The Loss Rank Criterion for Variable Selection in Linear Regression Analysis

Minh Ngoc Tran[*]

Department of Statistics and Applied Probability

National University of Singapore, Singapore 117546

`ngoctm@nus.edu.sg`

November 8, 2010

Abstract

Lasso and other regularization procedures are attractive methods for variable selection, subject to a proper choice of shrinkage parameter. Given a set of potential subsets produced by a regularization algorithm, a consistent model selection criterion is proposed to select the best one among this preselected set. The approach leads to a fast and efficient procedure for variable selection, especially in high-dimensional settings. Model selection consistency of the suggested criterion is proven when the number of covariates d is fixed. Simulation studies suggest that the criterion still enjoys model selection consistency when d is much larger than the sample size. The simulations also show that our approach for variable selection works surprisingly well in comparison with existing competitors. The method is also applied to a real data set.

Keywords

Model selection, lasso, loss rank principle, shrinkage parameter, variable selection

---

1

1 Introduction

Variable selection is probably the most fundamental and important topic in linear regression analysis. We consider the case where a large number (even larger than the sample size) of candidate covariates are introduced at the initial stage of modeling. One then has to select a smaller subset of the covariates to fit/interpret the data. If the number of potential covariates is not so large (as small as 30), one may use subset selection to select significant variables (Miller, 1990). However, with a large number of covariates, searching on model space is computationally infeasible. Lasso (Tibshirani, 1996) and other regularization procedures (e.g., the adaptive lasso of Zou (2006), SCAD of Fan and Li (2001)) are successful methods to overcome this problem. A Lasso-type procedure estimates the regression coefficient vector $\boldsymbol{\beta}$ by minimizing the sum of the squared error and a regularization term

$$\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda T(\boldsymbol{\beta}), \tag{1}$$

where X is an $(n \times d)$ non-random design matrix, $\boldsymbol{y}$ is an $n-$vector of responses, and $\lambda \geq 0$ is a shrinkage parameter that controls the amount of regularization. The regularization function $T(\boldsymbol{\beta})$ can take different forms according to different regularization procedures. The original and most popular one used in Lasso is the l_1 norm $T(\boldsymbol{\beta}) = \sum_{j=1}^d |\beta_j|$. As λ increases, the coefficients are continuously shrunk towards 0. When λ is sufficiently large, some coefficients are shrunk to exact 0, thus leading to sparse solutions. This feature makes the Lasso-type procedures very attractive for variable selection. Indeed, their model selection consistency has been shown (Zhao and Yu, 2006; Meinshausen and Buhlmann, 2006; Fan and Li, 2001): Under some conditions, there exists a "proper" sequence of shrinkage parameters $\{\lambda_n\}$ under which

$$\{j : \hat{\beta}_j^{\lambda_n} \neq 0\} = \mathcal{S}_T \ \text{ w.p.1 when sample size } n \text{ is large enough}, \tag{2}$$

where $\hat{\boldsymbol{\beta}}_{\lambda_n} = (\hat{\beta}_1^{\lambda_n}, ..., \hat{\beta}_d^{\lambda_n})^\top$ is the regularized estimator of $\boldsymbol{\beta}$ with shrinkage parameter λ_n, and $\mathcal{S}_T$ is the true model, i.e., $\mathcal{S}_T$ is the index set of true covariates. Therefore, it is convenient to use the Lasso-type procedures for variable selection purposes.

The remaining problem in practice is how to choose such proper λ_n. A widely-used criterion is the generalized cross-validation criterion (GCV) (Craven and Wahba, 1979; Tibshirani, 1996). However, theoretical properties of GCV for choosing λ for the purpose of variable selection have not been investigated yet. Furthermore, for choosing shrinkage parameter for the SCAD method (Fan and Li, 2001), a regularization method closely related to Lasso, GCV seems to be likely to choose shrinkage parameters that produce overfitted models (Wang et al., 2007). Zou et al. (2007) showed that the number of nonzero coefficients is an unbiased estimate for the degrees of freedom of the lasso. As a result, popular model selection criteria - like AIC, BIC and C_p - can be used for selecting λ. However, theoretical properties of the selected model remain unknown. The main

contribution of this paper is to propose a criterion for selecting shrinkage parameters in order for regularization procedures to produce the true model. (Throughout this paper, the true model is assumed to exist. We note, however, that whether or not the true model exists is still a controversial issue in the model selection literature, see Burnham and Anderson (2002)).

Although regularization procedures can be used for simultaneous variable selection and estimation, it seems to be impossible to tune the shrinkage parameter to achieve both model selection consistency and optimal estimation at the same time. For an orthogonal design, Leng et al. (2006) showed that the Lasso estimator that is optimal in terms of estimation does not give consistent model selection. This fact was also shown by Poetscher and Leeb (2009) for other regularized estimators. We are in this paper primarily concerned with the problem of variable selection, i.e., we use a Lasso-type procedure to produce a set of potential subsets and then select the best one among this preselected set using a model selection criterion. It was brought to our attention by a reviewer that the idea of using Lasso as a "selector" was also briefly mentioned in Friedman (2008); Efron et al. (2004). They discussed an approach to reduce estimation bias on the non-zero estimated coefficients in which the Lasso (with some method for choosing the shrinkage parameter) is used as a subset selector and then a different unpenalized procedure is used to estimate the coefficients w.r.t. the selected covariates. Our approach is to select the best subset of covariates - using a model selection criterion as the stopping rule - among a preselected set of potential subsets produced by a Lasso-type procedure. The preselected set consists of at most d subsets rather than 2^d possible subsets if using subset selection. After selecting the best subset, we of course can use an unpenalized procedure to estimate the coefficients in order to reduce estimation bias.

The model selection criterion we use is derived from the *loss rank principle* (LoRP), a general-purpose principle for model selection, introduced recently by Hutter (2007); Hutter and Tran (2010). LoRP selects a model that has the smallest *loss rank*. The loss rank of a model is defined as the number of other "fictitious" data that fit the model better than the training data (see Section 2 for a formal introduction). It was shown by Hutter and Tran (2010) that minimizing the loss rank is a suitable criterion for model selection, since it trades off between the quality of fit and the model complexity. LoRP seems to be a promising principle with enormous potential, leading to a rich field. Tran (2009) demonstrated the use of LoRP for selecting the ridge parameter in ridge regression. Tran and Hutter (2010) adapted the idea of LoRP for model selection in a classification context and showed its close connection with excellent model selection techniques based on Rademacher complexities (Koltchinskii, 2001; Bartlett et al., 2002). In this paper, we shall show that LoRP also successfully applies to selecting the shrinkage parameter for the purpose of variable selection.

The main contribution of this paper is to propose a criterion, called the loss rank (LR) criterion, for selecting shrinkage parameters for variable selection purposes. As long as the regularization procedure in use has the consistency property

(2), the shrinkage parameter selected by the LR criterion will produce the true model asymptotically with probability 1. This model selection consistency of the proposed criterion will be proven theoretically in the case where the number of covariates d is fixed and smaller than n. For cases with $d \gg n$, our simulation study suggests that this property still holds. The simulation also shows that our method for variable selection works surprisingly well. Benefiting from fast l_1-regularization algorithms, our method is able to correctly identify significant variables from thousands of candidates in several CPU seconds.

The paper is organized as follows. The main idea of LoRP is briefly reviewed in Section 2. The LR criterion is derived and its model selection consistency is proven in Section 3. Simulation studies and real-data application are presented in Section 4. Section 5 contains the conclusions and outlook. The proofs are relegated to the appendix.

2 The loss rank principle

In this section, we give a brief review of the loss rank principle (LoRP). The reader is referred to Hutter (2007); Hutter and Tran (2010) for the details.

Let us consider a training data set $D = (\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), ..., (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ from a regression model

$$y_i = f(x_i) + \epsilon_i.$$

We first consider discrete $\mathcal{Y}$. Suppose that we use a model M to fit the data D, e.g., M is a linear regression model with d covariates, or M is a k-nearest neighbors regression model. Imagine that in experiment situations we can conduct the experiment many times with fixed design points $\boldsymbol{x}$. We then would get many other (fictitious) output $\boldsymbol{y}'$. Observe that if the model M is complex/flexible (large d, small k), then M fits the training data $(\boldsymbol{x}, \boldsymbol{y})$ well and it also fits $(\boldsymbol{x}, \boldsymbol{y}')$ well (with respect to some loss function). Here, for simplicity, we only consider the squared loss $\mathrm{Loss}_M(\boldsymbol{y}|\boldsymbol{x}) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = \sum_1^n (y_i - \hat{y}_i)^2$ where $\hat{\boldsymbol{y}}$ is the fitted vector under model M. Therefore the loss rank of M defined by

$$\mathrm{Rank}_M(D) := \#\{\boldsymbol{y}' \in \mathcal{Y}^n : \mathrm{Loss}_M(\boldsymbol{y}'|\boldsymbol{x}) \leq \mathrm{Loss}_M(\boldsymbol{y}|\boldsymbol{x})\}$$

will be large for complex M. Conversely, as argued by Hutter and Tran (2010), if M is small/rigid, that both $\mathrm{Loss}_M(\boldsymbol{y}|\boldsymbol{x})$ and $\mathrm{Loss}_M(\boldsymbol{y}'|\boldsymbol{x})$ are large also leads to a large loss rank. Thus, it is natural to choose a model with the smallest loss rank as a good model. By doing this, we trade off between the fit and the model complexity.

In the case of continuous $\mathcal{Y}$, say for instance $\mathcal{Y} = \mathbb{R}$, it is natural to use the concept of volume instead of the counting measure in the definition of loss rank, i.e.,

$$\mathrm{Rank}_M(D) := \mathrm{Vol}\{\boldsymbol{y}' \in \mathbb{R}^n : \mathrm{Loss}_M(\boldsymbol{y}'|\boldsymbol{x}) \leq \mathrm{Loss}_M(\boldsymbol{y}|\boldsymbol{x})\}.$$

Consider the case of linear models where the fitted vector $\hat{\boldsymbol{y}}$ is linear in $\boldsymbol{y}$, i.e., $\hat{\boldsymbol{y}} = M(\boldsymbol{x})\boldsymbol{y}$ where the regression matrix $M = M(\boldsymbol{x})$ depends only on $\boldsymbol{x}$ (using the

same symbol M for both model and regression matrix will not cause any confusion in the following). Then $\text{Loss}_M(\boldsymbol{y}|\boldsymbol{x}) = \|\boldsymbol{y} - M\boldsymbol{y}\|^2 = \boldsymbol{y}^\top A\boldsymbol{y}$ with $A = (I_n - M)^\top(I_n - M)$ and the loss rank is

$$\text{Rank}_M(D) = \text{Vol}\{\boldsymbol{y}' \in \mathbb{R}^n : \boldsymbol{y}'^\top A\boldsymbol{y}' \leq L\} \text{ where } L := \boldsymbol{y}^\top A\boldsymbol{y}.$$

Suppose at the moment that $\det(A) \neq 0$. The set $\{\boldsymbol{y}' \in \mathbb{R}^n : \boldsymbol{y}'^\top A\boldsymbol{y}' \leq L\}$ is an ellipsoid in $\mathbb{R}^n$, so that its volume is

$$\text{Rank}_M(D) = \frac{v_n L^{n/2}}{\sqrt{\det A}}$$

where $v_n = \pi^{n/2}/\Gamma(\frac{n}{2}+1)$ is the volume of the unit sphere in $\mathbb{R}^n$. Because the logarithm is monotone increasing and v_n depends only on n, it is equivalent to consider

$$\text{LR}_M(D) = \tfrac{n}{2}\log(\boldsymbol{y}^\top A\boldsymbol{y}) - \tfrac{1}{2}\log\det A. \tag{3}$$

Principle 1. *Given a class of linear models $\mathcal{M} = \{M\}$, the best model among $\mathcal{M}$ is the one with the smallest loss rank*

$$M^{best} = \text{argmin}_{M \in \mathcal{M}}\{\tfrac{n}{2}\log(\boldsymbol{y}^\top A\boldsymbol{y}) - \tfrac{1}{2}\log\det A\} \tag{4}$$

where $A = (I_n - M)^\top(I_n - M)$, provided that $\det A > 0$.

Now we consider the case where $\det A = 0$ (e.g., projective regression) or A is nearly singular (e.g., ridge regression when the ridge parameter is very close to 0). In such cases, $\text{Rank}_M(D)$ is infinity or extremely large. Following the principle of ridge regression, we add, in order to prevent the loss rank from being infinity or extremely large, a small penalty $\alpha\|\boldsymbol{y}\|^2$ to the loss

$$\text{Loss}_M^\alpha(\boldsymbol{y}|\boldsymbol{x}) := \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2 + \alpha\|\boldsymbol{y}\|^2 = \boldsymbol{y}^\top S_\alpha\boldsymbol{y}, \quad S_\alpha = A + \alpha I_n$$

where $\alpha > 0$ is a small number to be determined later. Now, S_α being not singular yields

$$\text{Rank}_M^\alpha(D) = \text{Vol}\{\boldsymbol{y}' \in \mathbb{R}^n : \boldsymbol{y}'^\top S_\alpha\boldsymbol{y}' \leq L\} = \frac{v_n L^{n/2}}{\sqrt{\det S_\alpha}} \text{ where } L := \boldsymbol{y}^\top S_\alpha\boldsymbol{y}.$$

Taking logarithm and neglecting a constant independent of M, we define the loss rank of model M (dependent on α) as

$$\text{LR}_M^\alpha(D) = \tfrac{n}{2}\log(\boldsymbol{y}^\top S_\alpha\boldsymbol{y}) - \tfrac{1}{2}\log\det S_\alpha. \tag{5}$$

How do we deal with the extra parameter α? We are seeking a model of smallest loss rank, so it is natural to minimize $\text{LR}_M^\alpha(D)$ in α (see Hutter and Tran (2010) for a more detailed interpretation). Therefore, we finally define the loss rank of model M as

$$\text{LR}_M(D) = \inf_{\alpha>0}\text{LR}_M^\alpha(D) = \inf_{\alpha>0}\{\tfrac{n}{2}\log(\boldsymbol{y}^\top S_\alpha\boldsymbol{y}) - \tfrac{1}{2}\log\det S_\alpha\}. \tag{6}$$

Principle 2. *(Hutter and Tran, 2010) Given a class of linear models $\mathcal{M} = \{M\}$, the best model among $\mathcal{M}$ is the one with the smallest loss rank*

$$M^{best} = \operatorname{argmin}_{M \in \mathcal{M}} \mathrm{LR}_M(D) = \operatorname{argmin}_{M \in \mathcal{M}} \inf_{\alpha > 0} \{ \tfrac{n}{2} \log(\boldsymbol{y}^\top S_\alpha \boldsymbol{y}) - \tfrac{1}{2} \log \det S_\alpha \} \quad (7)$$

where $S_\alpha = A + \alpha I_n = (I_n - M)^\top (I_n - M) + \alpha I_n$.

Many attractive properties of LoRP have been pointed out: LoRP reduces to Bayesian model selection in some special cases and has an interpretation in terms of MDL principle (Hutter and Tran, 2010); in the classification context, LoRP has a close connection with (and works in some cases better than) model selection techniques based on Rademacher complexities (Tran and Hutter, 2010). By virtue of LoRP, the loss rank criterion for selecting the Lasso parameter will be derived in the next section.

3 The LR criterion

Let us go back to the variable selection problem in linear regression analysis. We build on the notation of Hutter and Tran (2010). Given a (large) set of d potential covariates $X_1,...,X_d$ and a response variable Y, we consider the problem of choosing the important covariates for explaining Y as a linear function of $X_1,...,X_d$. It is assumed as usual that the covariates are linearly independent and that $E(Y|X_1,...,X_d)$ is a linear combination of $X_1,...,X_d$ with some of the coefficients are zero. We are primarily interested in identifying the non-zero coefficients.

Suppose that the response vector $\boldsymbol{y}$ and the design matrix X have been centered, so that the intercept is omitted from models. Denote by $\mathcal{S}_T = \{j_1^*,...j_{d^*}^*\}$ and $\mathcal{S} = \{j_1,...j_{d_0}\}$ the true model and a candidate model, respectively. Under model $\mathcal{S}$, we can write

$$\boldsymbol{y} = X\boldsymbol{\beta}_\mathcal{S} + \sigma\boldsymbol{\epsilon}$$

where $E(\boldsymbol{\epsilon}) = 0$, $\mathrm{cov}(\boldsymbol{\epsilon}) = I_n$, $\sigma > 0$. We shall consider $\boldsymbol{\beta}_\mathcal{S} = (\beta_{\mathcal{S}1},...,\beta_{\mathcal{S}d})^\top$ as a point in $\mathbb{R}^d$ with $\beta_{\mathcal{S}j} = 0$ if $j \notin \mathcal{S}$. Denote by $\Theta(\mathcal{S}) := \{\theta_\mathcal{S} = (\boldsymbol{\beta}_\mathcal{S}, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+\}$ the parameter space of model $\mathcal{S}$ and by $X_\mathcal{S}$ the $(n \times d_0)$ design matrix obtained from X by removing the jth column for all $j \notin \mathcal{S}$.

3.1 The LR criterion

Let $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_1^\lambda,...,\hat{\beta}_d^\lambda)^\top$ be the regularized estimator of $\boldsymbol{\beta}$ w.r.t. a certain shrinkage parameter λ, i.e., $\hat{\boldsymbol{\beta}}_\lambda$ is the solution of (1). Denote by $\mathcal{S}_\lambda = \{j : \hat{\beta}_j^\lambda \neq 0\}$ the index set corresponding to the non-zero coefficients, by $\mathrm{df}_\lambda = |\mathcal{S}_\lambda|$ the number of non-zero coefficients, and by $X_{\mathcal{S}_\lambda}$ the design matrix corresponding to the selected covariates. We assume at the moment that $\mathrm{df}_\lambda \leq n$ and further assume that matrices $X_{\mathcal{S}_\lambda}$ are full rank. The case where $\mathrm{df}_\lambda > n$ will be dealt with later on.

Fitting model $\mathcal{S}_\lambda$ by least squares, we denote the OLS estimator and the variance estimator by

$$\hat{\boldsymbol{\beta}}_{\mathcal{S}_\lambda} = (X_{\mathcal{S}_\lambda}^\top X_{\mathcal{S}_\lambda})^{-1} X_{\mathcal{S}_\lambda}^\top \boldsymbol{y} \; ; \; \hat{\sigma}_{\mathcal{S}_\lambda}^2 = \tfrac{1}{n} \|\boldsymbol{y} - X_{\mathcal{S}_\lambda} \hat{\boldsymbol{\beta}}_{\mathcal{S}_\lambda}\|^2,$$

respectively. The fitted vector under model $\mathcal{S}_\lambda$

$$\hat{\boldsymbol{y}}_{\mathcal{S}_\lambda} = X_{\mathcal{S}_\lambda} \hat{\boldsymbol{\beta}}_{\mathcal{S}_\lambda} = M_{\mathcal{S}_\lambda} \boldsymbol{y} \text{ with } M_{\mathcal{S}_\lambda} := X_{\mathcal{S}_\lambda} (X_{\mathcal{S}_\lambda}^\top X_{\mathcal{S}_\lambda})^{-1} X_{\mathcal{S}_\lambda}^\top$$

is, conditionally on $\mathcal{S}_\lambda$, linear[1] in $\boldsymbol{y}$. Then from (5), the loss rank of model $\mathcal{S}_\lambda$ with parameter α is

$$\mathrm{LR}_\lambda^\alpha \equiv \mathrm{LR}_{\mathcal{S}_\lambda}^\alpha = \tfrac{n}{2} \log(\boldsymbol{y}^\top S_\alpha^\lambda \boldsymbol{y}) - \tfrac{1}{2} \log \det(S_\alpha^\lambda)$$

where $S_\alpha^\lambda = (I - M_{\mathcal{S}_\lambda})^\top (I - M_{\mathcal{S}_\lambda}) + \alpha I = (1+\alpha) I - M_{\mathcal{S}_\lambda}$. Because projection matrix $M_{\mathcal{S}_\lambda}$ has df_λ eigenvalues 1 and $n - \mathrm{df}_\lambda$ eigenvalues 0, S_α^λ has df_λ eigenvalues α and $n - \mathrm{df}_\lambda$ eigenvalues $1 + \alpha$. Thus, $\det S_\alpha^\lambda = \alpha^{\mathrm{df}_\lambda} (1+\alpha)^{n-\mathrm{df}_\lambda}$. Let $\rho_\lambda := \|\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathcal{S}_\lambda}\|^2 / \|\boldsymbol{y}\|^2$, we have

$$\mathrm{LR}_\lambda^\alpha = \tfrac{n}{2} \log \boldsymbol{y}^\top \boldsymbol{y} + \tfrac{n}{2} \log(\rho_\lambda + \alpha) - \tfrac{\mathrm{df}_\lambda}{2} \log \alpha - \tfrac{n - \mathrm{df}_\lambda}{2} \log(1 + \alpha).$$

Taking derivative w.r.t α, it is easy to see that $\mathrm{LR}_\lambda^\alpha$ is minimized at $\alpha_m = \frac{\rho_\lambda \mathrm{df}_\lambda}{(1-\rho_\lambda)n - \mathrm{df}_\lambda}$ provided that $1 - \rho_\lambda > \mathrm{df}_\lambda / n$. This condition is ensured by Assumption (A3) below. Finally, after some algebra, the loss rank of model $\mathcal{S}_\lambda$ as defined in (6) can be explicitly expressed as

$$\mathrm{LR}_\lambda = \mathrm{LR}_\lambda^{\alpha_m} = \tfrac{n}{2} \log \|\boldsymbol{y}\|^2 - \tfrac{n}{2} \mathrm{KL}(\tfrac{\mathrm{df}_\lambda}{n} \| 1 - \rho_\lambda). \tag{8}$$

where $\mathrm{KL}(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ is the Kullback-Leibler divergence between the Bernoulli distributions with parameters $p, q \in (0,1)$. The optimal shrinkage parameter(s) λ (for variable selection purposes) chosen by the LR criterion will be

$$\hat{\lambda}_{\mathrm{LR}} \in \mathrm{argmin}_{\lambda \geq 0} \mathrm{LR}_\lambda = \mathrm{argmax}_{\lambda \geq 0} \mathrm{KL}(\tfrac{\mathrm{df}_\lambda}{n} \| 1 - \rho_\lambda). \tag{9}$$

Often, LR_λ reaches its minimum in an interval $(\hat{\lambda}_l, \hat{\lambda}_u)$ (see Figure 1). Any λ in this interval produces the same model. This can be explained as follows. When λ increases from 0 to infinity, the number of non-zero coefficients of $\hat{\boldsymbol{\beta}}_\lambda$ will be a non-increasing step function of λ (Efron et al., 2004); in other words, the covariates are in turn removed from the models. As a result, by its definition LR_λ is also a step function. Note that our emphasis is on variable selection rather than on coefficient estimation.

---

[1]Strictly speaking, $\hat{\boldsymbol{y}}_{\mathcal{S}_\lambda}$ is not linear in $\boldsymbol{y}$ because $\mathcal{S}_\lambda$ depends on $\boldsymbol{y}$. However, we can consider preselected subsets $\mathcal{S}_\lambda$ as fixed models. If instead we first derive the LR criterion for a general fixed model $\mathcal{S}$ and then apply to $\mathcal{S}_\lambda$, we get the same results.

3.2 Model selection consistency for fixed d

In order to prove the model selection consistency of the LR criterion, we assume in this section that d is fixed and $d \leq n$. We need the following assumptions

(A1) There exists a deterministic sequence of reference shrinkage parameters λ_n such that $\mathcal{S}_{\lambda_n} \to \mathcal{S}_T$ w.p.1.

(A2) $\boldsymbol{\epsilon}$ is Gaussian $N(0, I_n)$.

(A3) For each candidate λ, ρ_λ is bounded away from 0 and 1, i.e., there are constants c_1, c_2 such that $0 < c_1 \leq \rho_\lambda \leq c_2 < 1$ w.p.1.

Comments. $\rho_\lambda = \|\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathcal{S}_\lambda}\|^2 / \|\boldsymbol{y}\|^2$ is a measure of fit. In extreme cases where the resulting model $\mathcal{S}_\lambda$ is too big or too small, ρ_λ will be close to 0 and 1, respectively. Therefore, it is reasonable to consider only λ in which ρ_λ is bounded away from 0 and 1. Note that for every $\mathcal{S}_\lambda$ we have that

$$\rho_\lambda = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathcal{S}_\lambda}\|^2}{\|\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathcal{S}_\lambda}\|^2 + \|\hat{\boldsymbol{y}}_{\mathcal{S}_\lambda}\|^2} = \frac{\hat{\sigma}^2_{\mathcal{S}_\lambda}}{\hat{\sigma}^2_{\mathcal{S}_\lambda} + \|\hat{\boldsymbol{y}}_{\mathcal{S}_\lambda}\|^2/n}.$$

For λ such that $\mathcal{S}_\lambda$ is the true model $\mathcal{S}_T$, (A3) follows from a mild sufficient condition

$$0 < \liminf_{n \to \infty} (\tfrac{1}{n}\|\hat{\boldsymbol{y}}_{\mathcal{S}_T}\|^2) \leq \limsup_{n \to \infty} (\tfrac{1}{n}\|\hat{\boldsymbol{y}}_{\mathcal{S}_T}\|^2) < \infty \text{ and } \hat{\sigma}^2_{\mathcal{S}_T} \to \sigma^2 > 0 \text{ w.p.1}$$

where $\hat{\boldsymbol{y}}_{\mathcal{S}_T}$ is the fitted vector under the true model. Moreover, if the intercept is included in the models, we have that $n(\bar{\boldsymbol{y}})^2 \leq \|\hat{\boldsymbol{y}}_{\mathcal{S}_\lambda}\|^2 \leq \|\boldsymbol{y}\|^2$. (A3) then follows from a very mild condition

$$0 < \liminf_{n \to \infty} (\bar{\boldsymbol{y}})^2 \leq \limsup_{n \to \infty} (\tfrac{1}{n}\|\boldsymbol{y}\|^2) < \infty \text{ and } \hat{\sigma}^2_{\mathcal{S}} \to \text{constant} > 0 \ \forall \mathcal{S} \text{ w.p.1.}$$

Assumption (A1) is satisfied by some regularization procedures, for example, Lasso (Zhao and Yu, 2006) and SCAD (Fan and Li, 2001). Normality assumption (A2) is not a necessary condition for consistency. This assumption can be relaxed, but then a more complicated proof technique is needed.

We have the following lemma.

Lemma 3. *The loss rank of model $\mathcal{S}_\lambda$ can be rewritten as*

$$\text{LR}_\lambda = \tfrac{n}{2} \log(n\hat{\sigma}^2_{\mathcal{S}_\lambda}) + \tfrac{n}{2} H(\tfrac{\text{df}_\lambda}{n}) + \tfrac{\text{df}_\lambda}{2} \log \tfrac{1-\rho_\lambda}{\rho_\lambda} \tag{10}$$

where $H(p) := -p\log p - (1-p)\log(1-p)$ is the entropy of p. Under Assumption (A3), the loss rank LR_λ has the form

$$\text{LR}_\lambda = \tfrac{n}{2} \log \hat{\sigma}^2_{\mathcal{S}_\lambda} + \tfrac{\text{df}_\lambda}{2} \log n + \tfrac{n}{2} \log n + O_{\text{P}}(1), \tag{11}$$

where $O_{\text{P}}(1)$ denotes a bounded random variable w.p.1.

Proof. With $\rho_\lambda = n\hat{\sigma}_{\mathcal{S}_\lambda}^2 / \|\boldsymbol{y}\|^2$, rearranging terms in (8) we get (10). The fact that $\frac{H(p)}{p} + \log p \to 1$ as $p \to 0$ implies that (note that $\mathrm{df}_\lambda \leq d$ and d is fixed)

$$\tfrac{n}{2} H(\tfrac{\mathrm{df}_\lambda}{n}) = \tfrac{\mathrm{df}_\lambda}{2} \log n + \tfrac{\mathrm{df}_\lambda}{2}(1 - \log \mathrm{df}_\lambda) + o(1).$$

Under Assumption (A3), the last term of (10) is bounded. This completes the proof. ∎

The above lemma is used to prove model selection consistency of the LR criterion.

Theorem 4 (Model selection consistency of the LR criterion). *Assume that d is fixed. Under Assumptions (A1)-(A3), the shrinkage parameter selected by the LR criterion will produce the true model w.p.1 when n is large enough, i.e.,*

$$\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}_T) \to 1$$

where $\hat{\lambda}_{\mathrm{LR}}$ is determined in (9).

The idea of the proof is to bound the probabilities of picking under- and overfitted models. A model $\mathcal{S}$ is said to be underfitted if $\mathcal{S}$ misses at least one true covariate (i.e., $\mathcal{S} \not\supseteq \mathcal{S}_T$), overfitted if $\mathcal{S}$ contains all true covariates and at least one untrue (i.e., $\mathcal{S} \supsetneq \mathcal{S}_T$). There is a finite number of such $\mathcal{S}$, so it is sufficient to prove that $\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}) \to 0$ for each of them. The detailed proof is relegated to the appendix.

We can of course use other model selection criteria rather than LoRP for choosing the best subset among the preselected set produced by the regularization procedure. The most widely-used selection criteria in statistics are probably AIC (Akaike, 1973) and BIC (Schwarz, 1978). AIC is asymptotically optimal in terms of loss efficiency but likely to select overfitted models, while BIC is asymptotically optimal in terms of model selection consistency; see Shao (1997); Yang (2005). Therefore one may use BIC as another stopping rule besides LoRP. The shrinkage parameter chosen by BIC will be

$$\hat{\lambda}_{\mathrm{BIC}} \in \mathrm{argmin}_{\lambda \geq 0} \mathrm{BIC}_\lambda \quad \text{where} \quad \mathrm{BIC}_\lambda := \tfrac{n}{2} \log \hat{\sigma}_{\mathcal{S}_\lambda}^2 + \tfrac{\mathrm{df}_\lambda}{2} \log n. \tag{12}$$

We see from Lemma 3 that, up to a constant, the LR criterion is asymptotically equivalent to BIC. It follows from the proof of Theorem 4 that using BIC also leads to the same model selection consistency, i.e., $\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{BIC}}} = \mathcal{S}_T) \to 1$ as $n \to \infty$. However, finite-sample simulation studies in the next section show that the LR criterion works better than BIC, especially when $d \gg n$.

3.3   Large d small n

High-dimensional variable selection problems in which $d \gg n$ is currently of great interest to scientists. In order for such a problem to be solvable, an essential assumption needed is that it is d^*−sparse (Candes and Tao, 2007), i.e., the number

of true covariates d^* must be smaller than n. Under this solvability assumption, it is clear that we can safely ignore irrelevant cases in which the number of covariates df_λ under consideration is larger than n. Then the LR criterion (8) is still valid. In practice, therefore, we propose to ignore those λ under which $\mathrm{df}_\lambda > n$ and apply the LR criterion as usual. A theoretically rigorous treatment is beyond the scope of the present paper, which we intend to do in a future paper. However, a systematic simulation study in the next section suggests that the LR criterion still works surprisingly well and enjoys model selection consistency.

4 Numerical examples

In this section, we present simulation studies for the LR criterion, compare the LR criterion to other methods, and also apply it to a real data set. The regularization procedure we use is Lasso. The Lasso solution paths are computed by the LARS algorithm of Efron et al. (2004). A widely-used method for choosing the Lasso parameter is GCV (Craven and Wahba, 1979; Tibshirani, 1996)

$$\mathrm{GCV}_\lambda = \frac{1}{n} \frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\lambda\|^2}{(1 - \frac{1}{n}\mathrm{DF}_\lambda)^2}$$

where $\mathrm{DF}_\lambda := \mathrm{tr}[X(X^\top X + \lambda W^-)^{-1} X^\top \boldsymbol{y}]$, $W = \mathrm{diag}(|\hat{\beta}_j^\lambda|)$ and W^- is a generalized inverse of W. Another one is the BIC-type criterion of Wang et al. (2007) (although its variable selection consistency requires the oracle property, a property not enjoyed by Lasso)

$$\widetilde{\mathrm{BIC}}_\lambda = \log \frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\lambda\|^2}{n} + \mathrm{DF}_\lambda \frac{\log n}{n}.$$

Note that $\hat{\boldsymbol{\beta}}_\lambda \neq \hat{\boldsymbol{\beta}}_{\mathcal{S}_\lambda}$. The former is the Lasso estimator whereas the latter is the OLS estimator resulting from fitting model $\mathcal{S}_\lambda$ by least squares. Our proposed criteria (8) and (12) are constructed based on $\hat{\boldsymbol{\beta}}_{\mathcal{S}_\lambda}$, not $\hat{\boldsymbol{\beta}}_\lambda$. This is the essential difference between our approach and the others.

Example 1: small d. We consider the following example which is taken from Tibshirani (1996):

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + \sigma\epsilon$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$, x_i are marginally $N(0,1)$ with the correlation between x_i and x_j equal to $0.5^{|i-j|}$, $\epsilon \sim N(0,1)$. We compare the performance of LR and BIC criterion to that of GCV and $\widetilde{\mathrm{BIC}}$. The performance is measured by the frequency of underfitting, overfitting and correct fitting and average number of zero coefficients over 100 replications.

Table 1 summarizes the simulation results for various factors n and σ. Although $\widetilde{\mathrm{BIC}}$ works slightly better than GCV, it still produces overfitted models most of the time. BIC does a good job and LR outperforms the others.

Table 1: The small-d case

σ	n	Method	Under-fitted(%)	Correctly fitted(%)	Overfitted(%)	Ave. No. of zeros
1	100	GCV	0	0	100	1.57
		$\widetilde{\text{BIC}}$	0	3	97	2.32
		BIC	0	89	11	4.88
		LR	0	97	3	4.97
	200	GCV	0	0	100	1.64
		$\widetilde{\text{BIC}}$	0	0	100	1.81
		BIC	0	94	6	4.93
		LR	0	100	0	5
3	100	GCV	0	0	100	1.34
		$\widetilde{\text{BIC}}$	0	0	100	1.53
		BIC	1	70	29	4.22
		LR	1	77	22	4.37
	200	GCV	0	0	100	1.69
		$\widetilde{\text{BIC}}$	0	0	100	2.09
		BIC	0	91	9	4.89
		LR	0	91	9	4.90

Example 2: large d. We consider cases of large d in this example with $d = 300$ and $n = 100, 200, 500$. We set up a *sparse recovery problem* in which most of coefficients are zero except $\beta_{30} = \beta_{60} = ... = \beta_{300} = 10$. The design matrix is simulated as in Example 1. Table 2 summarizes the simulation results for various factors $n = 100, 200, 500$ and $\sigma = 1, 3$. The LR criterion works surprisingly well in comparison with BIC and the others.

Let us take a closer look at the simulation results in Tables 1-2. Although the LR and BIC criteria are *asymptotically* equivalent to each other, the finite-sample simulation study shows that the LR criterion works better than BIC. A similar situation was also observed in Hutter and Tran (2010) for subset selection. This is probably because, contrarily to the BIC criterion, the penalty term of the LR criterion is data-adaptive. Some results in the model selection literature show that selection criteria with data-adaptive penalties are more encouraging than those with deterministic penalties; see Yang (2005) and references therein. We see that BIC seems to break down for the cases $d > n$ as it always produces overfitted models, but starts working well when $n > d$. The $O_P(1)$ term in (11) plays an important role here: it serves as a "corrector" to BIC. Note that BIC is just an approximation to the logarithm of posterior model probability (Schwarz, 1978), the approximation might be inaccurate if n is not large enough relative to d.

Example 3: Prostate cancer data. We consider a real data set in this example.

Table 2: The large-d case

σ	n	Method	Under-fitted(%)	Correctly fitted(%)	Overfitted(%)	Ave. No. of zeros
1	100	GCV	0	0	100	90.20
		$\widetilde{\text{BIC}}$	0	0	100	95.8
		BIC	0	0	100	202.01
		LR	0	30	70	288.24
	200	GCV	0	0	100	87.51
		$\widetilde{\text{BIC}}$	0	0	100	89.45
		BIC	0	0	100	102.02
		LR	0	86	14	289.83
	500	GCV	0	0	100	97.51
		$\widetilde{\text{BIC}}$	0	0	100	104.45
		BIC	0	40	60	287.30
		LR	0	100	0	290
3	100	GCV	0	0	100	78.35
		$\widetilde{\text{BIC}}$	0	0	100	87.40
		BIC	0	0	100	202.04
		LR	0	18	82	287.51
	200	GCV	0	0	100	92.02
		$\widetilde{\text{BIC}}$	0	0	100	96.51
		BIC	0	0	100	102.01
		LR	0	58	42	289.29
	500	GCV	0	0	100	93.31
		$\widetilde{\text{BIC}}$	0	0	100	96.52
		BIC	0	35	65	288.35
		LR	0	80	20	289.75

Stamey et al. (1989) studied the correlation between the level of prostate antigen (*lpsa*) and a number of clinical measures in men: log cancer volume (*lcavol*), log prostate weight (*lweight*), *age*, log of the amount of benign prostatic hyperplasia (*lbph*), seminal vesicle invasion (*svi*), log of capsular penetration (*lcp*), Gleason score (*gleason*), and percentage of Gleason scores 4 or 5 (*pgg45*). Following Tibshirani (1996), we assume a linear regression model between the response *lpsa* and the 8 covariates. We want to select a parsimonious model for the sake of scientific insight into the response-covariate relationship.

The data set of size 97 is standardized so that the intercept β_0 is excluded. Figure 1 presents the curves GCV_λ, $\widetilde{\text{BIC}}_\lambda$, LR_λ (1000 values of λ ranging from 0.01 to 10 in increments of .01 were used to search for the optimal λ). The λ selected by GCV, $\widetilde{\text{BIC}}$ are .5 and 1.1, and the corresponding models are $\{1, 2, 3, 4, 5, 7, 8\}$,
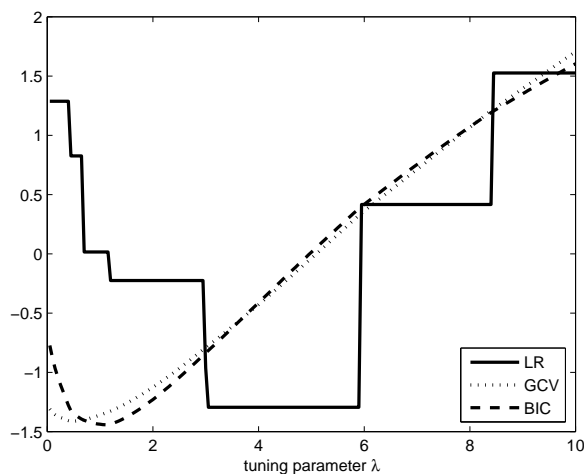
Figure 1: Prostate cancer data: LR_λ, $\widetilde{\text{BIC}}_\lambda$ and GCV_λ.

$\{1, 2, 3, 4, 5, 8\}$, respectively. The LR criterion is minimized in the interval $(3.1, 5.9)$. Any value in this interval produces the same model $\mathcal{S}_{\text{LR}} = \{1, 2, 5\}$. The BIC of these models are -19.20, -21.38, -25.19, respectively. That means the BIC also supports the choice of the LR criterion. (Note however that this does not mean that the BIC is an optimal criterion).

5   Conclusions and outlook

Regularization procedures are efficient methods for variable selection, subject to a proper choice of shrinkage parameter. By virtue of LoRP, a general-purpose principle for model selection, the LR criterion for variable selection in linear regression analysis was proposed. Variable selection consistency of the suggested criterion was pointed out theoretically and experimentally. Both theoretical and experimental results show that the proposed criterion is a very encouraging procedure for variable selection problem, especially in high-dimensional settings. Regularization procedures have now been extended to genaralized linear models and beyond, we intend to extend our approach to such frameworks in future work.

––––––––––

M. N. Tran, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546.
E-mail: ngoctm@nus.edu.sg

Appendix

Proof of Theorem 4. The main idea of the proof is taken from Chambaz (2006). Let us denote by $z_i = (x_{i1},...,x_{id},y_i)$ the i-th observation and by $\gamma(.,m,\sigma^2)$ the density of the Gaussian distribution with mean m and variance σ^2. Under model $\mathcal{S}$, the density of z_i is $p_{\theta_\mathcal{S}}(z_i) = \gamma(y_i, \sum_{j\in\mathcal{S}}\beta_j x_{ij}, \sigma^2)$. The log-likelihood is

$$l_n(\theta_\mathcal{S}) = \sum_{i=1}^n \log p_{\theta_\mathcal{S}}(z_i) = -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log\sigma^2 - \tfrac{1}{2\sigma^2}\sum_{i=1}^n\Big(y_i - \sum_{j\in\mathcal{S}}\beta_j x_{ij}\Big)^2.$$

It is easy to see that

$$\sup_{\theta\in\Theta(\mathcal{S})} l_n(\theta) = -\tfrac{n}{2}\log\hat{\sigma}_\mathcal{S}^2 - \tfrac{n}{2}(1 + \log(2\pi)).$$

By (11), the loss rank of model $\mathcal{S}_\lambda$ now can be written as

$$\mathrm{LR}_\lambda = -\sup_{\theta\in\Theta(\mathcal{S}_\lambda)} l_n(\theta) + \tfrac{\mathrm{df}_\lambda}{2}\log n + C(n) + O_\mathrm{P}(1)$$

where the constant term $C(n) = \tfrac{n}{2}\log n - \tfrac{n}{2}(1+\log(2\pi))$ is independent of $\mathcal{S}_\lambda$.

No underestimation. It is sufficient to prove that $\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}) \to 0$ for each $\mathcal{S} \not\supseteq \mathcal{S}_T$, as there is only a finite number of such $\mathcal{S}$.

$$
\begin{aligned}
\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}) &= \mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}, \mathrm{LR}_{\hat{\lambda}_{\mathrm{LR}}} \leq \mathrm{LR}_{\lambda_n}) \\
&= \mathrm{P}\Big(\tfrac{1}{n}\sup_{\theta\in\Theta(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}})} l_n(\theta) - \tfrac{1}{n}\sup_{\theta\in\Theta(\mathcal{S}_{\lambda_n})} l_n(\theta) \geq \tfrac{\log n}{2n}(\mathrm{df}_{\hat{\lambda}_{\mathrm{LR}}} - \mathrm{df}_{\lambda_n}) + o_\mathrm{P}(1), \; \mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}\Big) \\
&\leq \mathrm{P}\Big(\tfrac{1}{n}\sup_{\theta\in\Theta(\mathcal{S})} l_n(\theta) - \tfrac{1}{n}\sup_{\theta\in\Theta(\mathcal{S}_{\lambda_n})} l_n(\theta) \geq \tfrac{\log n}{2n}(|\mathcal{S}| - \mathrm{df}_{\lambda_n}) + o_\mathrm{P}(1)\Big) \\
&\leq \mathrm{P}\Big(\tfrac{1}{n}\sup_{\theta\in\Theta(\mathcal{S})} l_n(\theta) - \tfrac{1}{n}\sup_{\theta\in\Theta(\mathcal{S}_T)} l_n(\theta) \geq \tfrac{\log n}{2n}(|\mathcal{S}| - d^*) + o_\mathrm{P}(1)\Big) + P(\mathcal{S}_{\lambda_n} \neq \mathcal{S}_T) \\
&\leq \mathrm{P}\Big(\tfrac{1}{n}\sup_{\theta\in\Theta(\mathcal{S})} l_n(\theta) - \tfrac{1}{n}l_n(\theta^*) \geq \tfrac{\log n}{2n}(|\mathcal{S}| - d^*) + o_\mathrm{P}(1)\Big) + P(\mathcal{S}_{\lambda_n} \neq \mathcal{S}_T) \qquad (13)
\end{aligned}
$$

where $\theta^* \in \mathcal{S}_T$ denotes the true parameter. By the law of large numbers for the supremum of the likelihood ratios (see, e.g., Lemma B1 of Chambaz (2006))

$$\tfrac{1}{n}\sup_{\theta\in\Theta(\mathcal{S})} l_n(\theta) - \tfrac{1}{n}l_n(\theta^*) \to -\inf_{\theta\in\Theta(\mathcal{S})} \mathrm{KL}(p_{\theta^*}\|p_\theta) \quad \text{w.p.1.}$$

Because $\mathcal{S} \not\supseteq \mathcal{S}_T$, $\inf_{\theta\in\Theta(\mathcal{S})}\mathrm{KL}(p_{\theta^*}\|p_\theta) > 0$. This, together with the fact that $\tfrac{\log n}{2n}(|\mathcal{S}| - d^*) \to 0$ and Assumption (A1), shows that the left-hand side term of (13) goes to 0 as $n \to \infty$.

No overestimation. Fix an overfitted model $\mathcal{S} \supsetneq \mathcal{S}_T$, let us denote by

$$H(\theta) := \mathrm{KL}(p_{\theta^*}\|p_\theta) = E[\tfrac{1}{n}(l_n(\theta^*) - l_n(\theta))] \geq 0 \; \forall \theta \in \Theta(\mathcal{S})$$

14

$(H(\theta)$ is not necessarily positive) and $h_n(\theta):=\frac{l_n(\theta)-l_n(\theta^*)}{H(\theta)^{1/2}}$ with convention $\frac{0}{0}=0$. For every $\theta\in\Theta(\mathcal{S})$

$$
\begin{aligned}
l_n(\theta)-l_n(\theta^*)+nH(\theta) &= l_n(\theta)-l_n(\theta^*)-E[l_n(\theta)-l_n(\theta^*)] \\
&= H(\theta)^{1/2}(h_n(\theta)-Eh_n(\theta)) \\
&\leq H(\theta)^{1/2}\sup_{\nu\in\Theta(\mathcal{S})}(h_n(\nu)-Eh_n(\nu)).
\end{aligned}
\tag{14}
$$

By $\Theta(\mathcal{S}_T)\subset\Theta(\mathcal{S})$ and the property of supremum, for every $\epsilon>0$ there exists $\theta_0\in\Theta(\mathcal{S})$ such that

$$
\sup_{\theta\in\Theta(\mathcal{S})}(l_n(\theta)-l_n(\theta^*))\leq l_n(\theta_0)-l_n(\theta^*)+\epsilon
\tag{15}
$$

and also

$$
l_n(\theta_0)-l_n(\theta^*)\geq 0.
\tag{16}
$$

From (15) and (14)

$$
\sup_{\theta\in\Theta(\mathcal{S})}(l_n(\theta)-l_n(\theta^*))\leq H(\theta_0)^{1/2}\sup_{\theta\in\Theta(\mathcal{S})}(h_n(\theta)-Eh_n(\theta))+\epsilon.
\tag{17}
$$

From (16) and (14)

$$
nH(\theta_0)\leq l_n(\theta_0)-l_n(\theta^*)+nH(\theta_0)\leq H(\theta_0)^{1/2}\sup_{\theta\in\Theta(\mathcal{S})}(h_n(\theta)-Eh_n(\theta))
$$

or

$$
nH(\theta_0)^{1/2}\leq\sup_{\theta\in\Theta(\mathcal{S})}(h_n(\theta)-Eh_n(\theta)).
\tag{18}
$$

Now, since $\epsilon>0$ was chosen arbitrarily, (17) and (18) yield

$$
\sup_{\Theta(\mathcal{S})}l_n(\theta)-\sup_{\Theta(\mathcal{S}_T)}l_n(\theta)\leq\sup_{\Theta(\mathcal{S})}\{l_n(\theta)-l_n(\theta^*)\}\leq\frac{1}{n}\left(\sup_{\theta\in\Theta(\mathcal{S})}(h_n(\theta)-Eh_n(\theta))\right)^2.
\tag{19}
$$

We need the following bounded law of the iterated logarithm which is a consequence of Theorem 4.1, Dudley and W.Philipp (1983) or Lemma B2, Chambaz (2006).

Lemma 5. *There is a finite constant C so that*

$$
\limsup_n\frac{\sup_{\theta\in\Theta(\mathcal{S})}|h_n(\theta)-Eh_n(\theta)|}{\sqrt{n\log\log n}}\leq C\quad w.p.1.
$$

Now for every overfitted model $\mathcal{S}\supsetneq\mathcal{S}_T$, it is sufficient to prove that $P(\mathcal{S}_{\hat\lambda_{\mathrm{LR}}}=$

15

$\mathcal{S}) \to 0$. In fact,

$$P(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}) = P(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}, \ \mathrm{LR}_{\hat{\lambda}_{\mathrm{LR}}} \le \mathrm{LR}_{\lambda_n})$$

$$\le \ P\left(\sup_{\Theta(\mathcal{S})} l_n(\theta) - \sup_{\Theta(\mathcal{S}_{\lambda_n})} l_n(\theta) \ge \tfrac{\log n}{2}(|\mathcal{S}| - \mathrm{df}_{\lambda_n}) + O_{\mathrm{P}}(1) \right)$$

$$\le \ P\left(\sup_{\Theta(\mathcal{S})} l_n(\theta) - \sup_{\Theta(\mathcal{S}_T)} l_n(\theta) \ge \tfrac{\log n}{2}(|\mathcal{S}| - d^*) + O_{\mathrm{P}}(1) \right) + P(\mathcal{S}_{\lambda_n} \ne \mathcal{S}_T)$$

$$= \ P\left(\left[\frac{\log \log n}{\frac{d^*}{2} \log n}\right] \left[\frac{\sup_{\Theta(\mathcal{S})} l_n(\theta) - \sup_{\Theta(\mathcal{S}_T)} l_n(\theta)}{\log \log n}\right] \ge \tfrac{|\mathcal{S}|}{d^*} - 1 + o_{\mathrm{P}}(1) \right) + P(\mathcal{S}_{\lambda_n} \ne \mathcal{S}_T)$$

$$\le \ P\left(\left[\frac{\log \log n}{\frac{d^*}{2} \log n}\right] \left[\frac{\sup_{\Theta(\mathcal{S})} |h_n(\theta) - E h_n(\theta)|}{\sqrt{n \log \log n}}\right]^2 \ge \tfrac{|\mathcal{S}|}{d^*} - 1 + o_{\mathrm{P}}(1) \right) + P(\mathcal{S}_{\lambda_n} \ne \mathcal{S}_T) \quad (20)$$

where the last inequality follows from (19). Observe that $|\mathcal{S}| > d^*$ as $\mathcal{S} \supsetneq \mathcal{S}_T$. This, together with Lemma 5 and the fact that $\log \log n / (\frac{d^*}{2} \log n) \to 0$, implies that the first probability of (20) goes to zero. The second probability of (20) also goes to zero because of Assumption (A1). This completes the proof. ∎

References

Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory*, pages 267–281, Budapest, Hungary, Akademiai Kaidó.

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

Burnham K. P. and Anderson D. (2002). *Model selection and multimodel inference: a practical information-theoretic approach.* New York, Springer.

Candes E. and Tao T. (2007). The dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, 35, 2313–2351.

Chambaz A. (2006). Testing the order of a model. *The Annals of Statistics*, 34, 1166–1203.

Craven P. and Wahba G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the methods of generalized cross-validation. *Numerische Mathematik*, 31, 377–403.

Dudley R. M. and Philipp W. (1983). Invariance principles for sums of banach space valued random elements and empirical processes. *Z. Wahrsch. Verw. Gebiete*, 62, 509–552.

Efron B., Hastie T., Johnstone I. and Tibshirani R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407–499.

Fan J. and Li R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Friedman J. H. (2008). Fast sparse regression and classification. URL http://www-stat.stanford.edu/ jhf/ftp/GPSpaper.pdf.

Hutter M. (2007). The loss rank principle for model selection. In *Proc. 20th Annual Conf. on Learning Theory (COLT'07)*, volume 4539 of *LNAI*, 589–603, San Diego, Springer, Berlin.

Hutter M. and Tran M. N. (2010). Model selection with the loss rank principle. *Computational Statistics and Data Analysis*, 54, 1288–1306.

Koltchinskii V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transaction on Information Theory*, 47, 1902–1914.

Leng C., Lin Y. and Wahba G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16, 1273–1284.

Meinshausen N. and Buhlmann P. (2006). Consistent neighbourhood selection for high-dimensional graphs the lasso. *The Annals of Statistics*, 34, 1436–1462.

Miller A. (1990). *Subset Selection in Regression*. Chapman & Hall/CRC.

Poetscher B. M. and Leeb H. (2009). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis*, 100, 2065–2082.

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

Shao J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7, 221–264.

Stamey T., Kabalin J., McNeal J., Johnstone I., Freiha F., Redwine E. and Yang N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii. radical prostatectomy treated patients. *Journal of Urology*, 16, 1076–1083.

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

Tran M. N. (2009). Penalized maximum likelihood principle for choosing ridge parameter. *Communication in Statistics: Simulation and Computation*, 38, 1610–1624.

Tran M. N. and Hutter H. (2010). Model selection by loss rank for discrete data. *Submitted*.

Wang H., Li R. and Tsai C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553–568.

Yang Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation *Biometrika*, 92, 937–950.

Zhao P. and Yu B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7, 2541–2563.

Zou H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

Zou H., Hastie T. and Tibshirani R. (2007). On the degrees of freedom of the lasso. *The Annals of Statitics*, 35, 2173–2192.