Criteria for variable selection with dependence

Aurélie Boisbunon   **Stéphane Canu**   **Dominique Fourdrinier**

1   Context

Most theoretical tools for model selection rely on one main assumption: the independence of noise components. However, in many real regression examples this assumption is too strong and does not fit well the reality (see for instance the discussion on this issue in [1] and references therein). A way to take dependence into account is to model noise as a multivariate spherically symmetric random variable. This general framework includes the well-known multivariate Student and Kotz distributions. Our work consists in integrating this idea into model selection problems.

The problem of model selection implies three steps, as discussed in [2]: (a) the definition of a way to explore models, (b) the estimation of parameters for each model, and (c) the evaluation of the models. If we consider the example of Lasso [3], the L_1-penalization solves both step (a), with the regularization path algorithm [4], and step (b) simultaneously, while step (c) is usually performed with Mallows' C_p [5] or cross-validation [6].

We propose a new procedure for step (c) based on loss estimation. Loss estimation is a data-driven approach closely related to Stein's Unbiased Risk Estimation (SURE) [7], and has been extended to spherically symmetric distributions (see for instance [8]). Our estimator of loss is of the following form:

$$\delta_\rho(y) = \delta_0(y) - \rho(y), \tag{1}$$

where $y \in \mathbb{R}^n$ corresponds to the data, $\delta_0(y)$ is an unbiased estimator of loss (close to Mallows C_p and AIC [9] in the Gaussian case) and $\rho(y)$ is a correction function. We will give in the sequel some details on how to build ρ as well as an explicit formula in a more specific context.

Note that our procedure should not be considered alone. Indeed, as efficient as it can be, it may not give satisfactory results if step (a) and (b) are not well handled. For instance, methods such as Lasso give an interesting and low-computational solution for step (a) exploring models through a regularization path, but not all of them propose good estimates of the parameters (see [10] for instance). We suggest to use Firm Shrinkage [10], which leads to the same model exploration as Lasso for step (a) and whose estimates for step (b) are much less biased than those induced by the L_1-penalty. The choice of a regularization path for step (a) is driven by the increasing size of datasets, but we will discuss the limitations induced by searching only on a small number of models.

2   Loss estimation as a variable selection criterion

We focus our study on the linear regression model

$$Y = X\beta + \varepsilon, \tag{2}$$

where Y is a random vector in $\mathbb{R}^n$, X is a fixed design matrix containing p explanatory variables, $\beta \in \mathbb{R}^p$ is the unknown regression coefficient, and ε is the error vector in $\mathbb{R}^n$ with mean zero. We assume that the distribution of ε is spherically symmetric. In the case where it admits a density, this one is of the form $t \mapsto f(\| t \|^2)$ for a given function f mapping from $\mathbb{R}_+$ into $\mathbb{R}_+$, with $\|.\|$ the Euclidean norm. What characterizes our work is that we do not need to specify the form of f, thus our results are more robust in terms of distribution. Another important feature of this family of distributions is the dependence between the components of ε, with the exception of the Gaussian distribution. However, ε_i and ε_j are not correlated, but correlation could be handled by the more

general elliptical case. The family of spherically symmetric distributions covers, among others, the well known Student and Kotz distributions.

Model selection includes the specific problem of variable selection. In order to address this problem, we assume that only a small subset of variables in X are relevant to explain Y. The literature on this issue, corresponding to step (a), covers three main ways to perform the selection: the first one is to consider all possible subsets, leading to an exhaustive exploration, like in Subset Selection; the second one is to assess the significativity of each variable through statistical tests in order to decide whether to keep it or drop it, as is done in Forward and Backward Selection; the last one, which has been widely studied in the last 20 years, is the construction of a regularization path adding one variable at a time with respect to some criterion. We can cite from this latter approach methods such as the LARS algorithm for Lasso [3, 4] or Firm Shrinkage [10]. These methods generally depend on some hyperparameter tuning the level of sparsity, reducing the problem of variable selection to the optimization of this hyperparameter. The huge number of comparisons of the first two ways discussed here has led researchers to focus their interest on regularization path methods, especially in a context of large datasets.

In the sequel, we consider the Firm Shrinkage estimator, a method developed by Bruce and Gao in 1997 that reduces the large bias of Lasso. This way, Firm Shrinkage achieves a good estimation of parameter β and shares a low computational cost with Lasso. It is an arguable choice since there exist other Lasso-type methods with low bias, but it turns out to perform quite well on our simulations. Nevertheless, a comparison of such methods in terms of selection and estimation could lead to even better results. Note that Firm Shrinkage requires X to be orthogonal. Its estimator of β is

$$\hat{\beta}_i^{FS} = \begin{cases} 0 & |\hat{\beta}_i^{LS}| \leq \lambda \\ a(\hat{\beta}_i^{LS} - \lambda sign(\hat{\beta}_i^{LS}))/(a-1) & \lambda < |\hat{\beta}_i^{LS}| \leq a\lambda \\ \hat{\beta}_i^{LS} & |\hat{\beta}_i^{LS}| > a\lambda \end{cases}, \tag{3}$$

where $\hat{\beta}^{LS}$ is the least-squares solution, $\lambda > 0$ is the hyperparameter tuning the sparsity, and $a > 1$ a hyperparameter tuning the bias. We set a to 2 in order to avoid the estimation of a second parameter and because it yields good results (see [11]).

We now define more precisely our procedure of selection. Loss estimation is a decision theory tool originally designed for the problem of estimating the location parameter μ of a distribution $\mathbb{P}_\mu$, like the multivariate Gaussian (see for instance [7, 12]). It aims at evaluating the quality of some estimator of μ by estimating its loss function $L(\hat{\mu}, \mu)$ through the observations only, the real loss $L(\hat{\mu}, \mu)$ being inaccessible because of its dependence on the unknown parameter μ. The idea of loss estimation is closely related to Stein's Unbiased Risk Estimation (SURE), but the main difference lies in the fact that it does not only consider unbiased estimates. Unbiasedness is indeed not always an interesting property, and potential improvement can be achieved in terms of quadratic risk by adding a bias. We formalize this idea as follows. Let δ_0 be an unbiased estimator of loss $L(\hat{\mu}, \mu)$, that is to say δ_0 verifies the property $\mathbb{E}_\mu[\delta_0] = \mathbb{E}_\mu[L(\hat{\mu}, \mu)]$, where $\mathbb{E}_\mu$ is the expectation with respect to spherically symmetric distributions with location parameter μ. We look for corrections $\rho(y)$ in (1) such that

$$\mathcal{R}_\mu(\delta_\rho) = \mathbb{E}_\mu[(\delta_\rho - L(\hat{\mu}, \mu))^2] = \mathbb{E}_\mu[(\delta_0 - \rho(y) - L(\hat{\mu}, \mu))^2] \leq \mathbb{E}_\mu[(\delta_0 - L(\hat{\mu}, \mu))^2] = \mathcal{R}_\mu(\delta_0), \tag{4}$$

and with strict inequality at least for one value of μ. A common loss for L is the quadratic loss $L(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|^2$, since it allows easy computations. It is also interesting to use this loss function in combination with nearly unbiased estimators of μ, like Firm Shrinkage for $\mu = X\beta$, as it has a high probability of reaching its minimum for the true model, as soon as this latter one belongs to the set of models explored.

From now on we will restrict the study to the estimation of the quadratic risk of the Firm Shrinkage estimator. The corrective estimator of loss is

$$\delta_\rho(y) = \delta_0(y) - \rho(y) \tag{5}$$

where

$$\delta_0 = \frac{2df - n}{n - p} \|y - X\hat{\beta}^{LS}\|^2 + \|y - X\hat{\beta}^{FS}\|^2 \tag{6}$$

and ρ is a correction function. Note that the number of degrees of freedom for Firm Shrinkage is $df = k + \#\{i \setminus \lambda < |\beta_i^{LS}| \leq a\lambda\}/(a-1)$ with k the number of selected variables. We consider

2

correction functions of the form

$$\rho(y) = \|y - X\hat{\beta}^{LS}\|^4 \gamma(y) \tag{7}$$

where $\gamma(y)$ is a twice weakly differentiable function. In particular, we study the function

$$\gamma(y) = C \Big(k \max_{i \leq p} \{(q_i^T y)^2 \setminus |q_i^T y| \leq \lambda\} + \sum_{j \leq p} (q_j^T y)^2 \mathbb{1}_{\{|q_j^T y| \leq \lambda\}} \Big)^{-1}, \tag{8}$$

where q_i corresponds to the ith column of matrix Q in the QR decomposition of X. The constant C leading to the lower approximate quadratic risk in (4) is

$$C = \frac{2(p - 2 - 2k(k+1)/p + 10^{-4}\lambda^2(n+4)(n - 2p + 2df)/(n-p))}{(n - p + 4)(n - p + 6)} \tag{9}$$

3   Simulations

The following example is inspired by [13]: y is a vector of $n = 40$ observations from the random variable Y, and X contains $p = 5$ explanatory variables. The regression coefficient β is set to $(2, 0, 0, 4, 0)^T$. The error vector ε is drawn from two different distributions with variance 1: the Gaussian distribution, which is the usual assumption for most criteria, and the Student distribution with $\nu = 4$ degrees of freedom, corresponding to our assumption of spherical symmetry. We replicate this error vector 5000 times, this way we obtain 5000 regularization paths.

The estimators of loss δ_0 and δ_ρ are those described in (6) and (5) with the choices of the corrective function in (7), (8) and (9). We compare their selection to the real quadratic loss $L(\hat{\mu}, \mu)$ as well as to the classical Akaike's Information Criterion (AIC) and Schwarz' Information Criterion (BIC) with a Gaussian assumption, and the distribution-free leave-one-out cross-validation (LOOCV).

Tables 1 and 2 present for each method the empirical probabilities of selecting the subsets over the 5000 replicates. We iterated the experience ten times to estimate the means and standard deviations shown in the tables. Only the ten most voted subsets are displayed.

Table 1: Empirical probabilities (%) of selection with Firm Shrinkage (Gaussian case).

Subset	δ_0	δ_ρ	AIC	BIC	LOOCV	$L(\hat{\mu}, \mu)$
$\{4\}$	20.18 (0.59)	26.12 (0.56)	20.18 (0.59)	40.05 (0.83)	14.42 (16.18)	14.17 (0.43)
$\{1,4\}$	39.02 (0.74)	**44.41 (0.60)**	39.02 (0.74)	39.37 (0.49)	32.71 (12.27)	54.29 (0.56)
$\{2,4\}$	2.09 (0.22)	3.08 (0.24)	2.09 (0.22)	1.51 (0.20)	3.66 (1.55)	0.00 (0.11)
$\{3,4\}$	2.11 (0.21)	3.23 (0.32)	2.11 (0.21)	1.54 (0.16)	3.17 (1.09)	0.00 (0.15)
$\{4,5\}$	2.05 (0.13)	2.74 (0.19)	2.05 (0.13)	1.47 (0.20)	3.60 (1.56)	0.00 (0.07)
$\{1,2,4\}$	7.57 (0.34)	1.33 (0.15)	7.57 (0.34)	3.66 (0.26)	5.68 (3.06)	7.46 (0.33)
$\{1,3,4\}$	7.83 (0.40)	1.30 (0.13)	7.83 (0.40)	3.73 (0.19)	6.93 (2.99)	7.63 (0.32)
$\{1,4,5\}$	7.73 (0.40)	2.13 (0.20)	7.73 (0.40)	3.73 (0.36)	6.49 (3.73)	7.87 (0.27)
$\{1,2,3,4\}$	2.62 (0.20)	5.03 (0.39)	2.62 (0.20)	0.00 (0.12)	2.56 (1.35)	1.96 (0.22)
$\{1,2,3,4,5\}$	1.34 (0.13)	3.75 (0.23)	1.34 (0.13)	0.00 (0.04)	11.58 (6.84)	1.04 (0.18)

Subset	δ_0	δ_ρ	AIC	BIC	LOOCV	$L(\hat{\mu}, \mu)$
$\emptyset$	9.94 (0.65)	8.87 (0.43)	9.94 (0.65)	20.90 (0.74)	7.21 (3.12)	14.62 (0.45)
$\{4\}$	15.77 (0.37)	19.11 (0.29)	15.77 (0.37)	24.33 (0.45)	12.63 (8.99)	14.88 (0.50)
$\{1,4\}$	32.08 (0.74)	**38.01 (0.62)**	32.08 (0.74)	35.15 (0.82)	26.35 (11.77)	46.08 (0.78)
$\{1,2,4\}$	6.08 (0.21)	0.00 (0.14)	6.08 (0.21)	2.74 (0.16)	5.82 (2.93)	4.65 (0.21)
$\{1,3,4\}$	5.97 (0.19)	0.00 (0.19)	5.97 (0.19)	2.75 (0.27)	5.62 (3.48)	4.72 (0.20)
$\{1,4,5\}$	6.21 (0.36)	1.63 (0.22)	6.21 (0.36)	2.83 (0.20)	6.58 (3.39)	4.50 (0.16)
$\{1,2,3,4\}$	2.12 (0.19)	4.03 (0.29)	2.12 (0.19)	0.00 (0.08)	2.84 (1.47)	1.34 (0.13)
$\{1,2,4,5\}$	2.01 (0.25)	2.23 (0.23)	2.01 (0.25)	0.00 (0.06)	2.64 (1.14)	1.30 (0.15)
$\{1,3,4,5\}$	2.09 (0.19)	2.15 (0.19)	2.09 (0.19)	0.00 (0.08)	2.67 (1.19)	1.37 (0.19)
$\{1,2,3,4,5\}$	1.04 (0.13)	3.10 (0.16)	1.04 (0.13)	0.00 (0.05)	11.45 (3.47)	0.00 (0.11)

Table 2: Empirical probabilities (%) of selection with Firm Shrinkage (Student case).

Here, we can see that the corrective estimator δ_ρ selects the right model $\{1, 4\}$ with higher probability than the unbiased estimator δ_0 and the classical methods even for the usual *i.i.d.* Gaussian case, and is closer to the real loss results. Improvement might be even larger if we consider a more general form of correction, like $\delta^* = \alpha(\delta_0 - \|y - X\hat{\beta}^{FS}\|^4\gamma)$ for instance, as was done in [13]. But the most striking result is the low empirical probabilities of all the methods. The real loss $L(\hat{\mu}, \mu)$ manages to select the right subset only around half of the time, bounding from above the probability of selection with our criterion. On the contrary, the results in [13] showed that systematic exploration with subset selection leads to select $\{1, 4\}$ with an empirical probability of 83% for their corrective estimator (which is slightly different from the one in (1)), that is to say twice the probability obtained with Firm Shrinkage. This important difference is a consequence of the regularization path, which sometimes introduces irrelevant variables first and fails to select the right subset. It also explains the selection by δ_0 and AIC of subsets of size 3 containing the two relevant variables, with empirical probabilities around 7%, or subsets of bigger size by δ_ρ and LOOCV.

4 Discussion

We proposed a new data-driven procedure of model selection based on loss estimation and valid for the whole family of spherically symmetric distributions, allowing dependence between noise components. From this procedure, we derived a criterion for the Firm Shrinkage estimator, a regularization path method with low bias, evaluated by a quadratic loss. In the experience we drove, this criterion performs better than classical criterion like AIC, BIC and LOOCV, even under the usual *i.i.d.* Gaussian assumption for the noise.

Note that this procedure can be applied to other estimators and loss functions, and this way it could be adapted to problems such as classification or reinforcement learning.

References

[1] D. Fourdrinier and S. Pergamenshchikov. Improved model selection method for a regression function with dependent noise. *Annals of the Institute of Statistical Mathematics*, 59(3):435–464, 2007.

[2] I. Guyon, A. Saffari, G. Dror, and G. Cawley. Model selection: beyond the bayesian/frequentist divide. *The Journal of Machine Learning Research*, 11:61–87, 2010.

[3] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32(2):407–451, 2004.

[5] C.L. Mallows. Some comments on cp. *Technometrics*, pages 661–675, 1973.

[6] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. ISSN 0035-9246.

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

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

[9] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, volume 1, pages 267–281. Springer Verlag, 1973.

[10] A.G. Bruce and H.Y. Gao. Understanding waveshrink: Variance and bias estimation. *Biometrika*, 83(4):727, 1996.

[11] AT Walden, DB Percival, and EJ Mccoy. Spectrum estimation by wavelet thresholding of multitaper estimators. 1995.

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

[13] D. Fourdrinier and MT Wells. Comparaisons de procédures de sélection d'un modèle de régression: une approche décisionnelle. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 319(8):865–870, 1994.