Journal of Statistical Planning and Inference 143 (2013) 1457–1465

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Grouping strategies and thresholding for high dimensional linear models rejoinder

Mathilde Mougeot, Dominique Picard *, Karine Tribouley

Université Paris-Diderot, CNRS LPMA, Sophie Germain, 75013 Paris, France

ARTICLE INFO

Available online 2 April 2013

Keywords:
Structured sparsity
Grouping
Learning theory
Non-linear methods
Block-thresholding
Coherence
Wavelets

We are very grateful to all the discussants for the stimulating comments and questions that they have raised concerning the advantages/drawbacks of using two step procedures, the role of structured sparsity, the possible gains or losses of grouping or using grouping strategies. Many points have been raised, developed or clarified; many open questions have been posed pointing towards exciting future directions for this line of work. We concentrate our discussion on the following points:

- Apart from the enthusiasm of developing a procedure simple to implement and with comparatively good performances in theory and in practice, one of our ideas in this work was that, in high dimensional models, sparsity conditions and conditions on the matrix of covariates are deeply connected. It is not entirely simple how to express this link in the best way. It seems that considering the case of structure in the covariates – with many possible meanings of this – as well as a procedure with simple operations and simple requirements would possibly be an opportunity to shed some light over this question. It is noticeable that in all the discussions this point was raised in one way or another.
- We also discuss what sort of gain and which amplitude we can expect by grouping the covariates inducing structure, although it may seem doubtful in some situations.
- We come back to the BRG procedure which provides an algorithm for grouping the covariates in view to boost the rates of convergence. We emphasize that there is basically two steps in this procedure: scattering and gathering. The scattering step aims at organizing the covariates into groups to reduce the coherence indicator. This step based on a diversity principle brings quite often interesting information. The gathering step aims at reducing the 'group sparsity' of the coefficients. The adjustment of the balance between these two steps is not simple and there is probably room for improvement there in future researches.
- Our results provide upper bounds for the rates of convergence of the algorithm. Obviously a natural question (raised in different ways in the discussions) lies in the sharpness of these results, in terms of coherence property for instance. We explain in a theoretical framework that the bounds obtained are somehow unavoidable, meaning that the procedure

* Corresponding author.
E-mail address: dominique.picard@univ-paris-diderot.fr (D. Picard).

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.03.007

could possibly be improved, but to the price of changing the thresholds in the different steps. We also confront our procedure with a practical very unfavorable situation and observe that it behaves quite correctly despite very poor theoretical prediction.
- We comment the properties of the thresholding methods in terms of finding the support of the coefficients.
- Finally we investigate over an example the aptitude of the grouping strategy to catch features in the data by inducing diversity.

1. Comments on sparsity conditions/RIP-type conditions

When the general linear model $Y = X\beta + W$ is considered ($\beta \in \mathbb{R}^k$ is the unknown parameter, X is a known matrix $n \times k$ and W is the vector of the errors), there is obviously an identification problem for the parameter β when k is a large number compared to n. Generally both important conditions are assumed to solve the problem:

1. The first one (which somehow seems unavoidable) consists in assuming that the parameter β only contains a few contributing elements.
2. The second condition concerns the matrix X and basically asks, under different forms, that some near orthogonality arises among columns, when only a small number of them is considered.

In the following paragraphs of this section, we discuss in more details these conditions, their consequences and the way they can be entangled.

1.1. Case without structure on the covariates

Sparsity condition. It can be explained in a quite simple way, either by assuming that the number of contributing coefficients is small i.e.

$$\exists S > 0, \quad \{\ell \in \{1, \dots, k\}, \beta_\ell \neq 0\} \leq S,$$

or by assuming that there exists $q \in (0, 1)$ such that

$$\sum_{\ell} |\beta_\ell|^q \leq M^q \tag{1}$$

which is an extension of the l_0 norm. These assumptions summarize that only a small number of coefficients are significant.

Condition against the multi-colinearity problem. Another set of conditions, which are not so obviously necessary at first glance, are conditions we could call *restricted identity properties* of the Gram matrix. Recall that the Gram-matrix associated to the subsect $\mathcal{C}$ of $\{1, \dots k\}$ is defined by $\Gamma(\mathcal{C}) = n^{-1} X_{\mathcal{C}}^t X_{\mathcal{C}}$ where $X_{\mathcal{C}}$ is the restriction of the matrix X to the columns with indices in $\mathcal{C}$. Roughly speaking the *restricted identity property* means that $\Gamma(\mathcal{C})$ is almost the identity matrix as soon as the cardinality $m = \mathcal{C}$ is small enough. The most famous example of such a property is the standard *restricted isometry property* depending on the parameters m_0 and ν. For any $m \leq m_0$ and subset $\mathcal{C}$ with cardinality less than m, it assumes that

$$\forall x \in \mathbb{R}^m, \quad (1-\nu)\|x\|_{l_2}^2 \leq x^t \Gamma(\mathcal{C}) x \leq (1+\nu)\|x\|_{l_2}^2. \tag{2}$$

Another type of such conditions consists in bounding the extra diagonal terms of the Gram matrix, introducing the *coherence* (suppose to simplify that $\Gamma_{jj} = 1$ for any j)

$$\gamma_n = \sup_{\ell \neq m} |\Gamma_{\ell m}| = \sup_{\ell \neq m} |\Gamma_{\ell m}(\{1, \dots, k\})|.$$

Of course, coherence conditions are simpler to verify but also more stringent since $\gamma_n \leq \gamma$ for some γ implies that the restricted isometry property (2) is verified for ν and $m_0 = \lfloor \nu/\gamma \rfloor$.

These conditions are used as a remedy to a fundamental issue of high dimensional problems, as summarized in Obozinski's discussion: *one of the known sources of instability of many sparse methods is that if an irrelevant variable is too correlated with a relevant one, it could be spuriously included in the estimated support, and, conversely, a relevant variable too correlated with another incurs the risks of being ignored.*

1.2. Grouping the covariates

Grouping consists in re-arranging the k predictors into p ($p \leq k$) groups of variables $X_{\mathcal{G}_1}, \dots, X_{\mathcal{G}_p}$ where $\mathcal{G}_1, \dots, \mathcal{G}_p$ is a partition of $\{1, \dots, k\}$. For any j in $\{1, \dots, p\}$, $\mathcal{G}_j$ has size t_j and for each $\ell = 1, \dots, k$, the predictor X_ℓ is now registered as $X_{(j,t)}$ where $j \in \{1, \dots, p\}$ is the index of the group $\mathcal{G}_j$ and $t = r_j(\ell) \in \{1, \dots, t_j\}$ is the rank of ℓ inside the group $\mathcal{G}_j$.

Enforcing the structure by the experiment. Grouping procedures are generally used in the presence of structural relationship for data (see for instance, the multi-task case, or the sign-coherent groups as in Chiquet et al., 2012) or when the desired result is constrained (i.e. select or discard all the variables of the same group). The groups are formed in advance