. Brown. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Mathematical Statistics*, 42:855–903, 1971.

[9] D. Cellier and D. Fourdrinier. Shrinkage estimators under spherical symmetry for the general linear model. *Journal of Multivariate Analysis*, 52:338Ű–351, 1995.

[10] D. Cellier, D. Fourdrinier, and C. Robert. Robust shrinkage estimators of the location parameter for elliptically symmetric distributions. *Journal of Multivariate Analysis*, 29:39–52, 1989.

[11] M.L. Clevenson and J.V. Zidek. Simultaneous estimation of the mean of independent poisson laws. *Journal of the American Statistical Association*, 70:698Ű–705, 1975.

[12] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[13] D. K. Dey and C. Srinivasan. Estimation of covariance matrix under stein's loss. *The Annals of Statistics*, 13:1581–1591, 1985.

[14] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–Ű1244, 1995.

[15] B. Efron. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 81:461–470, 2004.

[16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[17] K. T. Fang, K. S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, New York, 1990.

[18] D. Fourdrinier and P. Lepelletier. Estimating a general function of a quadratic function. *Annals of the Institute of Statistical Mathematics*, 60:85–119, 2008.

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

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

[21] D. Fourdrinier and M. T. Wells. Comparaisons de procédures de sélection d'un modèle de régression: Une approche décisionnelle. *C.R. Acad. Sci. Paris Serie I*, 319:865–870, 1994.

[22] D. Fourdrinier and M. T. Wells. Estimation of a loss function for spherically symmetric distributions in the general linear model. *Annals of Statistics*, 23(2):571–592, 1995.

[23] D. Fourdrinier and M. T. Wells. Loss estimation for spherically symmetric distributions. *Journal of Multivariate Analysis*, 53:311–331, 1995.

[24] L. R. Haff. An identity for the wishart distribution with applications. *Journal of Multivariate Analysis*, 9:531–544, 1979.

[25] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[26] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[27] F. Hsieg and J. T. G. Hwang. Admissibility under the frequentist's validity constraint in estimating the loss of the least-squares estimator. *Journal of Multivariate Analysis*, 44:279–285, 1993.

[28] H. M. Hudson. A natural identity for exponential families with applications in multiparameter estimation. *Annals of Statistics*, 6:473Ű–484, 1978.

[29] I. Johnstone. On inadmissibility of some unbiased estimates of loss. In S. S. Gupta and J. 0. Berger, editors, *Statistical Decision Theory and Related Topics IV*, volume 1, pages 361–379. Springer-Verlag, New York, 1988.

[30] J. Kiefer. Conditional confidence approach in multi-decision problems. In P. R. Krishnaiah, editor, *Multivariate Analysis 4*. Academic Press, New York, 1975.

[31] J. Kiefer. Admissibility of conditional confidence procedures. *Annals of Statistics*, 4:836–865, 1976.

[32] J. Kiefer. Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, 72:789–827, 1977.

[33] T. Kubokawa and M. S. Srivastava. Robust improvement in estimation of a covariance matrix in an elliptically contoured distribution. *The Annals of Statistics*, 27:600–609, 1999.

[34] E. L. Lehmann and H. Sheffé. Completeness, similar regions and unbiased estimates. *Sankhyā*, 17:305–340, 1950.

[35] C. Lele. Inadmissibility of loss estimators. *Statistics and Decision*, 10:309–322, 1992.

[36] C. Lele. Admissibility results in loss estimation. *Annals of Statistics*, 21:378–390, 1993.

[37] K. L. Lu and J. O. Berger. Estimation of normal means: frequentist estimation of loss. *Annals of Statistics*, 17:890–906, 1989.

[38] C. Mallows. Some comments on Cp. *Technometrics*, 15:661–675, 1973.

[39] N. Du Plessis. *An Introduction to Potential Theory.* Hafner, Darien, CT, 1970.

[40] A. L. Rukhin. Estimated loss and admissible loss estimators. In S. S. Gupta and J. 0. Berger, editors, *Statistical Decision Theory and Related Topics 4*, volume 1, pages 409–418. Springer-Verlag, New York, 1988.

[41] E. Sandved. Ancillary statistics and estimation of the loss in estimation problems. *Annals of Mathematical Statistics*, 39:1756–1758, 1968.

[42] G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[43] C. Stein. Estimating the covariance matrix. Unpublished manuscript.

[44] C. Stein. Estimation of the mean of multivariate normal distribution. *Annals of Statistics*, 9:1135–1151, 1981.

[45] C. Stein. Lectures on the theory of estimation of many parameters. In A. Ibragimov and M. S. Nikulin, editors, *Studies in the Statistical Theory of Estimation*, volume 74, pages 4–65. Proceedings of Scientific Seminars of the Steklov Institute, Leningrad, 1988.

[46] A. Takemura. An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population. *Tsukuba Journal of Mathematics*, 8:367–376, 1984.

[47] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[48] A. T. K. Wan and G. Zou. On unbiased and improved loss estimation for the mean of a multivariate normal distribution with unknown variance. *Journal of Statistical Planning and Inference*, 119:17–22, 2004.

[49] J. Ye. On meausuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131, 1998.

[50] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[51] H. Zou, T. Hastie, and R. Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35:2173–2192, 2007.