AIC, Cp and estimators of loss for elliptically symmetric distributions

Aurélie Boisbunon, Stephane Canu, Dominique Fourdrinier, William Strawderman, Martin T. Wells

AIC and C_p as estimators of loss for spherically symmetric distributions*

Aurélie Boisbunon, Stéphane Canu, Dominique Fourdrinier, William Strawderman and Martin T. Wells

Abstract: In this article, we develop a modern perspective on Akaike's Information Criterion and Mallows' C_p for model selection. Despite the differences in their respective motivation, they are equivalent in the special case of Gaussian linear regression. In this case they are also equivalent to a third criterion, an unbiased estimator of the quadratic prediction loss, derived from loss estimation theory. Our first contribution is to provide an explicit link between loss estimation and model selection through a new oracle inequality. We then show that the form of the unbiased estimator of the quadratic prediction loss under a Gaussian assumption still holds under a more general distributional assumption, the family of spherically symmetric distributions. One of the features of our results is that our criterion does not rely on the specificity of the distribution, but only on its spherical symmetry. Also this family of laws offers some dependence property between the observations, a case not often studied.

Keywords and phrases: variable selection, C_p, AIC, loss estimator, unbiased estimator, SURE estimator, Stein identity, spherically symmetric distributions.

1. Introduction

The problem of model selection has generated a lot of interest for many decades now and especially recently with the increased size of datasets. In such a context, it is important to model the data observed in a sparse way. The principle of parsimony helps to avoid classical issues such as overfitting or computational error. At the same time, the model should capture sufficient information in order to comply with some objectives of good prediction, good estimation or good selection and thus it should not be too sparse. This principle has been summarized by many statisticians as a trade-off between goodness of fit to data and complexity of the model (see for instance Hastie, Tibshirani, and Friedman, 2008, Chapter 7). From the practitioner point of view, model selection is often implemented through cross-validation (see Arlot and Celisse, 2010, for a review on this topic) or the minimization of criteria whose theoretical justification relies on hypothesis made within a given framework. In this paper, we review two of the most commonly used criteria, namely Mallows' C_p and Akaike's AIC, together with the associated theory under Gaussian distributional assumptions, and then we propose a generalization towards spherically symmetric distributions.

*This research was partially supported by ANR ClasSel grant 08-EMER-002, by the Simons Foundation grant 209035 and by NSF Grant DMS-12-08488.

We will focus on the linear regression model

$$Y = X\beta + \sigma\varepsilon, \tag{1}$$

where Y is a random vector in $\mathbb{R}^n$, X is a fixed and known full rank design matrix containing p observed variables $\mathbf{x}^j$ in $\mathbb{R}^n$, β is the unknown vector in $\mathbb{R}^p$ of regression coefficients to be estimated, σ is the noise level and ε is a random vector in $\mathbb{R}^n$ representing the model noise, with mean zero and covariance matrix proportional to the identity matrix (we assume the proportion coefficient to be equal to one when ε is Gaussian). One subproblem of model selection is the problem of variable selection: only a subset of the variables in X gives sufficient and non-redundant information on Y and we wish to recover this subset as well as correctly estimate the corresponding regression coefficients.

Early works treated the model selection problem from the hypothesis testing point of view. For instance Forward Selection and Backward Elimination were stopped using Student's critical values. This practice changed with Mallows' automated criterion known as C_p (Mallows, 1973). Mallows' idea was to propose an unbiased estimator of the scaled expected prediction error $\mathbb{E}_\beta[\|X\hat{\beta}_I - X\beta\|^2/\sigma^2]$, where $\hat{\beta}_I$ is an estimator of β based on the selected variables set $I \subset \{1, \dots, p\}$, $\mathbb{E}_\beta$ denotes the expectation with respect to the sampling distribution in model (1) and $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$. This way, assuming Gaussian *i.i.d.* error terms, Mallows came to the following criterion

$$C_p = \frac{\|Y - X\hat{\beta}_I\|^2}{\hat{\sigma}^2} + 2\widehat{df} - n, \tag{2}$$

where $\hat{\sigma}^2$ is an estimator of the variance σ^2 based on the full linear model fitted with the least-squares estimator $\hat{\beta}^{LS}$, that is, $\hat{\sigma}^2 = \|Y - X\hat{\beta}^{LS}\|^2/(n-p)$, and $\widehat{df}$ is an estimator of df, the degrees of freedom, also called the effective dimension of the model (see Hastie and Tibshirani, 1990; Meyer and Woodroofe, 2000). For the least squares estimator, df is the number k of variables in the selected subset I.

Mallows' C_p relies on the assumption that, if for some subset I of explanatory variables the expected prediction error is low, then we can assume those variables to be relevant for predicting Y. In practice, the rule for selecting the "best" candidate is the minimization of C_p. However, Mallows argues that this rule should not be applied in all cases, and that it is better to look at the shape of the C_p-plot instead, especially when some explanatory variables are highly correlated.

In 1974, Akaike followed Mallows' spirit to propose automatic criteria that would not need a subjective calibration of the significance level as in hypothesis testing. His proposal was more general with application to many problems such as variable selection, factor analysis, analysis of variance, or order selection in auto-regressive models (Akaike, 1974). Akaike's motivation however was different from Mallows. He considered the problem of estimating the density $f(\cdot|\beta)$ of an outcome variable Y, where f is parametrized by $\beta \in \mathbb{R}^p$, by $f(\cdot|\hat{\beta})$. His aim was to generalize the principle of maximum likelihood enabling a selection

between several maximum likelihood estimators $\hat{\beta}_I$. Akaike showed that all the information for discriminating $f(\cdot|\hat{\beta}_I)$ from $f(\cdot|\beta)$ could be summed up by the Kullback-Leibler divergence $D_{KL}(\hat{\beta}_I, \beta) = \mathbb{E}[\log f(Y_{\text{new}}|\beta)] - \mathbb{E}[\log f(Y_{\text{new}}|\hat{\beta}_I)]$ where the expectation is taken over new observations. This divergence can in turn be approximated by its second-order variation when $\hat{\beta}_I$ is sufficiently close to β, which actually corresponds to the distance $\|\hat{\beta}_I - \beta\|^2_{\mathcal{I}}/2$ where $\mathcal{I} = -\mathbb{E}[(\partial^2 \log f/\partial\beta_i\partial\beta_j)_{i,j=1}^p]$ is the Fisher-information matrix and for a vector $\mathbf{z}$, its weighted norm $\|\mathbf{z}\|_{\mathcal{I}}$ is defined by $(\mathbf{z}^t\mathcal{I}\mathbf{z})^{1/2}$. By means of asymptotic analysis and by considering the expectation of D_{KL} the author arrived at the following criterion

$$\text{AIC} = -2\sum_{i=1}^{n} \log f(y_i|\hat{\beta}_I) + 2k, \tag{3}$$

where k is the number of parameters of $\hat{\beta}_I$. In the special case of a Gaussian distribution, AIC and C_p are equivalent up to a constant for model (1) (see Section 2.2). Hence Akaike described his AIC as a generalization of C_p for other distributional assumptions. Unlike Mallows, Akaike explicitly recommends the rule of minimization of AIC to identify the best model from data. Note that Ye (1998) proposed to extend AIC to more complex settings by replacing k by the estimated degrees of freedom $\widehat{df}$.

Both C_p and AIC have been criticized in the literature, especially for the presence of the constant 2 tuning the adequacy-complexity trade-off and favoring complex models in many situations, and many authors have proposed some correction (see Schwarz (1978); Burnham and Anderson (2002); Foster and George (1994); Shibata (1980)). Despite these critics, these criteria are still quite popular among practicioners. Also they can be very useful in deriving better criteria of the form $\delta = \delta_0 - \gamma$, where δ_0 is equal to C_p, AIC or an equivalent and γ is a correction function based on data. This framework, referred to as loss estimation, has been successfully used by Johnstone (1988) and Fourdrinier and Wells (1995), among others, to propose good criteria for selecting the best submodel.

Another possible criticism of C_p and AIC regards their strong distributional assumptions. Indeed, C_p's unbiasedness has been shown under the Gaussian *i.i.d.* case, while AIC requires the specification of the distribution. However, in many practical cases, we might not have any prior knowledge or intuition about the form of the distribution, and we want the result to be robust to a wide family of distributions.

The purpose of the present paper is twofold:

- First, we show in Section 2 that the procedures C_p and AIC are equivalent to unbiased estimators of the quadratic prediction loss when Y is assumed to be Gaussian in model (1). This result is an important initial step for deriving improved criteria as is done in Johnstone (1988) and Fourdrinier and Wells (2012). Both references consider the case of improving the unbiased estimator of loss based on the data, which is consistent with other approaches on data-driven penalties based on oracle inequali-

ties (see for instance Birgé and Massart, 2007; Arlot and Massart, 2009). The derivation of better criteria will not be covered in the present article, but a relationship between oracle inequality and the statistical risk of the estimators of the prediction loss is provided section 2.2.2.

- Second, we derive the unbiased loss estimator for the wide family of spherically symmetric distributions and show that, for any spherical law, this unbiased estimator is the same as that derived under the Gaussian law. The family of spherically symmetric distribution is a large family which generalizes the multivariate standard normal law and includes multivariate versions of the Student, Cauchy, Kotz, and Pearson type II and type VII distributions among others. Also, the spherical assumption frees us from the independence assumption of the error terms in (1), while not rejecting it since the Gaussian law is spherical. Note however that spherical symmetry here means no correlation, the case of correlated observations being handled by elliptical symmetry. Furthermore, some members of the spherical family, like the Student law, have heavier tails than the Gaussian density allowing a better handling of potential outliers. Finally, the results of the present work do not depend on the specific form of the distribution. The last two points provide some distributional robustness.

2. Expression of AIC and C_p in the loss estimation framework

2.1. Basics of loss estimation

2.1.1. Unbiased loss estimation

The idea underlying the estimation of loss is closely related to Stein's Unbiased Risk Estimate (SURE, Stein, 1981). The theory of loss estimation was initially developed for problems of estimation of the location parameter of a multivariate distribution (see *e.g.* Johnstone, 1988; Fourdrinier and Strawderman, 2003). The principle is classical in statistics and goes as follow: we wish to evaluate the accuracy of a decision rule $\hat{\mu}$ for estimating the unknown location parameter μ (in the linear model (1), we have $\mu = X\beta$). Therefore we define a loss function, which we write $L(\hat{\mu}, \mu)$, measuring the discrepancy between $\hat{\mu}$ and μ. A typical example is the quadratic loss $L(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|^2$. Since $L(\hat{\mu}, \mu)$ depends on the unknown parameter μ, it is unknown as well and can thus be assessed through an estimation using the observations (see for instance Fourdrinier and Wells (2012) and references therein for more details on loss estimation). The main difference with classical approaches (such as SURE) is that we are interested in estimating the actual loss, not its expectation, that is the risk

$$R(\hat{\mu}, \mu) := \mathbb{E}_\mu[L(\hat{\mu}, \mu)]. \tag{4}$$

In this paper we only consider unbiasedness as our notion of optimality. Hence, let us start with the definition of an unbiased estimator of loss in a general setting. This definition of unbiasedness is the one used by Johnstone (1988).

Definition 1 (Unbiasedness)**.** *Let Y be a random vector in $\mathbb{R}^n$ with mean $\mu \in \mathbb{R}^n$, and let $\hat{\mu}$ be any estimator of μ. An estimator $\delta_0(Y)$ of the loss $L(\hat{\mu}, \mu)$ is said to be* unbiased *if, for all $\mu \in \mathbb{R}^n$,*

$$\mathbb{E}_\mu[\delta_0(Y)] = \mathbb{E}_\mu[L(\hat{\mu}, \mu)]$$

where $\mathbb{E}_\mu$ denotes the expectation with respect to the distribution of Y.

This definition of unbiasedness of an estimator of the loss is somehow non standard. The usual notion of unbiasedness is retrieved when considering $\delta_0(Y)$ as an estimator of the risk $R(\hat{\mu}, \mu)$ since the risk is the expectation of the loss, see (4). However, this terminology of loss estimation, due to Sandved (1968) and Li (1985), has been used by other authors (among others Johnstone (1988); Lele (1993); Fourdrinier and Strawderman (2003); Fourdrinier and Wells (2012)). Differences between loss estimation and risk estimation are enlightened by results from Li (1985). Li proves that SURE estimates the loss consistently over the true mean μ as n goes to infinity. He also constructs a simple example where μ is estimated by a particular form of James-Stein shrinkage estimators for which SURE tends asymptotically to a random variable, and hence is inconsistent for the estimation of the risk, which is not random. Another interesting result of Li (1985) is the consistency of the estimator of μ selected using the rule "minimize SURE" (which is equivalent to the rule "minimize the unbiased estimator δ_0"). Although this result has only been proved for the special case of James-Stein type estimators, this is encouraging for choosing such a rule to select the best model from data.

Although in practice SURE and unbiased loss estimators are the same, we believe this discussion makes it clear that the actual loss is a more relevant quantity of interest than the risk. The difference will be important when improving on unbiased estimators, as we will discuss in the perspectives of Section 4.

Obtaining an unbiased estimator of loss requires Stein's identity, the key theorem of loss estimation theory which we recall here for the sake of completeness and whose proof can be found in Stein (1981).

Theorem 1 (Stein's identity)**.** *Let $Y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$. Given $g : \mathbb{R}^n \to \mathbb{R}^n$ a weakly differentiable function, we have, assuming the expectations exist, the following equality*

$$\mathbb{E}_\mu\left[(Y-\mu)^t g(Y)\right] = \sigma^2\, \mathbb{E}_\mu\left[\mathrm{div}_Y g(Y)\right], \tag{5}$$

where $\mathrm{div}_Y g(Y) = \sum_{i=1}^n \partial g_i(Y)/\partial Y_i$ is the weak divergence of $g(Y)$ with respect to Y.

See *e.g.* Section 2.3 in Fourdrinier and Wells (1995) for the definition and the justification of weak differentiability.

2.1.2. Unbiased loss estimation for model (1)

When considering the Gaussian model in (1), we have $\mu = X\beta$, we set $\hat{\mu} = X\hat{\beta}$ and $L(\hat{\beta}, \beta)$ is defined as the quadratic loss $\|X\hat{\beta} - X\beta\|^2$. Special focus will be

given to the quadratic loss since it is the most commonly used and allows simple calculations. In practice, it is a reasonable choice if we are interested in both good selection and good prediction at the same time. Moreover, the quadratic loss allows us to link loss estimation with C_p and AIC.

In the following theorem, an unbiased estimator of the quadratic loss under a Gaussian assumption is provided.

Theorem 2. *Let $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$. Let $\hat{\beta} = \hat{\beta}(Y)$ be a function of the least squares estimator of β such that $X\hat{\beta}$ is weakly differentiable with respect to Y. Let $\hat{\sigma}^2 = \|Y - X\hat{\beta}^{LS}\|^2/(n-p)$. Then*

$$\delta_0(Y) = \|Y - X\hat{\beta}\|^2 + (2\,\mathrm{div}_Y(X\hat{\beta}) - n)\hat{\sigma}^2 \tag{6}$$

is an unbiased estimator of $\|X\hat{\beta} - X\beta\|^2$.

Proof. The risk of $X\hat{\beta}$ at $X\beta$ is

$$\begin{aligned}\mathbb{E}_\beta[\|X\hat{\beta} - X\beta\|^2] \;&=\; \mathbb{E}_\beta[\|X\hat{\beta} - Y\|^2 + \|Y - X\beta\|^2] \\ &\quad + \mathbb{E}_\beta[2(Y - X\beta)^t(X\hat{\beta} - Y)].\end{aligned} \tag{7}$$

Since $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$, we have $\mathbb{E}_\beta[\|Y - X\beta\|^2] = n\,\sigma^2$ leading to

$$\mathbb{E}_\beta[\|X\hat{\beta} - X\beta\|^2] = \mathbb{E}_\beta[\|Y - X\hat{\beta}\|^2] - n\,\sigma^2 + 2\,\mathrm{tr}(\mathrm{cov}_\beta(X\hat{\beta}, Y - X\beta)).$$

Moreover, applying Stein's identity for the right-most part of the expectation in (7) with $g(Y) = X\hat{\beta}$ and assuming that $X\hat{\beta}$ is weakly differentiable with respect to Y, we can rewrite (7) as

$$\mathbb{E}_\beta[\|X\hat{\beta} - X\beta\|^2] = \mathbb{E}_\beta[\|Y - X\hat{\beta}\|^2] - n\,\sigma^2 + 2\,\sigma^2\,\mathbb{E}_\beta\Big[\mathrm{div}_Y X\hat{\beta}\Big].$$

Since $\hat{\sigma}^2$ is an unbiased estimator of σ^2 independent of $\hat{\beta}^{LS}$, the right-hand side of this last equality is also equal to the expectation of $\delta_0(Y)$ given by Equation (6). Hence, according to Definition 1, the statistic $\delta_0(Y)$ is an unbiased estimator of $\|X\hat{\beta} - X\beta\|^2$. □

Remark 1. *It is often of interest to use robust estimators of β and σ^2. In such a case, the hypotheses of the theorem need to be modified to insure the independence between estimators $\hat{\beta}$ and $\hat{\sigma}^2$ which were implicit in the statement of the theorem. We will see in Remark 2 of the next section that, by the use of Stein identity for the general spherical case, the implicit assumption of independence is no longer needed.*

In the following subsections, we discuss our choices for the measure of complexity of a model (through the divergence term) and for the estimation of the variance, and we relate them to other choices from the literature.

2.1.3. Degrees of freedom

It turns out that the divergence term of $\delta_0(Y)$ in (6), $\mathrm{div}_Y X\hat{\beta}$, is related to the estimator of the degrees of freedom $\widehat{df}$ used in Equation (2) for the definition of C_p, and to the number k of parameters proposed for AIC in (3).

A convenient way to establish this connection is to follow Ye (1998) in defining the (generalized) degrees of freedom of an estimator as the trace of the scaled covariance between the prediction $X\hat{\beta}$ and the observation Y

$$df = \frac{1}{\sigma^2}\mathrm{tr}\left(\mathrm{cov}_\beta(X\hat{\beta}, Y)\right). \tag{8}$$

This definition has the advantage of encompassing the effective degrees of freedom proposed for generalized linear models and the standard degrees of freedom used when dealing with the least squares estimator.

When it applies, Stein's identity yields

$$df = \mathbb{E}_\beta[\mathrm{div}_Y X\hat{\beta}].$$

Setting

$$\widehat{df} = \mathrm{div}_Y X\hat{\beta},$$

the statistic $\widehat{df}$ appears as an unbiased estimator of the (generalized) degrees of freedom. In the case of linear estimators, there exists a hat matrix, that is, a matrix H such that $X\hat{\beta} = HY$ and we have

$$\mathrm{div}_Y X\hat{\beta} = \mathrm{div}_Y(HY) = \sum_{i=1}^{n} \frac{\partial \sum_{j=1}^{n} H_{i,j}Y_j}{\partial Y_i} = \sum_{i=1}^{n} H_{i,i} = \mathrm{tr}(H),$$

so that

$$\widehat{df} = \mathrm{tr}(H).$$

This definition of $\widehat{df}$ is the one used by Mallows for the extension of C_p to ridge regression. Note that, in this case, $\widehat{df}$ no longer depends on Y and thus equals its expectation ($df = \widehat{df}$). When H is a projection matrix (*i.e.* when $H^2 = H$) as it is for the least squares estimator, then we have

$$\mathrm{tr}(H) = k,$$

where k is the rank of the projector which is also the number of linearly independent parameters, and thus $df = k$.

In this case the definition of degrees of freedom agrees with intuition. It is the number of parameters of the model that are free to vary. When H is no longer a projector, $\mathrm{rank}(H)$ is no longer a valid measure of complexity since it can be equal to n while for admissible estimators $\mathrm{tr}(H)$ is the trace norm of H (also known as the nuclear norm). The trace norm is also the minimum convex envelope of $\mathrm{rank}(H)$ over the unit ball of the spectral norm, used as a convex proxy for the rank in some optimization problem (see for instance

Recht, Fazel, and Parrilo (2010) and the references therein). As a convex norm, $\mathrm{tr}(H)$ is a continuous measure of the complexity of the associated mapping. For nonlinear estimators, the divergence $\mathrm{div}_Y X\hat{\beta}$ is the trace of the Jacobian matrix (its nuclear norm) of the mapping that produced a set of fitted values for Y. According to Ye (1998), it can be interpreted as "the cost of the estimation process" or as "the sum of the sensitivity of each fitted value to perturbations".

Other measures of the complexity, involving the trace or the determinant of the Fisher information matrix $\mathcal{I}$, have been used for instance by Takeuchi (1976) and Bozdogan (1987). However, such measures depend on the specific form of the distribution considered which does not suit our context. The Vapnik-Chervonenkis dimension (VC-dim, Vapnik and Chervonenkis (1971)), is another way to capture the complexity of the model. However, it can be difficult to compute for nonlinear estimators while it is equivalent to our divergence term for linear estimators. Hence, Cherkassky and Ma (2003) proposed to estimate the VC-dim by $\widehat{df}$.

2.1.4. Estimators of the variance

The issue of estimating the variance was clearly pointed out in Cherkassky and Ma (2003). The authors proposed two estimators of the variance, one for the full model together with $\hat{\beta}^{LS}$ the least-squares estimator

$$\hat{\sigma}^2_{full} = \frac{\|Y - X\hat{\beta}^{LS}\|^2}{n-p}, \tag{9}$$

and the second one for the model restricted to a subset $I \subset \{1, \dots, p\}$

$$\hat{\sigma}^2_{restricted} = \frac{\|Y - X\hat{\beta}_I\|^2}{n-k}, \tag{10}$$

where k is the size of I and $\hat{\beta}_I$ is linear in Y. Note that other estimators of σ^2 can be found in the literature (see for instance Arlot and Bach, 2009). As mentioned earlier, we are concerned with unbiasedness of the loss, so that the estimator of σ^2 should be unbiased and independent of $\mathrm{div}_Y X\hat{\beta}$. Hence the choice between $\hat{\sigma}^2_{full}$ and $\hat{\sigma}^2_{restricted}$ should be made with respect to what we believe is the true model, either the full model in (1) or the restricted model

$$Y = X_I\beta_I + \sigma\,\varepsilon.$$

In the general case where $\hat{\beta}_I$ is not necessarily linear in Y, the variance is usually estimated by (10) where k is replaced by $\widehat{df}$. However, in such a case, it is not clear whether this estimator is independent of Y or not. Also, the use of $\hat{\sigma}^2_{restricted}$ might not lead to an equivalence between the unbiased estimator of loss (24) and AIC. Indeed, in such a case the unbiased estimator of loss would have the form $k\| Y - X\hat{\beta}_I\|^2/(n-k)$, while AIC would result in $\log\hat{\sigma}^2_{restricted} + 2k$. This is just model selection based on the minimum standard error. On the contrary, when considering the estimator $\hat{\sigma}^2_{full}$, the equivalence is pretty clear, as we will see in the next section.

2.2. Links between loss estimation and model selection

2.2.1. Selection criteria

In order to make the following discussion clearer, we recall here the formula of the three criteria of interest for the Gaussian assumption, namely the unbiased estimator of loss $\delta_0(Y)$, Mallows' C_p and the extended version of AIC proposed by Ye (1998):

$$\begin{aligned} \delta_0(Y) &= \|Y - X\hat{\beta}\|^2 + (2\,\mathrm{div}_Y(X\hat{\beta}) - n)\hat{\sigma}^2 \\ C_p &= \frac{\|Y - X\hat{\beta}\|^2}{\hat{\sigma}^2} + 2\,\mathrm{div}_Y(X\hat{\beta}) - n \\ \mathrm{AIC} &= \frac{\|Y - X\hat{\beta}\|^2}{\hat{\sigma}^2} + 2\,\mathrm{div}_Y(X\hat{\beta}). \end{aligned}$$

Using the estimator $\hat{\sigma}^2_{full}$ of the variance in (9), we thus obtain the following link between δ_0, C_p and AIC:

$$\delta_0(Y) = \hat{\sigma}^2_{full} \times C_p = \hat{\sigma}^2_{full} \times (\mathrm{AIC} - n). \tag{11}$$

These links between different criteria function for model selection are due to the fact that, under our working hypothesis (linear model, quadratic loss, normal distribution $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$ for a fixed design matrix X), they can be seen as unbiased estimators of related quantities of interest.

Note that there is also an equivalence with other model selection criteria, investigated for instance in Li (1985), Shao (1997) and Efron (2004).

2.2.2. Quality of model selection procedures

Given a loss estimator $\widehat{L}(\hat{\beta}_I, \beta)$ (such as δ_0), the corresponding model selection procedure consists in finding the best subset of variables, $\hat{I}$, that is,

$$\hat{I} = \arg\min_I \ \widehat{L}(\hat{\beta}_I, \beta)$$

where $\hat{\beta}_I$ is the chosen estimator associated to the subset I of variables. To assess the quality of this model selection procedure, from a nonasymptotic point of view (fixed n), Donoho and Johnstone (1994) introduced the notion of an oracle. An oracle is supposed to determine the ideal subset, $I^\star$, that is,

$$I^\star = \arg\min_I \ L(\hat{\beta}_I, \beta).$$

A good model selection procedure approaches the ideal performance. More formally, the associated oracle inequality states that, with high probability,

$$L(\hat{\beta}_{\hat{I}}, \beta) \leq L(\hat{\beta}_{I^\star}, \beta) + \epsilon. \tag{12}$$

That is to say for any α, $0 < \alpha < 1$, there is a positive ϵ such that

$$\mathbb{P}\left(L(\hat{\beta}_{\hat{I}}, \beta) - L(\hat{\beta}_{I^\star}, \beta) \geq \epsilon\right) \leq \alpha. \tag{13}$$

Noticing that $\widehat{L}(\hat{\beta}_{\hat{I}}, \beta) \leq \widehat{L}(\hat{\beta}_{I^\star}, \beta)$ we can write

$$\begin{aligned}
\mathbb{P}\left(L(\hat{\beta}_{\hat{I}}, \beta) - L(\hat{\beta}_{I^\star}, \beta) \geq \epsilon\right) &\leq \mathbb{P}\left(|\widehat{L}(\hat{\beta}_{I^\star}, \beta) - L(\hat{\beta}_{I^\star}, \beta)| \geq \frac{\epsilon}{2}\right) \\
&\quad + \mathbb{P}\left(|\widehat{L}(\hat{\beta}_{\hat{I}}, \beta) - L(\hat{\beta}_{\hat{I}}, \beta)| \geq \frac{\epsilon}{2}\right) \\
&\leq \sum_I \mathbb{P}\left(|\widehat{L}(\hat{\beta}_I, \beta) - L(\hat{\beta}_I, \beta)| \geq \frac{\epsilon}{2}\right)
\end{aligned}$$

considering all the subsets I. Then by Chebyshev's inequality

$$\begin{aligned}
\mathbb{P}\left(L(\hat{\beta}_{\hat{I}}, \beta) - L(\hat{\beta}_{I^\star}, \beta) \geq \epsilon\right) &\leq \sum_I \frac{4}{\epsilon^2}\, \mathbb{E}\left[|\widehat{L}(\hat{\beta}_I, \beta) - L(\hat{\beta}_I, \beta)|^2\right] \\
&= \frac{4}{\epsilon^2} \sum_I \mathcal{R}\big(\widehat{L}(\hat{\beta}_I, \beta)\big),
\end{aligned}$$

where $\mathcal{R}\big(\widehat{L}(\hat{\beta}_I, \beta)\big)$ is the statistical risk of a further quadratic loss function

$$\mathcal{R}\big(\widehat{L}(\hat{\beta}_I, \beta)\big) := \mathbb{E}\left[|\widehat{L}(\hat{\beta}_I, \beta) - L(\hat{\beta}_I, \beta)|^2\right]. \tag{14}$$

Hence, for any α we can find ϵ, say,

$$\epsilon = 2\sqrt{\sum_I \frac{\mathcal{R}\big(\widehat{L}(\hat{\beta}_I, \beta)\big)}{\alpha}} \tag{15}$$

such that the oracle inequality (12) is satisfied. It is clear that a way of controlling the oracle bound of the selector is to choose a loss estimator with small quadratic risk. In the case where the loss estimator is of the form squared residual plus a penalty function the oracle condition above translates into a sufficient condition on the behavior of the penalty term. Classical oracle inequality analysis gives an exponential bound on the left-hand side of (13). The development of such a result requires a concentration type inequality for the maximum of Gaussian processes (see Massart (2007)) to give an exponential upper bound on (12) that tends to zero. Our goal here is complementary to the standard oracle inequality approach; that is, we developed a novel upper bound that links the quality of the model selection procedure to the risk assessment of a loss estimator. This idea is further elucidated in Barron, Birgé, and Massart (1999); Birgé and Massart (2007) and related work. Note that, in particular, C_p has been proven to satisfy an oracle inequality by Baraud (2000).

Bartlett, Boucheron, and Lugosi (2002) studied model selection strategies based on penalized empirical loss minimization in the context of bounded loss.

They prove, using concentration inequality techniques, the equivalence between loss estimation and data-based complexity penalization. It was shown that good loss estimates may be converted into a data-based penalty function and the performance of the estimate is governed by the quality of the loss estimate. Furthermore, it was shown that a selected model that minimizes the empirical loss achieves an almost optimal trade-off between the approximation error and the expected complexity. The key point to stress is that the results of Bartlett et al. are concordant with the oracle bound in (15), that is there is a fundamental dependence on the notions of good complexity regularization and good loss estimation.

2.2.3. Model selection

The final objective is to select the "best" model among those at hand. This can be performed by minimizing either of the three proposed criteria, that is the unbiased estimator of loss δ_0, C_p and AIC. The idea behind this heuristic, as shown in the previous section, is that the best model in terms of prediction is the one minimizing the estimated loss. Now, from (11), it can be easily seen that the three criteria differ from each other only up to a multiplicative and/or additive constant. Hence the models selected by the three criteria will be the same.

We would like to point out that Theorem 2 does not rely on the linearity of the link between X and Y so that this work can easily be extended to nonlinear links at no extra cost. Therefore δ_0 generalizes C_p to nonlinear models. Moreover, following its definition (3), AIC implementation requires the specification of the underlying distribution. In this sense it is considered as a generalization of C_p for non Gaussian distributions. However, in practice, we might only have a vague intuition of nature of the underlying distribution and we might not be able to give its specific form. We will see in the following section that δ_0, which is equivalent to the Gaussian AIC as we have just seen, can be also derived from a more general distribution context, that of spherically symmetric distributions, with no need to specify the precise form of the distribution.

3. Unbiased loss estimators for spherically symmetric distributions

3.1. Multivariate spherical distributions

In the previous section, results were given under the Gaussian assumption with covariance matrix $\sigma^2 I_n$. In this section, we extend this distributional framework.

The characterization of the normal distribution as expressed by Kariya and Sinha (1989) allows two directions for generalization. Indeed, the authors assert that a random vector is Gaussian, with covariance matrix proportional to the identity matrix, if and only if its components are independent and its law is spherically symmetric. Hence, we can generalize the Gaussian assumption by

either keeping the independence property and consider other laws than the Gaussian, or by relaxing the independence assumption to the benefit of spherical symmetry. In the same spirit, Fan and Fang (1985) pointed out that there are two main generalizations of the Gaussian assumption in the literature: one generalization comes from the interesting properties of the exponential form and leads to the exponential family of distributions, while the other is based on the invariance under orthogonal transformation and results in the family of spherically symmetric distributions (which can be generalized by elliptically contoured distributions). These generalizations go in different directions and have lead to fruitful works. Note that their only common member is the Gaussian distribution. The main interest of choosing spherical distributions is that the conjunction of invariance under orthogonal transformation together with linear regression with less variables than observations brings robustness. The interest of that property is illustrated by the fact that some statistical tests designed under a Gaussian assumption, such as Student and Fisher tests, remain valid for spherical distributions Wang and Wells (2002); Fang and Anderson (1990). This robustness property is not shared by independent non-Gaussian distributions, as mentioned in Kariya and Sinha (1989).

Note that, from the model in (1), the distribution of Y is the distribution of $\sigma\,\varepsilon$ translated by $\mu = X\beta$: Y has a spherically symmetric distribution about the location parameter μ with covariance matrix equal to $n^{-1}\sigma^2\mathbb{E}[\|\varepsilon\|^2]I_n$ where I_n is the identity matrix. We write $\varepsilon \sim \mathcal{S}_n(0, I_n)$ and $Y \sim \mathcal{S}_n(\mu, \sigma^2 I_n)$. Examples of this family besides the Gaussian distribution $\mathcal{N}_n$ are the Student distribution $\mathcal{T}_n$, the Kotz distribution $\mathcal{K}_n$, or else variance mixtures of Gaussian distributions $\mathcal{GM}_n$ whose densities are respectively given by

$$
\begin{aligned}
\mathcal{N}_n(y;\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\|y-\mu\|^2}{2\sigma^2}} \\
\mathcal{T}_n(y;\mu,\sigma^2,\nu) &= \frac{\Gamma[(n+\nu)/2]}{(\pi\sigma^2)^{n/2}\Gamma(\nu/2)\nu^{n/2}} \left[1 + \frac{\|y-\mu\|^2}{\nu\sigma^2}\right]^{-\frac{(\nu+n)}{2}} \\
\mathcal{K}_n(y;\mu,\sigma^2,N,r) &= \frac{\Gamma(n/2) r^{N-1+n/2}}{\pi^{\frac{n}{2}}\sigma^{n+2-2N}\Gamma\left(N-1+\frac{n}{2}\right)} \|y-\mu\|^{2(N-1)} e^{-r\frac{\|y-\mu\|^2}{\sigma^2}} \\
\mathcal{GM}_n(y;\mu,\sigma^2,G) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \int_0^\infty \frac{1}{v^{n/2}} e^{-\frac{\|y-\mu\|^2}{2v\sigma^2}} G(dv)\,,
\end{aligned}
$$

where $\nu \geq 1$, $2N + n > 2$, $r > 0$, and $G(\cdot)$ is a probability measure on the scale parameter v. Here, Γ denotes the Gamma function. Note that the Gaussian distribution is a special case of Kotz distribution with $r = 1/2$ and $N = 1$, and of Gaussian mixtures, while it is the limiting distribution of the Student law when ν tends to infinity. Figure 1 shows the shape of these densities in two dimensions ($n = 2$).

As we will see in the sequel, the unbiased estimator of the quadratic loss δ_0 (24) remains unbiased for any of these distributions with no need to specify its form. It thus brings distributional robustness. For more details and examples of the spherical family, we refer the interested reader to Chmielewski (1981) for a

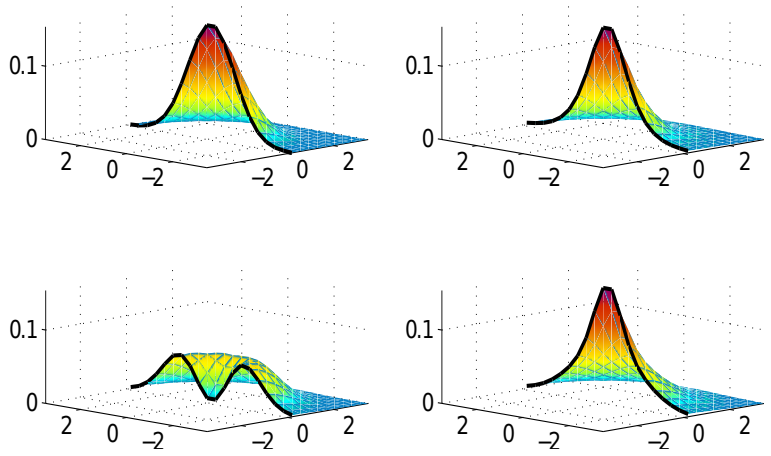

FIGURE 1. *Examples of spherically symmetric densities for a two-dimensional random variable with $\mu = 0$ and $\sigma^2 = 1$: Gaussian (top left), Student with $\nu = 1$ or Cauchy (top right), Kotz with $N = 2$ and $r = 1$ (bottom left), mixture of centered Gaussian with $\mathbb{P}[v = 0.1] = 0.3$ and $\mathbb{P}[v = 5] = 0.7$ (bottom right). Note that the Student and the mixture of centered Gaussian distributions have heavier tails than the Gaussian law.*

historical review and Fang and Zhang (1990) for a comprehensive presentation.

3.2. The canonical form of the linear regression model

An efficient way of dealing with the linear regression model under spherically symmetric distributions is to use its canonical form (see Fourdrinier and Wells (1995) for details). This form will allow us to give more straightforward proofs.

Considering model (1), the canonical form consists in an orthogonal transformation of Y. Using partitioned matrices, let $Q = (Q_1\ Q_2)$ be an $n \times n$ orthogonal matrix partitioned such that the first p columns of Q (*i.e.* the columns of Q_1) span the column space of X. For instance, this is the case of the Q-R factorization where $X = QR$ with R an $n \times p$ upper triangular matrix. Now, according to (1), let

$$Q^t Y = \begin{pmatrix} Q_1^t \\ Q_2^t \end{pmatrix} Y = \begin{pmatrix} Q_1^t \\ Q_2^t \end{pmatrix} X\beta + \sigma Q^t \varepsilon = \begin{pmatrix} \theta \\ \mathbf{0} \end{pmatrix} + \sigma Q^t \varepsilon \tag{16}$$

with $\theta = Q_1^t X\beta$ and $Q_2^t X\beta = \mathbf{0}$ since the columns of Q_2 are orthogonal to whose of X. It follows from the definition that $(Z^t\, U^t)^t := Q^t Y$ has a spherically symmetric distribution about $(\theta^t\, \mathbf{0}^t)^t$. In this sense, the model

$$\begin{pmatrix} Z \\ U \end{pmatrix} = \begin{pmatrix} \theta \\ \mathbf{0} \end{pmatrix} + \sigma \begin{pmatrix} Q_1^t \varepsilon \\ Q_2^t \varepsilon \end{pmatrix}$$

is the canonical form of the linear regression model (1).

This canonical form has been considered by various authors such as Cellier, Fourdrinier, and Robert (1989); Cellier and Fourdrinier (1995); Maruyama (2003); Maruyama and Strawderman (2005, 2009); Fourdrinier and Strawderman (2010); Kubokawa and Srivastava (1999). Kubokawa and Srivastava (2001) addressed the multivariate case where θ is a mean matrix (in this case Z and U are matrices as well).

For any estimator $\hat{\beta}$, the orthogonality of Q implies that

$$\|Y - X\hat{\beta}\|^2 \quad = \quad \|\hat{\theta} - Z\|^2 + \|U\|^2 \tag{17}$$

where $\hat{\theta} = Q_1^t X\hat{\beta}$ is the corresponding estimator of θ. In particular, for the least squares estimator $\hat{\beta}^{LS}$, we have

$$\|Y - X\hat{\beta}^{LS}\|^2 = \|U\|^2. \tag{18}$$

In that context, we recall the Stein-type identity given by Fourdrinier and Wells (1995).

Theorem 3 (Stein-type identity)**.** *Given $(Z, U) \in \mathbb{R}^n$ a random vector following a spherically symmetric distribution around $(\theta, \mathbf{0})$, and $g : \mathbb{R}^p \to \mathbb{R}^p$ a weakly differentiable function, we have*

$$\mathbb{E}_\theta[(Z-\theta)^t g(Z)] = \mathbb{E}_\theta[\|U\|^2 \mathrm{div}_Z g(Z)/(n-p)], \tag{19}$$

provided both expectations exist.

Note that the divergence in Theorem 2 is taken with respect to Y while the Stein type identity (19) requires the divergence with respect to Z (with the assumption of weak differentiability). Their relationship can be seen in the following lemma.

Lemma 1. *We have*

$$\mathrm{div}_Y X\hat{\beta}(Y) = \mathrm{div}_Z \hat{\theta}(Z, U)\,. \tag{20}$$

Proof. Denoting by $\mathrm{tr}(A)$ the trace of any matrix A and by $J_f(x)$ the Jacobian matrix (when it exists) of a function f at x, we have

$$\mathrm{div}_Y X\hat{\beta}(Y) = \mathrm{tr}\big(J_{X\hat{\beta}}(Y)\big) = \mathrm{tr}\big(Q^t\, J_{X\hat{\beta}}(Y)\, Q\big)$$

by definition of the divergence and since Q^t is an orthogonal matrix. Now, setting $W = Q^t\, Y$, *i.e.* $Y = Q\, W$, applying the chain rule to the function

$$\widehat{T}(W) = Q^t X\hat{\beta}(Q\, W)$$

gives rise to

$$J_{\widehat{T}}(W) = J_{Q^t X\hat{\beta}}(Y)\, Q = Q^t\, J_{X\hat{\beta}}(Y)\, Q, \tag{21}$$

noticing that Q^t is a linear transformation.

Also, as according to (16)

$$W = \begin{pmatrix} Z \\ U \end{pmatrix} \quad \text{and} \quad \widehat{T} = \begin{pmatrix} \hat{\theta} \\ \mathbf{0} \end{pmatrix},$$

the following decomposition holds

$$J_{\widehat{T}}(W) \;=\; \begin{pmatrix} J_{\hat{\theta}}(Z) & J_{\hat{\theta}}(U) \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where $J_{\hat{\theta}}(Z)$ and $J_{\hat{\theta}}(U)$ are the parts of the Jacobian matrix in which the derivatives are taken with respect to the components of Z and U respectively. Thus

$$\operatorname{tr}\big(J_{\widehat{T}}(W)\big) = \operatorname{tr}\big(J_{\hat{\theta}}(Z)\big) \tag{22}$$

and, therefore, gathering (21) and (22), we obtain

$$\operatorname{tr}\big(J_{\hat{\theta}}(Z)\big) = \operatorname{tr}\big(Q^t\, J_{X\hat{\beta}}(Y)\, Q\big) = \operatorname{tr}\big(QQ^t\, J_{X\hat{\beta}}(Y)\big) = \operatorname{tr}\big(J_{X\hat{\beta}}(Y)\big),$$

which is (20) by definition of the divergence. □

3.3. Unbiased estimator of loss for the spherical case

This section is devoted to the generalization of Theorem 2 to the class of spherically symmetric distributions $Y \sim \mathcal{S}_n(X\beta, \sigma^2)$, given by Theorem 4. To do so we need to consider the statistic

$$\hat{\sigma}^2(Y) = \tfrac{1}{n-p}\|Y - X\hat{\beta}^{LS}\|^2. \tag{23}$$

It is an unbiased estimator of $\sigma^2\mathbb{E}_\beta[\|\varepsilon\|^2/n]$. Note that, in the normal case where $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$, we have $\mathbb{E}_\beta[\|\varepsilon\|^2/n] = 1$ so that $\hat{\sigma}^2(Y)$ is an unbiased estimator of σ^2.

Theorem 4 (Unbiased estimator of the quadratic loss under spherical assumption)**.** *Let $Y \sim \mathcal{S}_n(X\beta, \sigma^2 I_n)$ and let $\hat{\beta} = \hat{\beta}(Y)$ be an estimator of β depending only on $Q_1^t Y$. If $\hat{\beta}(Y)$ is weakly differentiable with respect to Y, then the statistic*

$$\delta_0(Y) = \|Y - X\hat{\beta}(Y)\|^2 + \big(2\,\mathrm{div}_Y(X\hat{\beta}(Y)) - n\big)\,\hat{\sigma}^2(Y)\,, \tag{24}$$

is an unbiased estimator of $\|X\hat{\beta}(Y) - X\beta\|^2$.

Proof. The quadratic loss function of $X\hat{\beta}$ at $X\beta$ can be decomposed as

$$\|X\hat{\beta} - X\beta\|^2 = \|Y - X\hat{\beta}\|^2 + \|Y - X\beta\|^2 + 2\,(X\hat{\beta} - Y)^t(Y - X\beta)\,. \tag{25}$$

The expectation of the second term in the right hand side of (25) has been considered in (23). As for the third term, by orthogonally invariance of the

inner product,

$$\begin{aligned}(X\hat{\beta}-Y)^t(Y-X\beta) &= (Q^tX\hat{\beta}-Q^tY)^t(Q^tY-Q^tX\beta)\\ &= \begin{pmatrix}Q_1^tX\hat{\beta}-Q_1^tY\\ Q_2^tX\hat{\beta}-Q_2^tY\end{pmatrix}^t\begin{pmatrix}Q_1^tY-Q_1^tX\beta\\ Q_2^tY-Q_2^tX\beta\end{pmatrix}\\ &= \begin{pmatrix}\hat{\theta}-Z\\ -U\end{pmatrix}^t\begin{pmatrix}Z-\theta\\ U\end{pmatrix}\\ &= (\hat{\theta}-Z)^t(Z-\theta)-\|U\|^2\,. \end{aligned} \tag{26}$$

Now, since $\hat{\theta}=\hat{\theta}(Z,U)$ depends only on Z, by Stein type identity, we have

$$\begin{aligned}\mathbb{E}\Big[(\hat{\theta}-Z)^t(Z-\theta)\Big] &= \mathbb{E}\left[\frac{\|U\|^2}{n-p}\operatorname{div}_Z(\hat{\theta}-Z)\right]\\ &= \mathbb{E}\left[\frac{\|U\|^2}{n-p}\Big(\operatorname{div}_Z\hat{\theta}-p\Big)\right]\end{aligned} \tag{27}$$

so that

$$\begin{aligned}\mathbb{E}[(X\hat{\beta}-Y)^t(Y-X\beta)] &= \mathbb{E}\left[\frac{\|U\|^2}{n-p}\Big(\operatorname{div}_Z\hat{\theta}-p\Big)-\|U\|^2\right]\\ &= \mathbb{E}\left[\frac{\|U\|^2}{n-p}\operatorname{div}_Z\hat{\theta}-\frac{n}{n-p}\|U\|^2\right]\\ &= \mathbb{E}[\hat{\sigma}^2(Y)\operatorname{div}_YX\hat{\beta}-n\,\hat{\sigma}^2(Y)]\end{aligned} \tag{28}$$

by (18) and since $\operatorname{div}_Z\hat{\theta}=\operatorname{div}_YX\hat{\beta}$ by Lemma 1. Finally, gathering (25), (23) and (28) gives the desired result. □

From the equivalence between δ_0, C_p and AIC under a Gaussian assumption, and the unbiasedness of δ_0 under the wide class of spherically symmetric distributions, we conclude that C_p and AIC derived under the Gaussian distribution still can be considered as good selection criteria for spherically symmetric distributions, although their original properties may not have been verified in this context.

Remark 2. *Note that the extension of Stein's lemma in Theorem 3 implies that* $\widehat{df}=\operatorname{div}_YX\hat{\beta}$ *is also an unbiased estimator of df under the spherical assumption. Moreover, we would like to point out that the independence of* $\hat{\sigma}^2$ *used in the proof of Theorem 2 in the Gaussian case is no longer necessary. Also, to require that* $\hat{\beta}$ *depends on* Q_1^tY *only is equivalent to say that* $\hat{\beta}$ *is a function of the least squares estimator. When this hypothesis is not available, an extended Stein type identity can be derived (Fourdrinier, Strawderman, and Wells (2003)).*

4. Discussion

In this work, we studied the well-known model selection criteria C_p and AIC through a loss estimation approach and related them to an unbiased estimator

of the quadratic prediction loss under a Gaussian assumption. We then derived the unbiased estimator of loss under a wider distributional setting, the family of spherically symmetric distributions. Under this context, the unbiased estimator of loss is actually equal to the one derived under the Gaussian law. Hence, this implies that we do not have to specify the form of the distribution, the only condition being its spherical symmetry. We also conclude from the equivalence between unbiased estimators of loss, C_p and AIC that their form for the Gaussian case is able to handle any spherically symmetric distribution. The spherical family is interesting for many practical cases since it allows a dependence property between the components of random vectors whenever the distribution is not Gaussian. Some members of this family also have heavier tails than the Gaussian law, and thus the unbiased estimator derived here can be robust to outliers. A generalization of this work with elliptically symmetric distributions for the error vector would go even further by taking into account a general covariance matrix Σ. We intend to study this case in future work.

It is well known that unbiased estimators of loss are not the best estimators and can be improved. It was not our intention in this work to show better results of such estimators, but our result explains why their performances can be similar when departing from the Gaussian assumption. The improvement of these unbiased estimators requires a way to assess their quality. This can be done either using oracle inequalities or the theory of admissible estimators under a certain risk. These two points of view are closely related. Based on our definition of the risk (14) a selection rule δ_0 is inadmissible if we can find a better estimator, say δ_γ, that has a smaller risk function for all possible values of the parameter β, that is, with stict inequality for some β. The heuristic of loss estimation is that the closer an estimator is to the true loss, the more we expect their respective minima to be close. We are currently working on improved estimators of loss of the type $\delta_\gamma(Y) = \delta_0(Y) + \gamma(Y)$, where $\gamma - 2k\hat{\sigma}$ can be thought of as a data driven penalty. From another point of view, choosing a γ improvement term, decreases the oracle bound given in (15). The selection of such a γ term is an important research direction.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 46–54. NIPS, 2009.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117(4):467–493, 2000.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1):33–73, 2007.

H. Bozdogan. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*. Springer Verlag, 2002.

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

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.

V. Cherkassky and Y. Ma. Comparison of model selection for regression. *Neural Computation*, 15(7):1691–1714, 2003.

M.A. Chmielewski. Elliptically symmetric distributions: A review and bibliography. *International Statistical Review/Revue Internationale de Statistique*, 49:67–74, 1981.

D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425, 1994.

B. Efron. The estimation of prediction error. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

J.Q. Fan and K.T. Fang. Inadmissibility of sample mean and regression coefficients for elliptically contoured distributions. *Northeastern Mathematical Journal*, 1:68–81, 1985.

K.T. Fang and T.W. Anderson. *Statistical Inference in Elliptically Contoured and Related Distributions*. Allerton Pr, 1990.

K.T. Fang and Y.T. Zhang. *Generalized Multivariate Analysis*. Science Press Beijing, 1990.

D.P. Foster and E.I. George. The risk inflation criterion for multiple regression. *Annals of Statistics*, 22(4):1947–1975, 1994.

D. Fourdrinier and W.E. Strawderman. On Bayes and unbiased estimators of loss. *Annals of the Institute of Statistical Mathematics*, 55(4):803–816, 2003.

D. Fourdrinier and W.E. Strawderman. Robust generalized Bayes minimax estimators of location vectors for spherically symmetric distribution with unknown scale. *IMS Collections, Institute of Mathematical Statistics, A Festschrift for Lawrence D. Brown*, 6:249–262, 2010.

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.

D. Fourdrinier and M.T. Wells. On improved loss estimation for shrinkage estimators. *Statistical Science*, 27(1):61–81, 2012.

Dominique Fourdrinier, William E. Strawderman, and Martin T. Wells. Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *Journal of multivariate analysis*, 85(1):24–39, 2003.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd Edition)*, volume 1. Springer Series in Statistics, 2008.

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC, 1990.

I.M. Johnstone. On inadmissibility of some unbiased estimates of loss. *Statistical Decision Theory and Related Topics*, 4(1):361–379, 1988.

T. Kariya and B.K. Sinha. *Robustness of Statistical Tests*, volume 1. Academic Press, 1989.

T. Kubokawa and M. S. Srivastava. Robust improvement in estimation of a covariance matrix in an elliptically contoured distribution. *Annals of Statistics*, 27(2):600–609, 1999.

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

C. Lele. Admissibility results in loss estimation. *Annals of Statistics*, 21(1): 378–390, 1993.

K.C. Li. From Stein's unbiased risk estimates to the method of generalized cross validation. *Annals of Statistics*, 13(4):1352–1377, 1985.

C.L. Mallows. Some comments on Cp. *Technometrics*, 15(4):661–675, 1973.

Y. Maruyama. A robust generalized Bayes estimator improving on the James-Stein estimator for spherically symmetric distributions. *Statistics and Decisions*, 21:69–78, 2003.

Y. Maruyama and W. E. Strawderman. An extended class of minimax generalized Bayes estimators of regression coefficients. *Journal of Multivariate Analysis*, 100:2155–2166, 2009.

Yuzo Maruyama and William E. Strawderman. A new class of generalized Bayes minimax ridge regression estimators. *Ann. Statist.*, 33(4):1753–1770, 2005. ISSN 0090-5364.

P. Massart. *Concentration Inequalities and Model Selection: École d'Été de Probabilités de Saint-Flour XXXIII-2003*. Number 1896. Springer-Verlag, 2007.

M. Meyer and M. Woodroofe. On the degrees of freedom in shape-restricted regression. *Annals of Statistics*, 28(4):1083–1104, 2000.

B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3): 471–501, 2010.

E. Sandved. Ancillary statistics and estimation of the loss in estimation problems. *Annals of Mathematical Statistics*, 39(5):1756–1758, 1968.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2): 461–464, 1978.

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7: 221–242, 1997.

R. Shibata. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, 8(1):147–164, 1980.

C.M. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.

K. Takeuchi. Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18, 1976.

V.N. Vapnik and A.Y. Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Teoriya veroyatnostei i ee primeneniya*, 16(2): 264–279, 1971.

W. Wang and M.T. Wells. Power analysis for linear models with spherical errors. *Journal of Statistical Planning and Inference*, 108(1-2):155–171, 2002.

J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998.