Information criteria for structured sparse variable selection

Bastien Marquis[*1] **and Maarten Jansen**[1]

[1] *Université Libre de Bruxelles, departments of Mathematics*

In contrast to the low dimensional case, variable selection under the assumption of sparsity in high dimensional models is strongly influenced by the effects of false positives. The effects of false positives are tempered by combining the variable selection with a shrinkage estimator, such as in the lasso, where the selection is realized by minimizing the sum of squared residuals regularized by an ℓ_1 norm of the selected variables. Optimal variable selection is then equivalent to finding the best balance between closeness of fit and regularity, i.e., to optimization of the regularization parameter with respect to an information criterion such as Mallows's Cp or AIC. For use in this optimization procedure, the lasso regularization is found to be too tolerant towards false positives, leading to a considerable overestimation of the model size. Using an ℓ_0 regularization instead requires careful consideration of the false positives, as they have a major impact on the optimal regularization parameter. As the framework of the classical linear model has been analysed in previous work, the current paper concentrates on structured models and, more specifically, on grouped variables. Although the imposed structure in the selected models can be understood to somehow reduce the effect of false positives, we observe a qualitatively similar behavior as in the unstructured linear model.

Keywords: variable selection, structured data, sparsity, lasso, Mallows's Cp.

1 Introduction

Recent literature has had considerable attention for the uncertainties that follow from the process of model or variable selection. On one hand, it has been realized that the selection of variables should look forward, focussing on the application in which the selected model will be used, so as not to waste degrees of freedom on variables that are of little importance in the application [2]. On the other hand, post-model selection inference is looking backwards,

*Corresponding author: bastien.marquis@ulb.ac.be

investigating the effects of the model selection uncertainty on the inference in the selected model [7, 6].

The contribution of this paper is, however, situated on the effect of the uncertainty on the variable selection process itself. The numerous insignificant components in sparse, high dimensional models lead to false positives being a main source of uncertainty. Well established methods for high dimensional variable selection are explicitly based on controlling the false discovery rate [1] or even the absolute number of false positives [4]. The methods in this class tend to be minimax oriented, rather than data driven. Another way to deal with false positives is to reduce the impact of a false positive by using shrinkage selection. This is realized, for instance, in the lasso, where the variable selection objective is formulated as a trade off between the sum of the residual squares and the ℓ_1 norm of the selected variables. The ℓ_1 norm, i.e., the sum of the absolute values of the selected variables, should be seen as an alternative for the ℓ_0 norm, measuring the size of the selected set. Finding the minimum sum of squared residuals, regularized by the number of selected variables, is a combinatorial problem, and therefore intractable from the computational point of view. The ℓ_1 alternative leads to a quadratic programming problem whose solution is still a proper variable selection, as it contains many exact zeros. The nonzeros, however, are not found by least squares projection, but rather by shrunk versions of the least squares estimators. The intuition behind this is that dubious parameters can be included into the model, but with a value close to zero. If such a parameter happens to be a false positive, its inclusion into the model has a limited impact on any inference in that model. With a much faster algorithm than its ℓ_0 counterpart, the ℓ_1 regularized variable selection, equiped with an appropriate choice of the regularization parameter, is able to find a model with a similar degree of sparsity [3].

Existing variable selection consistency results do not consider the case where the regularization parameter has to be optimized in a data dependent way, using an information criterion. While for fixed or minimax values of the parameter, ℓ_1 regularization provides a valid alternative for ℓ_0, the equivalence holds no longer through the optimization process. This is explained by the ℓ_1 tolerance towards false positives: since the ℓ_1 procedure reduces the impact of a false positive, the optimal balance between the sum of the residual squares and the regularization shifts towards larger models.

In searching for the optimal regularization, ℓ_1 can still be used to actually come up with a selection, but for the evaluation of the quality of the selection, it makes a difference whether the estimation within the selection keeps the shrinkage of the ℓ_1 regularization. If the shrinkage estimator is replaced by a least squares projection, then the optimal balance should shift back towards smaller models. It is obvious that the estimation of the ℓ_0 balance requires

a different expression of the information criterion. The compensation for the difference between ℓ_0 and ℓ_1 regularization has been described as a "mirror" effect [5], further explained in Section 2. It has been explored in the context of unstructured selection in a linear model. In this paper, we extend the scope to structured selection, presented in Section 3. The actual contribution of the paper then follows in Section 4.

2 Mirror effect in unstructured selection in linear models

Consider the sparse linear model

$$\mathbf{Y} = \mathbf{K}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the design matrix $\mathbf{K}$ has size $n \times m$ with n smaller than m, and the number of nonzeros in $\boldsymbol{\beta}$ is unknown but smaller than n. Also, let $A_s \subset \{1, 2, \ldots, m\}$ be a selection with s nonzeros, obtained by a procedure, $\mathcal{S}(\mathbf{Y}, s)$, that selects among all possible subsets of size s. As an example, $\mathcal{S}(\mathbf{Y}, s)$ could be an implementation of the lasso, finetuned to have s nonzeros as result. Furthermore, let $\mathbf{K}_{A_s}$ denote the $n \times s$ submatrix consisting of the s columns in $\mathbf{K}$ corresponding to the selection. We investigate the quality of the least squares projection $\widehat{\boldsymbol{\beta}}_{A_s} = (\mathbf{K}_{A_s}^T \mathbf{K}_{A_s})^{-1} \mathbf{K}_{A_s}^T \mathbf{Y}$, assuming that $\mathbf{K}_{A_s}$ is non-singular. As a measure for quality, we adopt the prediction error, but a similar discussion would hold for any distance between selected and true model. The prediction error is defined as $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s})$, where

$$\mathrm{PE}(\widehat{\boldsymbol{\beta}}_A) = \frac{1}{n} E\left(\|\mathbf{K}\boldsymbol{\beta} - \mathbf{K}_A \widehat{\boldsymbol{\beta}}_A\|_2^2 \right). \tag{1}$$

Let A_s^o be the selection provided by an oracle observing $\mathbf{K}\boldsymbol{\beta}$ without noise, using the same procedure, i.e., $A_s^o = \mathcal{S}(\mathbf{K}\boldsymbol{\beta}, s)$. Then the least squares projection, $\widehat{\boldsymbol{\beta}}_{A_s^o} = (\mathbf{K}_{A_s^o}^T \mathbf{K}_{A_s^o})^{-1} \mathbf{K}_{A_s^o}^T \mathbf{Y}$, depends on the observations through $\mathbf{Y}$, but not through A_s^o. The prediction error $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o})$ is estimated unbiasedly by $\Delta_p(A_s^o)$, where $\Delta_p(A)$ is a non studentized version of Mallows's Cp criterion,

$$\Delta_p(\widehat{\boldsymbol{\beta}}_A) = \frac{1}{n} \|\mathbf{Y} - \mathbf{K}_A \widehat{\boldsymbol{\beta}}_A\|_2^2 + \frac{2|A|}{n} \sigma^2 - \sigma^2. \tag{2}$$

The selection $A_s = \mathcal{S}(\mathbf{Y}, s)$, however, depends on $\mathbf{Y}$. The expectation of (2) will not be equal to $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s})$. As the second and third term of (2) are constants, this is explained by the behavior of $\|\mathbf{Y} - \mathbf{K}_{A_s} \widehat{\boldsymbol{\beta}}_{A_s}\|_2^2$. In the case where the procedure consists of minimizing (2) on all selections of size s, i.e., $\mathcal{S}(\mathbf{Y}, s) = \arg\min_{|A|=s} \Delta_p(A)$, the deviation of the information criterion from

the error curve can be described as a reflection with respect to the oracular mirror $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o}) = E\Delta_p(\widehat{\boldsymbol{\beta}}_{A_s^o})$ [5], meaning that

$$\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s}) - \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o}) \approx \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o}) - E\Delta_p(\widehat{\boldsymbol{\beta}}_{A_s}) \tag{3}$$

An intuitive explanation follows by assuming that s is large enough to catch all really important variables into both A_s and A_s^o. Once the important variables are in the model, the remainder of the s variables are chosen to further minimize the distance between $\mathbf{K}_{A_s}\widehat{\boldsymbol{\beta}}_{A_s}$ and $\mathbf{Y}$. Among the remaining candidates, these variables perform best in fitting the signal $\mathbf{K}\boldsymbol{\beta}$ with the errors, and thus perform worst in staying close to signal without the errors. The contrast between the better-than-average appearance $E\Delta_p(\widehat{\boldsymbol{\beta}}_{A_s})$ and worse-than-average true prediction error follows from the fact that the optimisation over random variables $\Delta_p(\widehat{\boldsymbol{\beta}}_A)$ affects the statistics of the selected values. The oracle curve $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o})$ acts as mirror, because the selection A_s^o does not depend on $\mathbf{Y}$, thus leaving the statistics of the selected values unchanged.

3 Structured selection with grouped variables

The lasso, in addition to providing us with a selection A_s considering an appropriate regularisation parameter, can be extented or used to take into account structured models such as grouped variables [9], graphical models [10, 8] or even hierarchical information [11]. When the variables are under the hypothesis to have a natural group structure, the coefficients within a group should all be nonzero (or zero).

In its Lagrangian form, the lasso problem of a linear model is expressed as

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{K}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{4}$$

with λ being a regularisation parameter which can be adjusted to obtain the desired degree of sparsity. When $\mathbf{K}$ is orthogonal, the solution of (4) is simply a soft-thresholded version of the least-squares estimate whose threshold is λ. For the remainder of this paper, we consider the signal-plus-noise model $\mathbf{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $m = n$ and $\mathbf{K} = \mathbf{I}_n$. Then the best s term unstructured selection, mesured by the Cp-value, consists of the s largest elements from $\mathbf{Y}$.

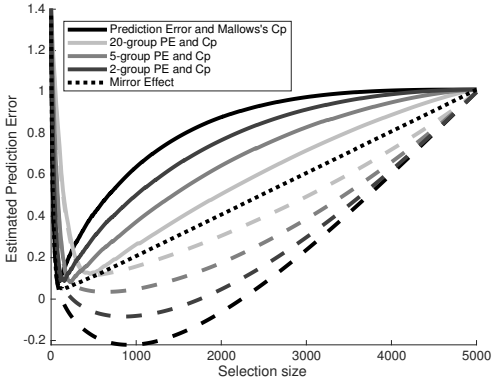
For group selection, the penalty in (4) can be modified to become the sum of the ℓ_2 norms of each group. This is known as group lasso and it aims to optimise the following expression, for the signal-plus-noise model with n_g groups,

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{n_g}\|\boldsymbol{\beta}_j\|_2 \tag{5}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{w_j}$ forms a group of w_j coefficients from $\boldsymbol{\beta}$ and $\sum_{j=1}^{n_g} w_j = n$. The solution of (5) is again a soft-thresholded version of $\mathbf{Y}$, although the threshold has the form $\lambda|Y_i|/\|\mathbf{Y}_j\|_2$ for observation i within group j. Hence without shrinkage, the best s_g group selection contains the values of $\mathbf{Y}$ from the s_g groups of observations whose ℓ_2 norms are the largest.

4 Mirror effect in group selection and discussion

In our simulation, 250 groups containing 20 coefficients $\boldsymbol{\beta}_j$ are generated so that $\boldsymbol{\beta} = (\boldsymbol{\beta}_j)_{j=1,\ldots,250}$ is a n-dimensional vector with $n = 5000$. Within group j, the $\boldsymbol{\beta}_j$ have the same probability p_j of being set to 0; for each j, a different value p_j is randomly drawn from the set $\mathbf{P} = (0.95, 0.80, 0.50, 0.05, 0.00)$ with respective probability $\mathbf{Q} = (0.02, 0.02, 0.01, 0.20, 0.75)$. The expected proportion of nonzeros is then $\langle \mathbf{P}, \mathbf{Q} \rangle = 1/20$ for the whole data $\boldsymbol{\beta}$. The nonzeros $\boldsymbol{\beta}$ are then distributed according to the zero inflated Laplace model $f_{\beta|\beta\neq 0}(\beta) = (a/2)\exp(-a|\beta|)$ where $a = 1/5$. The observations are $\mathbf{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a n-vector of independent, standard normal errors. Estimates $\widehat{\boldsymbol{\beta}}$ are calculated considering four configurations: groups of size 20 (initial setting), 5 and 2 (subgroups built from the original groups) and 1 (unstructured selection).



The PE and Cp curves, solid and dashed lines respectively, are represented as functions of the selection size, for different sizes of group. The dotted line depicts the mirror curve estimated for unstrutured variables [5].

Figure 1: Mirror effect and group size impact.

Figure 1 plots the prediction error and Mallows's Cp as a function of the selection size for unstructured and 20-5-2-grouped variable selection. In each case, we observe that the PE and Cp curves are reflexion of each other with respect to a mirror curve. It is interesting to note that, for the signal-plus-noise model, the unstructured and group mirror curves coincide once the PE and

Cp curves are drifting apart. Also, it seems the bigger the group size gets, the closer the corresponding PE and Cp curves are. Hence when the group size grows, the mirror effect becomes smaller.

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSSB*, 57:289–300, 1995.

[2] G. Claeskens and N. Hjort. The focused information criterion. *JASA*, 98: 900–916, 2003.

[3] D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[4] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[5] M. Jansen. Information criteria for variable selection under sparsity. *Biometrika*, 101(1):37–55, 2014.

[6] J. D. Lee, D. L. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[7] H. Leeb and B.M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591, 2006.

[8] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[9] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1): 49–67, 2007a.

[10] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007b.

[11] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37:3468–3497, 2009.