Journal of Statistical Planning and Inference 171 (2016) 147–174

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Orthogonal One Step Greedy Procedure for heteroscedastic linear models

Marc-Antoine Giuliani

Université Paris Diderot, CNRS LPMA, 8 place FM/13, 75013, Paris

ARTICLE INFO

Article history:
Received 17 April 2015
Received in revised form 12 October 2015
Accepted 13 October 2015
Available online 26 October 2015

Keywords:
High-dimensional linear models
Greedy
Heteroscedastic

ABSTRACT

This paper investigates the prediction problem in the general Gaussian linear model with correlated noise, under the assumption that the covariance matrix is known, and focuses particularly on the high dimensional setting. We adapt an overly greedy procedure, where the relevant covariates are selected initially in one pass on the data, without any iteration, nor optimization. A simple componentwise regression, followed by an adaptive thresholding, locates leaders among the regressors to reduce the initial dimensionality. A second adaptive thresholding is performed on the linear regression upon the leaders. These steps take into account the correlated structure of the noise, by using weights associated to the covariates in a modified norm induced by the covariance matrix of the noise. The consistency of the procedure is investigated, and rates are provided for a wide range of sparsity classes, with little restriction on the number of regressors. An extensive computational experiment is conducted to emphasize the fact that the good theoretical results are corroborated by quite good practical performances in the presence of correlated noise.

1. Introduction

Consider the following linear model

$$Z = \Psi\alpha + \eta,$$

where we observe the n-dimensional vector Z, and the $n \times p$ design matrix Ψ. The p-dimensional vector α is the signal to be estimated, while the n-dimensional vector η is an unobservable noise. The case where the number of regressors p is large compared to the number n of observations is the focus of a lot of attention in contemporary statistics. Indeed, such models have many practical applications ranging from genomics, where the number of possibly involved genes in a pathology can be huge compared to the little number of affected people, to image analysis, where the number of unknown pixels can be very large compared to the number of measurements. Natural language processing is another important field of applications: document-term matrices, where each line represents a text from a given corpus and each column a word belonging to one of the texts, leading necessarily to very high dimensional models.

The problem of estimating α in such a high dimensional setting is impossible to solve in full generality. But it can become feasible if some measure of the intrinsic dimension of the signal is in fact much smaller than the dimension of the ambient space $\mathbb{R}^p$. This is referred to as the sparsity of the signal. Many computationally reasonable and theoretically efficient algorithms have been proposed in the literature, using greedy methods (Mallat and Zhang, 1993; Tropp, 2004; Needell and Vershynin, 2009; Zhang, 2011) or the extraordinary explosive domain of ℓ_1 penalties which we can barely

E-mail address: giuliani@math.univ-paris-diderot.fr.

http://dx.doi.org/10.1016/j.jspi.2015.10.008
0378-3758/© 2015 Elsevier B.V. All rights reserved.

reference: Tibshirani (1994), Candes and Tao (2007) and van de Geer et al. (2011) being a few of the references on the topic. For a much more complete bibliography we can refer to Bühlmann and van de Geer (2011).

Besides the sparsity of the signal, other conditions appear to be also necessary to solve the problem, basically to prevent multi-collinearities for the columns of the matrix Ψ. Most of the results in the papers cited above are obtained under RIP type-conditions. Recall that the Gram-matrix associated to the subset $\mathcal{C}$ of $\{1, \dots, p\}$ is defined by $G(\mathcal{C}) = n^{-1} \Psi_{\mathcal{C}}^t \Psi_{\mathcal{C}}$ where $\Psi_{\mathcal{C}}$ is the restriction of the matrix Ψ to the columns with indices in $\mathcal{C}$. Roughly speaking the Restricted Identity Property (RIP) means that $G(\mathcal{C})$ is almost the identity matrix as soon as the cardinality $m = |\mathcal{C}|$ is small enough. However this condition can seem quite drastic if the problem is only to avoid too many multi-colinearities. Indeed, one could imagine for instance, replacing '$G(\mathcal{C})$ is almost the identity matrix' by the more flexible condition: $G(\mathcal{C})$ is an invertible matrix. And one might wonder how the results would be affected by such a less restrictive condition. The answer to this question is quite unclear, and one goal of this paper is to shed some light on this aspect.

The problem appears in quite a clear way for instance in models derived from inverse problems where the eigenvalues of the matrices $G(\mathcal{C})$ can depend in a crucial way on the set $\mathcal{C}$. An example of such a case occurs when Ψ is in fact the multiplication of a $n \times n$ symmetric definite positive matrix K by a $n \times p$ matrix X obeying RIP conditions. In practice this is corresponding to a compressed sensing situation where the responses are not only perturbed by noise but are also blurred by the filter K.

This paper will focus on the equivalent problem of the heteroscedastic setting, where instead of assuming that the noise components are independent identically distributed random variables, we suppose that the vector η has a covariance matrix Γ. In such a situation, the usual "low-dimensional" ($p\lambda_2 n$) intuition would be to "pre-whiten" the noise by multiplying Z and Ψ by $\Gamma^{-1/2}$, in order to consider a homoscedastic model. But doing so will modify the correlation among the covariates and may lead to very poorly conditioned design matrix. For example, even if the initial design Ψ verifies the RIP, there is a no reason to believe that $\Gamma^{-1/2}\Psi$ will still do so, and that we will not make estimation much worse by starting with this "whitening" operation. Our greedy procedure wants to avoid such a transformation, starting from a well conditioned design Ψ (with low coherence), we modify a greedy procedure to take into account the covariance Γ without risking to deteriorate the conditioning of Ψ.

Although most of the works cited above have been investigating the homoscedastic setting, several works have been conducted in this direction where the noise has a non trivial covariance, studying the behavior of the classical lasso estimator (Tibshirani, 1994), or the adaptive lasso estimator (Zou, 2006) in this correlated setting, avoiding to "pre-whiten" the noise. In Dette and Wagener (2013) and Wagener and Dette (2013) it is proved that the adaptive lasso is consistent and asymptotically normal in a heteroscedastic setting with p fixed and n growing, but with suboptimal variance. A correction is proposed with a weighted adaptive lasso estimator which has optimal asymptotic variance. In Wagener and Dette (2012) this analysis is extended to the more general bridge estimators. A modification of Lasso and Pseudo-Lasso in the context of linear instrumental variables models able to handle the heteroscedastic setting even with unknown covariance is proposed in Belloni et al. (2012), where sharp convergence rates are proved under the hypothesis that $\log p = o(n^{1/3})$. In Jia et al. (2010) it is shown that the lasso is sign consistent in a Poisson-like model when the signal to noise ratio is large enough. Furthermore the trade off implied by the "pre-whitening" operation has already been investigated in the context of Lasso pre-conditioning, for example in Wauthier et al. (2013), Jia and Rohe (2012), Qian and Jia (2012), Rohe (2015), or Rauhut and Ward (2011).

We will adapt to the heteroscedastic setting an overly greedy procedure studied in the white noise setting in the series of papers (Kerkyacharian et al., 2009; Mougeot et al., 2012, 2013, 2014). The LOL algorithm is an Orthogonal One-Step Greedy (OOSG) procedure, which extends the classical thresholding theory to high-dimensional linear models (even if it is not necessary, the algorithm is still usable for classical low dimensional models, where the number of observations is larger than the number of covariates). One-Step Greedy procedures are typical selection/estimation procedures in the sense of Foster and George (1994): in a first step they select a number N of covariates by independent screening (Fan and Lv, 2008), then perform least squares regression on those covariates, the resulting estimator being finally thresholded. The number N and the threshold are data driven, giving an adaptive procedure. This procedure can behave in the high dimensional setting almost as well as much more sophisticated procedures involving optimization steps. The strength of this kind of method is its extreme simplicity (and numerical efficiency). As a drawback, they rely for instance on coherence conditions instead of RIP assumptions. In the context of heteroscedasticity we will see that precisely the simplicity of these types of condition becomes helpful to disentangle with the parts linked to the covariance.

To adapt such a procedure to the covariance structure of the errors, we will modify the OOSG methodology by incorporating in the thresholds weights related to the size of the columns of the design in the norm induced by the covariance matrix of the noise. This simple modification allows us to obtain convergence rates (in Section 4) on weighted ℓ_q balls driven by the standard behavior of an inverse problem term involving additionally the coherence and the sparsity of the signal, together with a term taking into account the location of the signal among the regressors. Indeed, a basic effect of the presence of a non standard covariance is to bring disparity between the potential precisions of estimation of each coordinate of the signal.

To obtain such results we need to modify the concentration inequality on the norm of the orthogonal projection of the noise to take into account not only the dimension of the subspace generated by the selected covariates, but its position too (see Proposition 3). Since we work under assumptions which allow the selection procedure to behave well, this leads to rates driven by the location among the covariates of the support of the signal of interest α.