Variable clustering in high dimensional linear regression models

Loïc Yengo, Julien Jacques, Christophe Biernacki

Submission

Journal de la Société Française de Statistique

Variable clustering in high dimensional linear regression models

Titre: Classification de variables en régression linéaire

Loïc Yengo [1,2,3] , Julien Jacques [2,3] and Christophe Biernacki [2,3]

Abstract: For the last three decades, the advent of technologies for massive data collection have brought deep changes in many scientific fields. What was first seen as a blessing, rapidly turned out to be termed as the curse of dimensionality. Reducing the dimensionality has therefore become a challenge in statistical learning. In high dimensional linear regression models, the quest for parsimony has long been driven by the idea that a few relevant variables may be sufficient to describe the modeled phenomenon. Recently, a new paradigm was introduced in a series of articles from which the present work derives. We propose here a model that simultaneously performs variable clustering and regression. Our approach no longer considers the regression coefficients as fixed parameters to be estimated, but as unobserved random variables following a Gaussian mixture model. The latent partition is then determined by maximum likelihood and predictions are obtained from the conditional distribution of the regression coefficients given the data. The number of latent components is chosen using a BIC criterion. Our model has very competitive predictive performances compared to standard approaches and brings significant improvements in interpretability.

Résumé : Les trois dernières décennies ont vu l'avènement de profonds changements dans de nombreuses disciplines scientifiques. Certains de ces changements, directement liés à la collecte massive de données, ont donné naissance à de nombreux défis en apprentissage statistique. La réduction de la dimension en est un. En régression linéaire, l'idée de parcimonie a longtemps été associée à la possibilité de modéliser un phénomène grâce à un faible nombre de variables. Un nouveau paradigme a récemment été introduit dans lequel s'inscrivent pleinement les présents travaux. Nous présentons ici un modèle permettant simultanément d'estimer un modèle de régression tout en effectuant une classification des covariables. Ce modèle ne considère pas les coefficients de régression comme des paramètres à estimer mais plutôt comme des variables aléatoires non observées suivant une distribution de mélange gaussien. La partition latente des variables est estimée par maximum de vraisemblance. Le nombre de groupes de variables est choisi en minimisant le critère BIC. Notre modèle possède une très bonne qualité de prédiction et son interprétation est aiseée grâce à l'introduction de groupe de variables.

Keywords: Dimension reduction, Linear regression, Variable clustering

Mots-clés : Réduction de la dimension, Régression linéaire, Classification de variables

AMS 2000 subject classifications: , ,

[1] CNRS UMR 8199, Laboratoire génomique et maladies métaboliques, France.
E-mail: loic.yengo@good.ibl.fr

[2] CNRS UMR 8524, Laboratoire Paul Painlevé, Université Lille 1, France.

[3] Inria Lille Nord-Europe, Équipe MODAL, France.

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

1. Introduction

We consider the standard linear regression model defined as

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,\ i = 1,\ldots,n. \tag{1}$$

For an individual i, y_i is the observed response, x_{ij} is an observed value for the j-th covariate and ε_i is an error term often assumed to be normally distributed. The ε_i's are also assumed to be independent and identically distributed. β_j is the regression coefficient associated with the j-th covariate. We denote $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)$ as the vector of regression coefficients.

The dimension p of model (1) is tightly related to both its interpretability and ability to yield reliable predictions. It is well known that the more covariates we add to the model the harder becomes its interpretation. Besides, Stein established in [3] that the mean prediction squared error attributable to a linear regression model increases with its dimension. Reducing the model dimension therefore pursues the goal of minimizing prediction error while keeping the model interpretable. This problem, also referred to as the *bias-variance trade-off* [21], becomes increasingly challenging as the set of covariates exceeds the sample size. This high dimensional framework has fueled a number of researches during the last three decades.

Variable selection is one of the most popular approaches for reducing dimensionality. Although it has a direct impact on p, traditional stepwise algorithms for finding the best subset of predictors had a mitigated success because of their heavy computational burden [21]. At a more affordable computational cost, penalized approaches were introduced as an efficient alternative for variable selection. Penalized approaches impose a constraint on $\boldsymbol{\beta}$ that generally depends on a tuning parameter. This parameter can be selected over a grid of values either minimizing the out-of-sample prediction error (cross validation) or using information based criteria like AIC or BIC [15, 10]. Among the most emblematic methods belonging to this second family of approaches we can refer to the least absolute shrinkage and selection operator (LASSO) [18] and the elastic net [14].

Another relevant approach for reducing dimensionality consists of identifying patterns under which covariates can be pooled together. This idea was recently implemented in a gene expression study [24]. In that study, groups of genes were built from hierarchical clustering of gene expression levels. The authors created surrogate covariates by averaging gene expression levels within each group. Those new predictors were afterwards included in a linear regression model, replacing the primary variables. The major limitation in this approach is the independence between the prediction and clustering parts. Consequently, effects of the surrogate covariates can be diluted if they contain primary variables with either no effect or even opposite effects on the response. To sidestep the previous limitation, Bondell and Reich [4] introduced in 2008 the octogonal shrinkage and clustering algorithm for regression (OSCAR). The OSCAR methodology belongs to the family of penalized approaches. It imposes a constraint on $\boldsymbol{\beta}$ that is a weighted combination of the L^1 norm and the pairwise L^∞ norm. Upper-bounding the pairwise L^∞ norm enforces the covariates to have close coefficients. When the constraint is strong enough, closeness translates into equality achieving thus a grouping property. More recently, a generalization to the OSCAR methodology was proposed in [5]. One major advantage of this new approach, namely the pairwise

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

absolute clustering and sparsity (PACS) was the reduced computational cost. In the aftermath of OSCAR and PACS, other methodologies aiming at simultaneously performing parameter estimation and clustering were proposed. We can for instance refer to the approaches of Petry [20] and She [25] which also mixed L^1 and pairwise L^∞ penalties or those of Daye [16] and Shen [23] based on alternative penalties.

In line with the latter works, we introduce the clusterwise effect regression (CLERE), a new methodology aiming at simultaneously performing regression and clustering of covariates. CLERE considers each β_j no longer as a fixed parameter but as an unobserved random variable (Assumption *A1*) following a mixture of Gaussian distributions (Assumption *A2*) with an arbitrary number of components (Assumption *A3*). The means of each component in the mixture are moreover assumed unequal (Assumption *A4*). Under assumptions *A1* and *A2* our approach shows strong similarities with a Bayesian approach for variable selection known as the spike and slab model [22, 13]. Despite those similarities, assumptions *A3* and *A4* drive the main differences between the two methods. Indeed, in spike and slab models the number of components is restricted to two and the means of each component of the mixture are assumed equal to zero. In addition to those two differences, we recall an important issue which is that our primary goal is not variable selection like for spike and slab models but variable clustering. The clustering of the covariates is achieved using the probability of each β_j to be drawn from the same component of the mixture, given the data and the estimated parameters.

The present paper is organized as follows. Section 2 presents our model. In Section 3, a maximum likelihood strategy is presented to estimate the model parameters as well as a criterion to select the number of latent groups. Section 4 presents numerical experiments both on simulated and real data. In this section the predictive performances of our model are compared to standard approaches for dimension reduction in high dimensional linear regression models. The perspectives of this research are discussed in Section 5.

2. Model definition and notation

2.1. Model definition

As aforementioned, the number of predictors may be very large with respect to the number of samples. It is therefore impossible to uniquely estimate each coefficient β_j. However, we may assume the existence of g latent groups of covariates within which the β_j's are sufficiently close to one another that all of them may be summarized by their average. Among possible mathematical translations of the latter assumption, we propose to consider the β_j's no longer as fixed effect parameters but as unobserved independent random variables following a Gaussian mixture distribution:

$$\beta_j \sim \sum_{k=1}^{g} \pi_k \mathcal{N}\left(b_k, \gamma^2\right). \tag{2}$$

In other words, we assume for each β_j the existence of a multinomial distributed random variable, $\mathbf{z} = (z_{j1}, \ldots, z_{jg})$ of parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)'$, such as β_j is drawn from the k-th component of

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

the mixture when $z_{jk} = 1$. Our model can then be written as

$$\begin{cases} y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \\ \beta_j | \mathbf{z}_j \sim \mathcal{N}\left(\sum_{k=1}^{g} b_k z_{jk}, \gamma^2\right) \\ \mathbf{z}_j = (z_{j1}, \ldots, z_{jg}) \sim \mathcal{M}(\pi_1, \ldots, \pi_g). \end{cases} \tag{3}$$

Parameter β_0 is associated with a constant variable. Since our primary aim is variable clustering, we did not considered β_0 as random in model (3).

2.2. Notation

In all subsequent sections of the paper the following notation hold: $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\mathbf{X} = (x_{ij})$, $\mathbf{Z} = (z_{jk})$, $\mathbf{b} = (b_1 \ldots b_g)'$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)'$.Moreover, $\log p(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta})$ denotes the log-likelihood of model (3) assessed for the parameter $\boldsymbol{\theta} = \left(\beta_0, \boldsymbol{b}, \boldsymbol{\pi}, \sigma^2, \gamma^2\right)$ and $\mathscr{Z}$ the set of $p \times g$-matrices defined as

$$\left(z_{jk}\right)_{1 \leq j \leq p, 1 \leq k \leq g} \in \mathscr{Z} \implies \forall j \in \{1, \ldots, p\}, \begin{cases} \exists! k \text{ such as } z_{jk} = 1 \\ \text{if } k' \neq k \text{ then } z_{jk} = 0. \end{cases}$$

2.3. Bayes or Empirical Bayes?

With such a hierarchical definition, model (3) can be interpretated as a Bayesian approach. However, to be fully Bayesian a prior distribution for $\boldsymbol{\theta} = \left(\beta_0, \boldsymbol{b}, \boldsymbol{\pi}, \sigma^2, \gamma^2\right)$ would have been necessary. Instead, we propose to estimate $\boldsymbol{\theta}$ by maximizing the (marginal) log-likelihood, $\log p(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta})$. This partially Bayesian approach is referred to as *Empirical Bayes* (EB) [7]. Our choice for an EB approach was motivated by the number of parameters we have to estimate. This number, $2(g+1)$, is often small with respect to the sample size n. In this situation, posterior distributions obtained with an EB approach and with a fully Bayesian approach are expected to be close [19].

2.4. Degeneracy of the likelihood

To prevent degeneracy of the likelihood, which often occurs in mixture models [2], constraints are generally imposed to the space of hidden variables [9]. In this work the following constraint is proposed:

$$\forall k = 1, \ldots, g \ \sum_{j=1}^{p} z_{jk} \geq 1. \tag{4}$$

This constraint basically requires none of the groups to be empty.

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

3. Estimation, prediction, clustering and model selection

3.1. Maximum Likelihood Estimation

The log-likelihood $\log p(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})$ is defined as

$$\log p(\mathbf{y}|\mathbf{X};\boldsymbol{\theta}) = \log\left[\sum_{\mathbf{Z}\in\mathscr{Z}}\int_{\mathbb{R}^p} p(\mathbf{y},\boldsymbol{\beta},\mathbf{Z}|\mathbf{X};\boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}\right]. \tag{5}$$

The likelihood cannot be calculated analytically as it involves integration over unobserved data $(\boldsymbol{\beta},\mathbf{Z})$. A direct maximization for estimating $\boldsymbol{\theta}$ is consequently impossible.

The expectation-maximization (EM) algorithm [17] has been introduced to perform MLE in the presence of unobserved data. The EM algorithm is an iterative method, which starts with initial estimates of the parameters and updates these estimates at each iteration until convergence is achieved. We propose in the following subsections its implementation in the special case of model (3).

3.1.1. Initialization

The algorithm is initialized using primary estimates ${\beta_j}^{(0)}$ of each β_j. The latter can be either obtained from univariate regression coefficients or from penalized approaches like the LASSO or the ridge regression. Model (2) is then fitted using $\beta^{(0)} = \left(\beta_1^{(0)},\ldots,\beta_p^{(0)}\right)$ as observed data to produce starting values $\mathbf{b}^{(0)}$, $\boldsymbol{\pi}^{(0)}$ and ${\gamma^2}^{(0)}$ respectively for parameters $\mathbf{b}$, $\boldsymbol{\pi}$ and γ^2. An initial partition $\mathbf{Z}^{(0)} = \left(z_{jk}^{(0)}\right) \in \mathscr{Z}$ is determined as

$$\forall j \in \{1,\ldots,p\},\ z_{jk}^{(0)} = \begin{cases} 1 & \text{if } k = \operatorname{argmin}_{k'\in\{1,\ldots,g\}}\left({\beta_j}^{(0)} - b_{k'}^{(0)}\right)^2 \\ 0 & \text{otherwise.} \end{cases}$$

β_0 and σ^2 are initialized using $\beta^{(0)}$ as following:

$$\beta_0^{(0)} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j^{(0)}x_{ij}\right) \text{ and } {\sigma^2}^{(0)} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta_0^{(0)} - \sum_{j=1}^{p}\beta_j^{(0)}x_{ij}\right)^2.$$

The EM algorithm only ensures to converge towards a local maximum of the likelihood. Our approach is therefore potentially subjected to this limitation. Nevertheless, the stochasticity introduced during the *E*-step (see Section 3.1.2) tends to lessen the impact of the starting point. This has already been studied in a general context [8]. Indeed, we illustrate further in Section 4.1.2 that the choice of the starting point is not critical to our method.

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

3.1.2. (Stochastic) Expectation step

At iteration d of the algorithm, the log-likelihood of the complete data $\log p(\mathbf{y},\boldsymbol{\beta},\mathbf{Z}|\mathbf{X};\boldsymbol{\theta}^{(d)})$ has the following expression:

$$\begin{aligned}\log p(\mathbf{y},\boldsymbol{\beta},\mathbf{Z}|\mathbf{X};\boldsymbol{\theta}^{(d)}) &= \log p(\mathbf{y}|\boldsymbol{\beta},\mathbf{X};\beta_0^{(d)},\sigma^{2(d)}) + \log p(\boldsymbol{\beta},\mathbf{Z}|\mathbf{X};\mathbf{b}^{(d)},\boldsymbol{\pi}^{(d)},\gamma^{2(d)}) \\ &= -\frac{n}{2}\log\left(2\pi\sigma^{2(d)}\right) - \frac{1}{2\sigma^{2(d)}}\sum_{i=1}^{n}\left(y_i - \beta_0^{(d)} - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \\ &\quad -\frac{p}{2}\log\left(2\pi\gamma^{2(d)}\right) + \sum_{j=1}^{p}\sum_{k=1}^{g} z_{jk}\left(\log\pi_k^{(d)} - \frac{\left(\beta_j - b_k^{(d)}\right)^2}{2\gamma^{2(d)}}\right).\end{aligned}$$

In classical EM algorithm, the E-step requires, at each iteration, the calculation of the expectation of the log-likelihood of the full data $\log p\left(\mathbf{y},\boldsymbol{\beta},\mathbf{Z}|\mathbf{X};\boldsymbol{\theta}^{(d)}\right)$, with respect to the conditional distribution of unobserved data given observed data. This quantity generally denoted as $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(d)}\right)$, does not have a closed form in model (3). We therefore approximate $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(d)}\right)$ using Monte Carlo simulations. This stochastic version of the EM algorithm was introduced in [11] under the name of Monte Carlo EM (MCEM) algorithm. A Gibbs sampling scheme is proposed to generate draws from the probability distribution $p\left(\boldsymbol{\beta},\mathbf{Z}|\mathbf{y},\mathbf{X};\boldsymbol{\theta}^{(d)}\right)$. In model (3), Gibbs sampling requires the definition of the conditional distributions $p\left(\boldsymbol{\beta}|\mathbf{Z},\mathbf{y},\mathbf{X};\boldsymbol{\theta}^{(d)}\right)$ and $p\left(\mathbf{Z}|\boldsymbol{\beta},\mathbf{y},\mathbf{X};\boldsymbol{\theta}^{(d)}\right)$. The latter distributions are given in Equations (6) and (7). Details about how those distributions were derived are given in Section 6.

$$\left\{\begin{array}{l}\boldsymbol{\beta}|\mathbf{Z},\mathbf{y};\boldsymbol{\theta}^{(d)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(d)},\boldsymbol{\Sigma}^{(d)}\right) \\ \boldsymbol{\mu}^{(d)} = \left[\mathbf{X}'\mathbf{X} + \frac{\sigma^{2(d)}}{\gamma^{2(d)}}\mathrm{I}_p\right]^{-1}\mathbf{X}'\left(\mathbf{y} - \beta_0^{(d)}\mathbb{I}_n\right) + \frac{\sigma^{2(d)}}{\gamma^{2(d)}}\left[\mathbf{X}'\mathbf{X} + \frac{\sigma^{2(d)}}{\gamma^{2(d)}}\mathrm{I}_p\right]^{-1}\mathbf{Z}\mathbf{b}^{(d)} \\ \boldsymbol{\Sigma}^{(d)} = \sigma^{2(d)}\left[\mathbf{X}'\mathbf{X} + \frac{\sigma^{2(d)}}{\gamma^{2(d)}}\mathrm{I}_p\right]^{-1}\end{array}\right. \tag{6}$$

and

$$p\left(z_{jk} = 1|\boldsymbol{\beta};\boldsymbol{\theta}^{(d)}\right) \propto \pi_k^{(d)}\exp\left(-\frac{\left(\beta_j - b_k^{(d)}\right)^2}{2\gamma^{2(d)}}\right). \tag{7}$$

Now suppose we have sampled $\left\{\left(\boldsymbol{\beta}^{(1,d)},\mathbf{Z}^{(1,d)}\right),\ldots,\left(\boldsymbol{\beta}^{(M_d,d)},\mathbf{Z}^{(M_d,d)}\right)\right\}$ from $p\left(\boldsymbol{\beta},\mathbf{Z}|\mathbf{y},\mathbf{X};\boldsymbol{\theta}^{(d)}\right)$ and verifying the condition (4); the approximated E-step can then be written as follows:

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(d)}\right) = \mathbb{E}\left[\log p(\mathbf{y},\boldsymbol{\beta},\mathbf{Z}|\mathbf{X};\boldsymbol{\theta}^{(d)})|\mathbf{y},\mathbf{X};\boldsymbol{\theta}^{(d)}\right] \approx \frac{1}{M_d}\sum_{m=1}^{M_d}\log p(\mathbf{y},\boldsymbol{\beta}^{(m,d)},\mathbf{Z}^{(m,d)}|\mathbf{X};\boldsymbol{\theta}^{(d)}). \tag{8}$$

The computational time and the convergence of the algorithm is governed by the choice of M_d. In [11], the authors suggested using small values for M_d (around 20) when starting the algorithm and increases this value along with the number of iterations. In this paper however M_d was set to a constant large value.

3.1.3. Maximization step

The M-step consists of maximizing $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(d)}\right)$ with respect to $\boldsymbol{\theta}$. We get the following update equations:

$$\pi_k^{(d+1)} = \frac{1}{M_d p} \sum_{m=1}^{M_d} \sum_{j=1}^{p} z_{jk}^{(m,d)}, \tag{9}$$

$$b_k^{(d+1)} = \frac{1}{M_d p \pi_k^{(d+1)}} \sum_{m=1}^{M_d} \sum_{j=1}^{p} z_{jk}^{(m,d)} \beta_j^{(m,d)}, \tag{10}$$

$$\gamma^{2(d+1)} = \frac{1}{M_d p} \sum_{m=1}^{M_d} \sum_{j=1}^{p} \sum_{k=1}^{g} z_{jk}^{(m,d)} \left(\beta_j^{(m,d)} - b_k^{(d+1)}\right)^2, \tag{11}$$

$$\beta_0^{(d+1)} = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{p} \left(\frac{1}{M_d} \sum_{m=1}^{M_d} \beta_j^{(m,d)} \right) x_{ij} \right], \tag{12}$$

$$\sigma^{2(d+1)} = \frac{1}{n M_d} \sum_{m=1}^{M_d} \sum_{i=1}^{n} \left(y_i - \beta_0^{(d+1)} - \sum_{j=1}^{p} \beta_j^{(m,d)} x_{ij} \right)^2. \tag{13}$$

3.2. Prediction and Clustering

If $\mathbf{X}^v$ denotes a new design matrix for which we want to predict the response $\mathbf{y}^v$, then we can define the predicted response $\widehat{\mathbf{y}}$ as

$$\widehat{\mathbf{y}} = \mathbf{X}^v \mathbb{E}\left[\boldsymbol{\beta}|\mathbf{y},\mathbf{X};\widehat{\boldsymbol{\theta}}\right], \tag{14}$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$. The clustering of the covariates is achieved using the probability of each β_j to be drawn from the same component of the mixture, given the data and the estimated parameters. Therefore the j-th covariate is assigned to the k-th cluster if

$$\forall l = 1,\ldots,g, \mathbb{E}\left[z_{jk}|\mathbf{y},\mathbf{X};\widehat{\boldsymbol{\theta}}\right] \geq \mathbb{E}\left[z_{jl}|\mathbf{y},\mathbf{X};\widehat{\boldsymbol{\theta}}\right].$$

3.3. Model selection

Model (3) depends on a tuning parameter g, which is the assumed number of groups of covariates. In few situations, this number can be chosen *a priori*, however in a more general setting a strategy should be proposed to make this choice. The BIC is proposed as a means to select g. This criterion

was preferred to other criteria based on estimates of the out-of-sample prediction error like cross-validation (CV) because of its low computational cost. In model (3) the number of parameters equals $2(g+1)$. The BIC has therefore the following expression:

$$BIC = -2\log p(\mathbf{y}|\mathbf{X};\widehat{\boldsymbol{\theta}}) + 2(g+1)\log(n). \tag{15}$$

As the calculation of the likelihood is still intractable, we can derive from Equation (5), an approximation of the BIC criterion using Monte Carlo simulations.

4. Numerical experiments

In this section, we compare our approach CLERE with standard dimension reduction approaches in terms of prediction error. The methods selected for comparison are the variable selection using LARS algorithm [1], the ridge regression [6], the elastic net [14], the LASSO [18], PACS [5], the method of Park and colleagues [24] (subsequently denoted AVG) and the spike and slab model [13] (subsequently denoted SS). The first four methods are implemented in freely available R packages `lars` and `glmnet` (for ridge, LASSO and elastic net). Those packages were used with default options. For PACS a R script was released on Bondell's webpage[1]. This R script was however running very slowly. We therefore decided to reimplement it in C++. This led to a 30-fold speed-up in the computational time. Similarly to Bondell's script, our program uses two parameters named `lambda` and `betawt`. In [5], the authors suggest assigning `betawt` with the coefficients obtained from a ridge regression model after the tuning parameter was selected using AIC. In this simulation study we used the same strategy; however the ridge parameter was selected via 5-fold cross validation. 5-fold CV was preferred to AIC since selecting the ridge parameter using AIC always led to estimated coefficients equal to zero. Once `betawt` was selected, `lambda` was chosen via 5-fold cross validation among the following values: 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100, 200 and 500. All other default parameters of their script were unchanged. For the AVG method, we followed the algorithm described in [24] and implemented it in R. We used the R package `spikeslab` to run the spike and slab models. Especially, we used the function `spikeslab` from that package to detect influential variables. The number of iterations used to run the function `spikeslab` was 2000 (1000 discarded). When running CLERE, the number of EM iterations as well as the number M_d of Monte Carlo samples was set to 1000. The number of groups for CLERE was chosen between 1 and 9. In all experiments, CLERE was initialized using the estimated univariate regression coefficients as explained in Section 3.1.1. Our C++ implementations of PACS and CLERE are available on request.

4.1. Simulated data

4.1.1. Description

The simulated data are presented under three scenarios. For each scenario, 100 training data sets were simulated from the standard linear regression model (1). All training data sets consist of $n = 50$ simulated individuals with $p = 100$ variables. In each scenario, a validation set consisting

[1] `http://www4.stat.ncsu.edu/~bondell/Software/PACS/PACS.R.r`

of 5000 individuals was used to calculate the scaled mean squared prediction error.

If $(\mathbf{y}^t, \mathbf{X}^t)$ and $(\mathbf{y}^v, \mathbf{X}^v)$ are respectively the training and validation data sets, then the scaled mean squared prediction error MSE is calculated as:

$$\text{MSE} = \frac{\|\mathbf{y}^v - \widehat{\mathbf{y}}(\mathbf{X}^v, \mathbf{y}^t, \mathbf{X}^t)\|_2}{\|\mathbf{y}^v\|_2}, \tag{16}$$

where $\widehat{\mathbf{y}}(\mathbf{X}^v, \mathbf{y}^t, \mathbf{X}^t)$ is the predicted response and $\|.\|_2$ stands for the L^2 norm. For CLERE, predictions are obtained using Equation (14). Each of the methods selected for comparison provides a fitted value $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$. A predicted response under the design $\mathbf{X}^v$ is then calculated as $\mathbf{X}^v\widehat{\boldsymbol{\beta}}$. In all simulations, design matrices $\mathbf{X}^t$ and $\mathbf{X}^v$ were simulated as independently normally distributed:

$$\mathbf{x}_i \sim \mathscr{N}(\mathbf{0}, \mathbf{R}), \tag{17}$$

where $\mathbf{R} = \left(r_{jj'}\right)$ is a $p \times p$ matrix defined by $r_{jj'} = 0.5^{|j-j'|}$. In all scenarios, parameters β_0 and σ^2 equal respectively 0 and 100.

The three scenarios are presented below.

1. In scenario 1, the vector $\boldsymbol{\beta}$ of regression coefficients is given by:

$$\boldsymbol{\beta} = (\underbrace{0,\ldots,0}_{36},\underbrace{1,\ldots,1}_{28},\underbrace{3,\ldots,3}_{20},\underbrace{7,\ldots,7}_{12},\underbrace{15,\ldots,15}_{4})'.$$

2. In scenario 2, the vector $\boldsymbol{\beta}$ of regression coefficients is given by:

$$\boldsymbol{\beta} = (\underbrace{0,\ldots,0}_{36},\underbrace{4,\ldots,4}_{28},\underbrace{24,\ldots,24}_{20},\underbrace{124,\ldots,124}_{12},\underbrace{624,\ldots,624}_{4})'.$$

3. In scenario 3, the regression coefficients are chosen uniformally between -10 and +10 :

$$\forall j,\ \beta_j = -10 + (j-1) \times \frac{20}{99}.$$

Scenarios 1 and 2 were chosen to favor variable selection approaches. In those scenarios indeed 36 out of 100 covariates do not influence the response. Moreover the number of effective variables decreases with their effect size. Scenario 3 was proposed to illustrate the relative predictive performances of CLERE under the assumption that almost all covariates contribute to the response. We also considered three additional scenarios directly deriving from the previous ones. Those scenarios are further denoted as *alternative* scenario 1, 2 and 3. The *alternative* scenario s ($s \in \{1,2,3\}$) is obtained by randomly permuting the regression coefficients in scenario s. These additional scenarios were proposed to explore the performances of our methodology when correlated variables do not necesarily have equal or similar regression coefficients.

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

4.1.2. Impact of the initialization strategy

We consider in this subsection four initialization schemes based on four initial guesses for the unobserved regression coefficients. In addition to univariate regression, LASSO and ridge regression already mentioned in Section 3.1.1, we also added the elastic net as one possible means to generate initial estimates for the β_j's. We compared the distribution of the maximum likelihood reached and the distribution of prediction error (MSE) for the four initialization strategies using 100 data sets simulated according to scenario 1. As a reference, we also considered the case where the initial guesses were actually the true regression coefficients used to generate the data. Figure 1

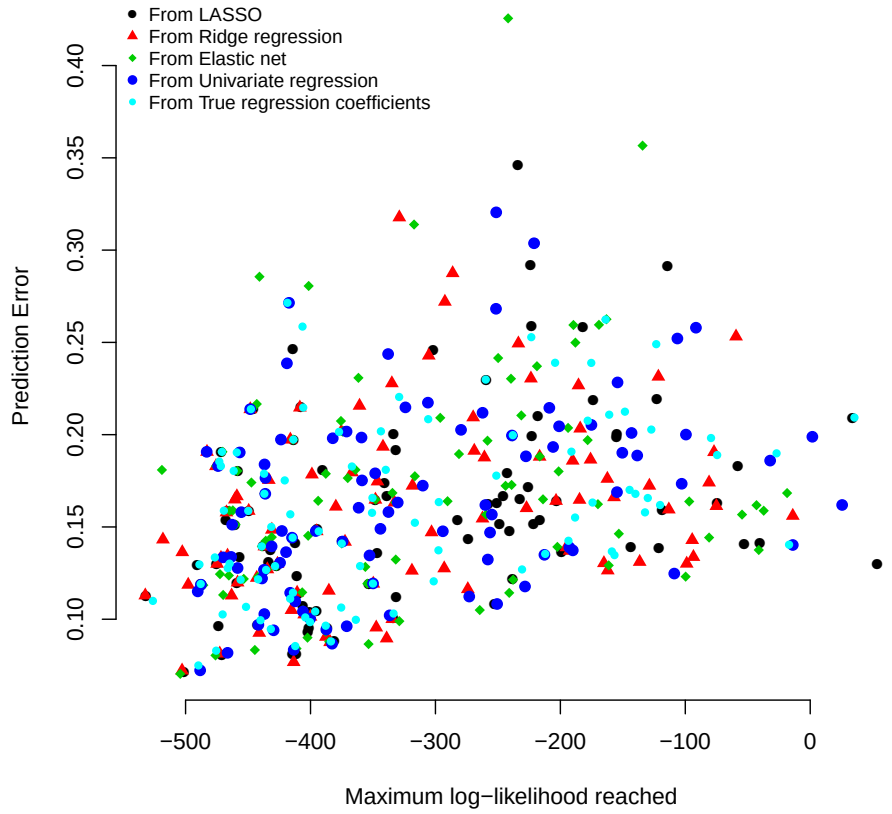

FIGURE 1. *Co-distribution of the MSE and the maximum log-likelihood reached for all the initialization strategies. No significant difference is noticeable in the maximum likelihood reached (F-value = 0.189 - P-value = 0.944) nor in the prediction error (F-value = 0.57 - P = 0.684).*

illustrates the results of that comparison. No significant difference was noticeable between the four initialization strategies. Indeed, none of them seemed to systematically lead to lower or higher likelihood. This is supported by the very large p-value (P = 0.944) obtained after performing

a Fisher's test to test for a potential difference between the four strategies. We can therefore argue that initialization is not a critical issue for our method. Similar results have already been observed for stochastic versions of the EM algorithm in [8] from which our implementation partially derives. We also compared the distribution of the MSE for each of the initialization strategies. No difference in terms of MSE was noticeable (P = 0.684).

4.1.3. Results

Table 1 summarizes the MSE calculated under each scenario. Using this measure, CLERE has the best average rank over all scenarios considered. We also considered a measure of model complexity being either the number of unique non-zero parameters or simply the number of parameters for CLERE. Using the latter measure, the present simulation study illustrates that CLERE selects the simplest model in all the scenarios considered.

	Without permutation			With permutation		
	100× averaged MSE (Std. Err)	Averaged number of parameters (Std. Dev)	MSE Rank	100× averaged MSE (Std. Err)	Averaged number of parameters (Std. Dev)	MSE Rank
Scenario 1						
LARS	51.1 (1.7)	49 (0)	9	74.1 (2.7)	49 (0)	8
LASSO	15.9 (0.4)	42.4 (4.8)	5	32.1 (0.88)	42.8 (7.3)	4
Ridge	59.7 (0.48)	100 (0)	7	61.6 (0.55)	100 (0)	6
Elastic net	14.3 (0.33)	48.8 (7.1)	3	29 (0.66)	52.3 (9.2)	3
CLERE (g = 5)	15.2 (0.48)	12 (0)	4	**25.3 (0.86)**	12 (0)	**1**
CLERE	16.7 (0.51)	**9.16 (4.8)**	6	25.9 (0.78)	**8.52 (4.8)**	2
AVG	**8.4 (0.33)**	31.5 (6.5)	**1**	36.1 (1.4)	38.2 (8.7)	5
PACS	10.4 (0.28)	35.5 (8.7)	2	74.3 (2.2)	34.1 (15)	9
SS	70.8 (0.7)	87.5 (5.7)	8	73.8 (0.62)	89 (5.5)	7
Scenario 2						
LARS	8.86 (0.68)	49 (0)	7	10.9 (0.56)	49 (0)	6
LASSO	1.15 (0.05)	33.3 (3)	5	3.82 (0.2)	40.5 (3)	3
Ridge	66.4 (0.4)	100 (0)	8	68.6 (0.44)	100 (0)	8
Elastic net	1.23 (0.058)	33.8 (3.1)	6	4.14 (0.22)	40.9 (3)	5
CLERE (g = 5)	0.023 (0.005)	12 (0)	2	0.26 (0.09)	12 (0)	2
CLERE	**0.014 (0.003)**	**15.4 (3)**	**1**	**0.14 (0.07)**	**14.4 (2.9)**	**1**
AVG	0.62 (0.057)	28 (5.4)	3	4.02 (0.19)	40.4 (2.9)	4
PACS	0.817 (0.075)	44.2 (8.6)	4	43.3 (2.9)	41.2 (8.5)	7
SS	98.1 (0.42)	85.4 (8.1)	9	98.9 (0.05)	85.4 (6.8)	9
Scenario 3						
LARS	74.8 (1.9)	49 (0)	8	139 (3.2)	49 (0)	9
LASSO	35.5 (0.75)	46 (6.5)	5	76.4 (1.2)	32.7 (14)	6
Ridge	53 (0.62)	100 (0)	6	73.8 (0.62)	100 (0)	4
Elastic net	24.3 (0.6)	65 (7.7)	4	**61.2 (1.2)**	53.8 (14)	**1**
CLERE (g = 5)	19.4 (0.86)	12 (0)	2	64.3 (2)	12 (0)	2
CLERE	23.8 (1.1)	**9.7 (6.1)**	3	64.7 (2.2)	**8.8 (5.9)**	3
AVG	**8.38 (0.54)**	33.2 (5.9)	**1**	76.1 (1.6)	35.6 (9.5)	5
PACS	70.5 (1.8)	28 (17)	7	94.2 (1.6)	29.7 (18)	8
SS	81.6 (0.47)	89.9 (3.7)	9	86.4 (0.45)	83 (9.5)	7

TABLE 1. *Averaged MSE for simulated data under the three scenarios. The average number of non-zero parameters estimated for each method was also reported. When not specified, the number of groups g is chosen using BIC criterion. For each scenario in the table we highlighted in bold font the lowest prediction error (equivalent to best MSE rank) or the lower number of parameters.*

In scenario 1 and 2, clusters of covariates were simulated. These clusters correspond to

covariates having equal regression coefficients. The predictive performances of all the methods were influenced by the separation between the clusters. Indeed, all methods increased their performances along with the clusters separation. This improvement was however much more noticeable for CLERE which outperformed its competitors in scenario 2.

The predictive performances of the methods were also influenced by the correlations between the covariates. This is illustrated by comparing each scenario with its *alternative* counterpart. CLERE robustly showed good performances in all *alternative* scenarios. Especially, it yields the best predictive performance in *alternative* scenarios 1 and 2. In Scenario 3, the regression coefficients were not separated at all. However, CLERE managed to yield competitive performances both under the initial and the *alternative* scenarios.

We also report for CLERE the distribution over 100 simulated data sets of the estimated b_k's. This is shown in Figure 2 under scenarios 1, 2 and their *alternative* counterparts. Estimates of the b_k's are known up to a permutation. It was therefore not straightforward to compare estimates from a data set to another. To achieve a global comparison of the estimates across all simulated data sets, we selected, for each data set, the permutation that minimized the bias.

4.2. Real data

4.2.1. Description

We used in this section the real data set *mice* from the `spls` R package. This data set consists of $n = 60$ mice for which the expression of 83 gene transcripts from liver tissues was measured and $p = 145$ microsatellite markers were genotyped. For more details about this data set please refer to [12]. One challenging issue of modern Genetics is to bridge gene expression levels with variations in the genomic sequence. Microsatellite markers are such variations. The latter markers are discrete quantitative variables taking values in $\{1,2,3\}$, while gene expression levels are real quantitative variables. Instead of considering each transcript as a response, we performed a principal component analysis (PCA) over the gene expression data to come up with a reduced number of outcomes. This PCA did not involve the microsatellite markers. The PCA was performed using the function `dudi.pca` implemented in the R package `ade4`. The first five principal components (PC) accounted for more than 92% of the total inertia. We then proposed a linear regression model for each of those selected PCs using the microsatellites markers as covariates. The selected PCs are subsequently denoted $PC1,\ldots,PC5$. Since no proper validation data sets were available, all methods were compared in terms of out-of-sample prediction error estimated via 5-fold cross-validation (CV).

4.2.2. Overall results

Table 2 summarizes the MSE for each selected PC and each method. Similarly to numerical experiments on simulated data, variable selection using the LARS algorithm yielded very large prediction error for each PC. All other methods had however comparable prediction error. Using the averaged rank as an indicator of overall performance, CLERE was the second best method. The first place was shared by ridge regression, PACS and the Spike and Slab method. CLERE showed the best predictive performances for PC4 and PC5 (no other method was best twice) and was among the most parsimonious methods with the Spike and Slab method and PACS.

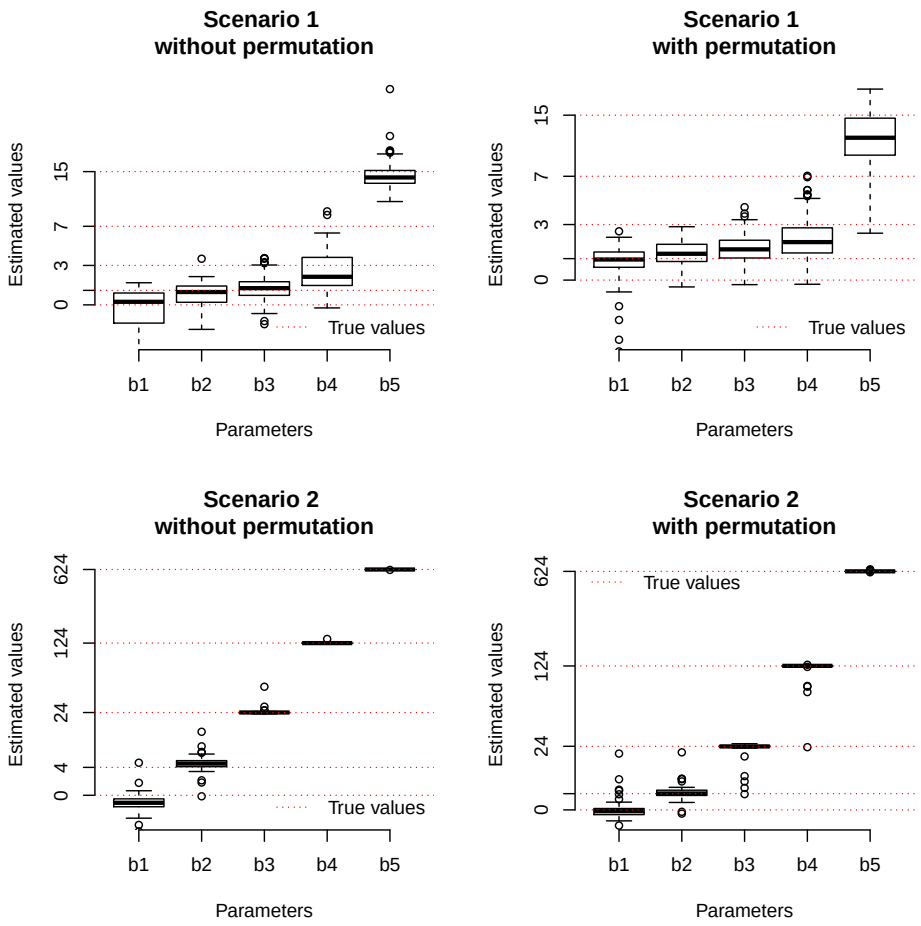

FIGURE 2. *Distributions of the maximum likelihood estimates for parameter* **b** *over 100 simulated data sets under scenarios 1 and 2.*

*4.2.3. Focus on PC*1

We have illustrated above that CLERE is a competitive method for prediction. In this sub-section we now present how CLERE can be used for interpretation purpose. A focus is therefore laid on *PC*1 as a single response variable. The data were no longer partitioned as previously did for cross-validation.

Using the whole data set, 3 groups were chosen using the BIC criterion (see Figure 3).

The estimated parameters are given in Table 3. Two groups with moderated positive effects and one group with strong negative effect were identified. In Section 3.2, we presented how to make predictions with CLERE using the vector $\mathbb{E}\left[\boldsymbol{\beta}|\mathbf{y},\mathbf{X};\widehat{\boldsymbol{\theta}}\right]$. The latter vector of expectations

		averaged 5-fold CV statistic (Std. Err)	Averaged number of parameters (Std. Dev.)	MSE Rank
*PC*1	LARS	293.57 (125.43)	47 (0)	8
	LASSO	1.26 (0.15)	6.8 (10.55)	4
	Ridge	1.07 (0.04)	145 (0)	3
	Elastic net	2.13 (0.81)	18.6 (21.3)	7
	CLERE	1.31 (0.1)	4 (0)	6
	AVG	1.29 (0.05)	11.4 (9.24)	5
	PACS	**1.03 (0.01)**	0.2 (0.45)	**1**
	SS	1.04 (0.01)	0.2 (0.45)	2
*PC*2	LARS	31.72 (10.76)	47 (0)	8
	LASSO	**0.95 (0.04)**	10.2 (4.71)	**1**
	Ridge	0.98 (0.08)	145 (0)	2
	Elastic net	1.14 (0.13)	43.2 (28.14)	5
	CLERE	1.17 (0.2)	4 (0)	6
	AVG	1.28 (0.14)	21.8 (3.11)	7
	PACS	1.03 (0.05)	5.2 (3.27)	4
	SS	0.99 (0.03)	1 (1.22)	3
*PC*3	LARS	18.74 (3.67)	47 (0)	8
	LASSO	1.66 (0.69)	22.4 (16.61)	6
	Ridge	**0.96 (0.14)**	145 (0)	**1**
	Elastic net	1.68 (0.68)	29.6 (26.49)	7
	CLERE	1.06 (0.2)	4 (0)	3
	AVG	1.61 (0.71)	21.6 (15.19)	5
	PACS	1.17 (0.11)	4.8 (5.07)	4
	SS	1.04 (0.14)	3 (2.24)	2
*PC*4	LARS	30.97 (6.97)	47 (0)	8
	LASSO	1.29 (0.11)	3.8 (4.32)	5
	Ridge	1.16 (0.03)	145 (0)	4
	Elastic net	1.38 (0.1)	12.2 (12.76)	7
	CLERE	**1.09 (0.03)**	4 (0)	**1**
	AVG	1.35 (0.1)	7.2 (4.66)	6
	PACS	1.13 (0.05)	1.6 (1.82)	3
	SS	1.11 (0.04)	0.6 (0.89)	2
*PC*5	LARS	17.26 (8.03)	47 (0)	8
	LASSO	1.07 (0.04)	0 (0)	3
	Ridge	1.07 (0.04)	145 (0)	3
	Elastic net	1.52 (0.49)	10 (21.81)	7
	CLERE	**1.05 (0.002)**	4 (0)	**1**
	AVG	1.16 (0.07)	4.4 (4.1)	6
	PACS	1.07 (0.04)	1.2 (0.45)	3
	SS	1.09 (0.05)	0 (0)	5

TABLE 2. *Out-of-sample prediction error estimated using 5-fold CV for each method and each PC for mice data from [12]. The averaged number of fitted parameters, as a measure of model complexity, is also reported. For each scenario in the table we highlighted in bold font the lowest prediction error (equivalent to best MSE rank) or the lower number of parameters.*

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

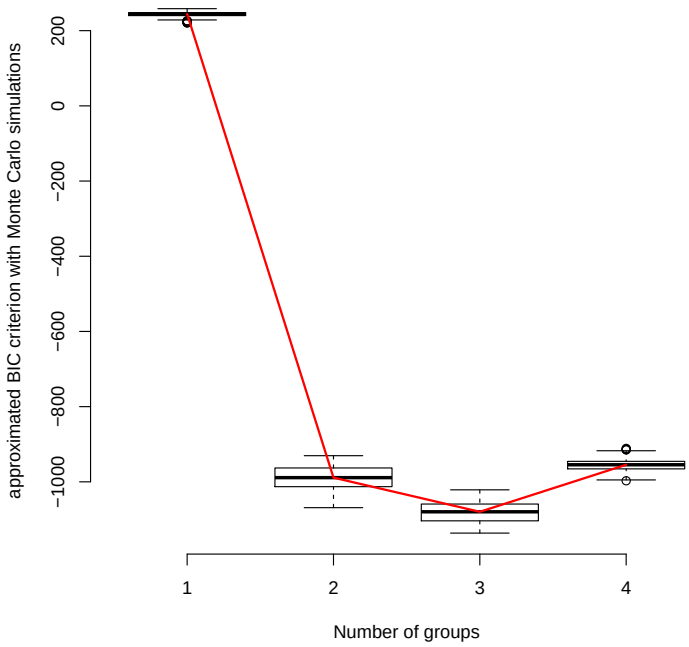

FIGURE 3. *Selection procedure for the number of groups. Here $g = 3$ was selected as it minimizes the BIC. The BIC is approximated using Monte Carlo simulations.*

$\widehat{\beta_0}$	$\widehat{b_1}$	$\widehat{b_2}$	$\widehat{b_3}$	$\widehat{\pi_1}$	$\widehat{\pi_2}$	$\widehat{\pi_3}$	$\widehat{\gamma^2}$	$\widehat{\sigma^2}$
2.32×10^{-2}	7.87×10^{-2}	-9.32×10^{-1}	7.63×10^{-2}	0.870	0.076	0.054	3.0×10^{-6}	7.35

TABLE 3. *Maximum likelihood estimate obtained for CLERE when fitting mice data using PC*1 *as response variable.*

can be interpreted as a vector of regression coefficients. Consequently, the small estimated value for parameter γ^2 ($\widehat{\gamma^2} = 3.0 \times 10^{-6}$) leads those expectations to be strongly concentrated around the $\widehat{b_k}$'s. CLERE yielded thus a very parsimonious regression model.

The second group, associated with $\widehat{b_2} = -0.931$, was of interest since it gathers the 11 variables showing the strongest impact on the response according to CLERE. In Table 4, we compared for those variables the regression coefficients obtained with LARS, LASSO, ridge regression, elastic net, AVG, PACS and SS. The five methods yielded sign and size consistent regression coefficients for almost all the markers highlighted in Table 4. One exception was however noticed for D13Mit16. In addition, CLERE showed that some variables discarded by other methods may still be of interest. Overall this analysis emphasized the ability of CLERE to consistently identify influential covariates using a very parsimonious model. Moreover, this analysis identifies the clusters of markers that may be relevantly investigated for a biological characterization.

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

Markers	Chromosome	Lars	LASSO	Ridge	Elastic net	CLERE	AVG	PACS	SS
D1Mit87	1	.	.	-0.0265	.	-0.9318	-0.0717	.	.
D3Mit19	3	-0.2347	-0.8962	-0.1940	-0.5670	-0.9316	-0.4253	.	-0.1219
D4Mit149	4	.	.	-0.0855	.	-0.9316	-0.0717	.	-0.2788
D4Mit237	4	-2.7478	-0.8661	-0.1714	-0.4767	-0.9318	-0.0717	.	.
D7Mit56	7	-0.2011	-0.0484	-0.1026	-0.1516	-0.9318	.	.	.
D7Mit76	7	.	-0.0116	-0.1026	-0.1514	-0.9317	.	.	.
D8Mit42	8	0.0119	.	-0.0430	.	-0.9319	.	.	.
D9Mit15	9	-3.1530	-1.6102	-0.2826	-1.0474	-0.9318	-0.4253	.	-0.1887
D13Mit16	13	1.2867	.	0.0530	0.0823	-0.9318	-0.0034	.	.
D15Mit174	15	-1.7012	-0.9335	-0.1149	-0.4312	-0.9319	-0.4253	.	-0.0253
D19Mit34	19	.	.	-0.0449	-0.0303	-0.9317	.	.	.

TABLE 4. *Microsatellite markers assigned to the cluster associated with parameter b_2. Regression coefficients for those variables are reported for all compared methods. For CLERE regression coefficients are obtained using* $\mathbb{E}\left[\beta|\mathbf{y},\mathbf{X};\widehat{\theta}\right]$. *"." means 0.*

5. Discussion

We proposed in this paper a new method for simultaneous variable clustering and regression. Our approach showed good predictive performances both on simulated and real data compared to its competitors (see Section 4). These good performances were accompanied by a lower complexity in terms of number of fitted parameters. CLERE also brought improvements in terms of interpretability since each fit provides a clustering of the covariates. This work comes in the aftermath of a series of recently published approaches aiming at reducing the dimension in linear regression models by collapsing the covariates into groups. Contrary to those previous works, our approach is not based on penalized least squares problem. However we assumed the existence of a latent structure within the variables that depends only on their unobserved regression coefficients. In such framework, no distributional assumption regarding the covariates is necessary for achieving the clustering. The latent structure is modeled using a Gaussian mixture model whose parameters are estimated via an EM like algorithm. A stochastic version, namely the MCEM, of the latter algorithm was proposed since the E-step was intractable. Even though MCEM has become a standard in many applications, it is noteworthy that its computational cost is not negligible. Indeed, running the estimation with 3 groups on the data set presented in Section 4.2.3 took 30 seconds for CLERE but less than 1 second for the other approaches. Although CLERE seemed to be relatively slow, the estimation time remained manageable. Improvements in speeding up the estimation through parallel computing is a natural perspective of this work, especially since we are aiming at tackling ultra-high dimensional regression problems in forthcoming research. We proposed in this paper the BIC criterion for choosing the number of latent groups. This criterion was preferred over different existing criteria such as the out-of-sample prediction error because of its small computational cost. Other information-based criteria will be explored in further works.

Variable selection is an appealing extension to our model. In fact, if a constraint is imposed on the parameter space, then CLERE can also be used as a variable selection tool. Such constraint may lead for instance to assume one group k to have its mean b_k and its associated variance equal to zero. This would be a new model which however may be easily derived from the approach presented here. Many applications deal with response variable that may not be continuous. Another promising extension of our model is therefore towards generalized linear models.

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

6. Appendix

6.1. Conditional distribution $p(\boldsymbol{\beta}|\mathbf{Z},\mathbf{y},\mathbf{X};\boldsymbol{\theta})$

In Section 3.1.2, a Gibbs sampling strategy is proposed to approximate the E-step of our EM like algorithm based on the conditional distribution $p(\boldsymbol{\beta}|\mathbf{Z},\mathbf{y},\mathbf{X};\boldsymbol{\theta})$ and $p(\mathbf{Z}|\boldsymbol{\beta},\mathbf{y},\mathbf{X};\boldsymbol{\theta})$. We present in this section how we obtained these distributions. Let C denotes the complete log-likehood:

$$C = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \sum_{j=1}^{p}\sum_{k=1}^{g} z_{jk}\left(-\frac{1}{2}\log\left(2\pi\gamma^2\right) - \frac{(\beta_j - b_k)^2}{2\gamma^2} + \log\pi_k\right) \tag{18}$$

$$\begin{aligned} C &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{p}{2}\log\left(2\pi\gamma^2\right) - \frac{1}{2\sigma^2}\left(\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right) - \frac{1}{2\gamma^2}\left(\boldsymbol{\beta}'\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{Z}\mathbf{b} + \mathbf{b}'\mathbf{Z}'\mathbf{Z}\mathbf{b}\right) \\ &\quad + \sum_{j=1}^{p}\sum_{k=1}^{g} z_{jk}\log\pi_k \\ &= K(\mathbf{Z},\sigma^2,\gamma^2,\boldsymbol{\pi}) - \frac{1}{2\sigma^2}\left[\boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{\gamma^2}I\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{y} + \frac{\sigma^2}{\gamma^2}\mathbf{Z}\mathbf{b}\right)\right], \end{aligned}$$

where $K(\mathbf{Z},\sigma^2,\gamma^2,\boldsymbol{\pi})$ is defined as

$$K(\mathbf{Z},\sigma^2,\gamma^2,\boldsymbol{\pi}) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{p}{2}\log\left(2\pi\gamma^2\right) - \frac{1}{2\sigma^2}\mathbf{y}'\mathbf{y} + \sum_{j=1}^{p}\sum_{k=1}^{g} z_{jk}\log\pi_k.$$

Let $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ respectively be defined as

$$\boldsymbol{\Sigma} = \sigma^2\left(\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{\gamma^2}I\right)^{-1}$$

and

$$\boldsymbol{\mu} = \left(\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{\gamma^2}I\right)^{-1}\left(\mathbf{X}'\mathbf{y} + \frac{\sigma^2}{\gamma^2}\mathbf{Z}\mathbf{b}\right).$$

Then

$$\begin{aligned} C &= K(\mathbf{Z},\sigma^2,\gamma^2,\boldsymbol{\pi}) - \frac{1}{2}\left[\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] \\ &= H(\mathbf{Z},\sigma^2,\gamma^2,\boldsymbol{\pi},\mathbf{b}) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\left[(\boldsymbol{\beta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right], \end{aligned}$$

where $H(\mathbf{Z},\sigma^2,\gamma^2,\boldsymbol{\pi},\mathbf{b})$ is defined as

$$H(\mathbf{Z},\sigma^2,\gamma^2,\boldsymbol{\pi},\mathbf{b}) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{p}{2}\log\left(\gamma^2\right) - \frac{1}{2\sigma^2}\mathbf{y}'\mathbf{y} + \sum_{j=1}^{p}\sum_{k=1}^{g} z_{jk}\log\pi_k + \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log(|\boldsymbol{\Sigma}|). \tag{19}$$

We can identify from Equation (19) the density function of a multidimensional normal distribution of parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Therefore since $p(\boldsymbol{\beta}|\mathbf{Z},\mathbf{y},\mathbf{X};\boldsymbol{\theta}) \propto p(\boldsymbol{\beta},\mathbf{Z},\mathbf{y}|\mathbf{X};\boldsymbol{\theta})$ we can derive Equation (6).

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

6.2. Conditional distribution $p(\mathbf{Z}|\boldsymbol{\beta},\mathbf{y},\mathbf{X};\boldsymbol{\theta})$

If we assume that for all $j \in \{1,\ldots,p\}$, $(z_{j1},\ldots,z_{j,g})\,|\boldsymbol{\beta},\mathbf{y},\mathbf{X};\boldsymbol{\theta}$ follows a multinomial distribution, then its associated probabilities can be deduced from Equation (18). Equation (7) derives therefore straightforwardly.

Acknowledgements

The authors are grateful to the editor and the referees for their valuable comments.

References

[1] Efron B., Hastie T., Johnstone I., and Tibshirani R. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[2] Biernacki C. Degeneracy in the Maximum Likelihood Estimation of Univariate Gaussian Mixtures for Grouped Data and Behaviour of the EM Algorithm. *Journal of Scandinavian Statistics*, 34:569–586, 2007.

[3] Stein C. Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statististics*, 9:1135–1151, 1981.

[4] Bondell H. D. and Reich B. J. Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics*, 64:115–123, 2008.

[5] Sharma D.B., Bondell H.D., and Zhang H.H. Consistent Group Identification and Variable Selection in Regression with Correlated Predictors. *Journal of Computational and Graphical Statistics. In Press.*, 2013.

[6] Hoerl A. E. and Kennard W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12:55–67, 1970.

[7] Casella G. An Introduction to Empirical Bayes Data Analysis. *The American Statistician*, 39(2):83–87, 1985.

[8] Celeux G., Chauveau D., and Diebolt J. Some Stochastic versions of the EM Algorithm. *Journal of Statistical Computation and Simulation*, 55:287–314, 1996.

[9] Policello G. Conditional Maximum Likelihood Estimation in Gaussian mixtures. In *Statistical Distributions in Scientific Work*, volume 79 of *NATO Advanced study Institutes Series*, pages 111–125. Springer Netherlands, 1981.

[10] Schwarz G. Estimating the Dimension of a Model. *Annals of Statistics*, 6:461–464, 1978.

[11] Wei C. G. and Tanner M. A. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990.

[12] Chun H. and Keles S. Expression Quantitative Trait Loci Mapping With Multivariate Sparse Partial Least Squares Regression. *Genetics*, 182:79–90, 2009.

[13] Ishwaran H. and Sunil Rao J. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, 33(2):730–773, 2005.

[14] Zou H. and Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[15] Zou H., Hastie T., and Tibshirani R. On the degrees of freedom of the lasso. *Annals of Statistics*, 35(5):2173–2192, October 2007.

[16] Daye Z. J. and Jeng X. J. Shrinkage and model selection with correlated variables via weighted fusion. *Computational Statistics & Data Analysis*, 53(4):1284–1298, February 2009.

[17] Dempster A. P., Laird M. N., and Rubin D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–22, 1977.

[18] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[19] Petrone S., Rousseau J., and Scricciolo C. Bayes and empirical Bayes: do they merge? *arXiv:1204.1470v1*, 2012.

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013

[20] Petry S. and Tutz G. Shrinkage and variable selection by polytopes. *Technical report No. 053, Department of Statistics, University of Munich*, 2009.

[21] Hastie T., Tibshirani R., and Friedman J. H. *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag, 2001.

[22] Mitchell T.J. and Beauchamp J.J. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

[23] Shen X. and Huang H. Grouping pursuit in regression. *Journal of American Statistical Association*, 105:727–739, 2010.

[24] Park M. Y., Hastie T., and Tibshirani R. Averaged gene expressions for regression. *Biostatistics*, pages 212–227, 2007.

[25] She Y. and Stanford University. *Sparse Regression with Exact Clustering*. Stanford University, 2008.

Soumis au Journal de la Société Française de Statistique
File: yengo_13DEC2012_JSFdS_revised.tex, compiled with jsfds, version : 2009/12/09
date: July 24, 2013