Computational Statistics and Data Analysis 80 (2014) 209–222

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm

Florian Rohart [a,b,*], Magali San Cristobal [a,b,c,d], Béatrice Laurent [a]

[a] UMR 5219, Institut de Mathématiques de Toulouse, INSA de Toulouse, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France
[b] INRA, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France
[c] Université de Toulouse INPT ENSAT, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France
[d] Université de Toulouse INPT ENVT, UMR1388 Génétique, Physiologie et Systèmes d'Elevage, F-31076 Toulouse, France

ARTICLE INFO

Article history:
Received 5 August 2013
Received in revised form 23 June 2014
Accepted 26 June 2014
Available online 5 July 2014

Keywords:
Linear mixed model
LmmLasso
Multiple hypothesis testing
High-dimension

ABSTRACT

Linear mixed models are especially useful when observations are grouped. In a high dimensional setting however, selecting the fixed effect coefficients in these models is mandatory as classical tools are not performing well. By considering the random effects as missing values in the linear mixed model framework, a ℓ^1-penalization on the fixed effects coefficients of the resulting log-likelihood is proposed. The optimization problem is solved via a multicycle Expectation Conditional Maximization (ECM) algorithm which allows for the number of parameters p to be larger than the total number of observations n and does not require the inversion of the sample $n \times n$ covariance matrix. The proposed algorithm can be combined with any variable selection method developed for linear models. A variant of the proposed approach replaces the ℓ^1-penalization with a multiple testing procedure for the variable selection aspect and is shown to greatly improve the False Discovery Rate. Both methods are implemented in the MMS R-package, and are shown to give very satisfying results in a high-dimensional simulated setting.

1. Introduction

The more extensive use of new technologies such as high-throughput DNA/RNA chips or RNA sequencing in biology generates an increasing number of highly dimensional data sets where the number of parameters p is much greater than the number of observations n. Consequently, the high dimensional framework generally means that the problem of parameters estimation cannot be solved. In order to address this curse of dimensionality, various constraints have been proposed in linear models. Most of them aim for a parsimonious model where many parameters are set to zero (sparse constraints), or use of a well-conditioned variance matrix on the observations. Many studies have addressed the problem of variable selection by using a linear model of the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is an $n \times p$ matrix containing the observations and $\boldsymbol{\epsilon}$ is a n-vector of i.i.d. random and usually Gaussian variables. One of the oldest methods is the Akaike Information Criterion (AIC), which is a penalization of the log-likelihood by a function of the number of parameters included in the model. More recently, the simple and powerful Lasso (Least Absolute Shrinkage and Selection Operator) method (Tibshirani, 1996) revolutionized the field. The Lasso works by applying a ℓ^1-penalization on the least squares estimate which shrinks some coefficients to exactly zero. Various extensions exist for the Lasso, for example group Lasso (Yuan and Lin, 2007), adaptive Lasso (Huang et al.,

* Correspondence to: Australian Institute for Bioengineering & Nanotechnology (AIBN), The University of Queensland, QLD 4072, Australia.
E-mail addresses: florian.rohart@gmail.com, f.rohart@uq.edu.au (F. Rohart).

http://dx.doi.org/10.1016/j.csda.2014.06.022
0167-9473/

2008) and a more stable version known as Bo-Lasso (Bach, 2009). However, penalizing the likelihood is not the only way to perform variable selection.

Indeed, statistical testing can also be used to determine the relevance of each parameter as in the False Discovery Rate (Benjamini and Hochberg, 1995; Bunea et al., 2006), or as in a more recent procedure that appears to provide better results in terms of variable selection (Rohart, 2011).

In all the previously described methods, observations are considered to be independent and identically distributed. These methods are therefore no longer appropriate when structured information, such as family relationships or common environmental effects, becomes available. In a linear mixed model, the observations are assumed to be clustered. The variance–covariance matrix $\boldsymbol{V}$ of the observations is therefore no longer diagonal but, in some cases, it is block diagonal. In the literature, most reports of linear mixed models relate to the estimation of variance components, using either maximum likelihood estimation (ML) (Henderson, 1973, 1953), or restricted maximum likelihood estimation (REML) which accounts for the loss in degrees of freedom due to fitting fixed effects (Patterson and Thompson, 1971; Harville, 1977; Henderson, 1984; Foulley et al., 2006). However, both methods assume that each fixed effect and each random effect is relevant. This assumption might be wrong and result in falsely estimated parameters. This might be especially the case in high-dimensional analysis. Contrary to linear models, the problem of selecting the fixed effect coefficients in a linear mixed model framework has rarely been addressed in a high dimensional setting.

Both Bondell et al. (2010) and Ibrahim et al. (2011) used penalized likelihoods to perform selection of both fixed and random effects. Bondell et al. (2010) introduced a constrained EM algorithm to solve the optimization problem, which becomes computationally complex in a high-dimensional context (it should be noted that their simulation studies were only designed for a low dimensional setting). Moreover, the methods of both Bondell et al. (2010) and Ibrahim et al. (2011) rely on Cholesky decompositions and, as pointed out by Müller et al. (2013), these decompositions are dependent on the order in which the random effects appear and are not permutation invariant (Pourahmadi, 2011). In the present paper, we primarily focus on analyzing data sets with only a few random effects and we therefore do not address the selection of both fixed and random effects.

Schelldorfer et al. (2011) have studied the selection of fixed effects in a high dimensional setting. Their paper introduced an algorithm based on ℓ^1-penalization of the maximum likelihood estimator in order to select the relevant fixed effect coefficients. As highlighted in their paper, their algorithm relies on the possibly time-consuming process of inverting the variance matrix of the observations $\boldsymbol{V}$.

The objective of this paper is two-fold. The first is to provide a more efficient way to select fixed effects in a linear mixed model. We consider the random effects as missing data, as previously described in Bondell et al. (2010) and Foulley (1997), and we introduce a ℓ^1-penalization on the log-likelihood of the complete data. A similar approach is studied in Groll and Tutz (2014) in the framework of Generalized Linear Mixed Models. We propose a multicycle Expectation Conditional Maximization algorithm (ECM) with convergence properties (Foulley, 1997; McLachlan and Krishnan, 2008; Meng and Rubin, 1993) to solve the optimization problem and provide theoretical results when the variances of the observations are known. The second objective is to increase the performance of variable selection. Due to its step design, the ECM algorithm can be combined with any variable selection method built for linear models. We propose to use a multiple testing procedure introduced in Rohart (2011) instead of the ℓ^1-penalization of the maximum likelihood estimator. We show that this procedure exhibits a higher percentage of recovery of the exact set of variables, a lower false discovery rate and a better estimation of $\boldsymbol{\beta}$, which induces a reduced mean squared error. As the selection of fixed effects in a high-dimensional linear mixed model framework has been rarely addressed before, we will mainly compare our results to those of Schelldorfer et al. (2011).

The proposed approach is then applied to a real data set from a project in which hundreds of pigs were studied, the aim being to shed light on the relationships between some of the phenotypes of interest and metabolomic data (Rohart et al., 2012). Linear mixed models are appropriate in this case because observations are in fact repeated data collected in different environments (groups of animals reared together in the same conditions). Some individuals were also genetically related, introducing a family effect. The data set consisted of 506 individuals from 3 breeds, 8 environments and 157 families, metabolomic data contained $p = 375$ variables, and the phenotype investigated was the Daily Feed Intake (DFI).

This paper is organized as follows. We first introduce the linear mixed model and its objective function to solve. We then describe the multicycle ECM algorithm used to solve the optimization problem. In Section 3, the algorithm described in Section 2 is extended to be used with any variable selection method developed for linear models. We assess the performance of the approach on a simulation study and demonstrate that the combination of this new algorithm with a multiple testing procedure for variable selection greatly improves the False Discovery Rate (Section 4). Finally, in Section 5, we illustrate the proposed approach on the metabolomic pigs data set.

2. Selection with ℓ^1-penalization

Let us introduce some notation that will be used throughout the paper. $\boldsymbol{Var}(\boldsymbol{a})$ denotes the variance–covariance matrix of the vector $\boldsymbol{a}$. For all $a > 0$, let $\boldsymbol{I}_a$ be the identity matrix of $\mathbb{R}^a$. For $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, denote I a subset of $\{1, \ldots, n\}$ and J a subset of $\{1, \ldots, p\}$. Let $\boldsymbol{A}_{I,J}$, $\boldsymbol{A}_{\cdot,J}$ and $\boldsymbol{A}_{I,\cdot}$ denote submatrices of $\boldsymbol{A}$ respectively composed of elements of $\boldsymbol{A}$ with rows in I and columns in J, columns in J and all rows, and rows in I and all columns. Moreover, for all $a > 0$, $b > 0$, denote $\mathbf{0}_a$ to be the vector of size a in which all coordinates are 0 and $\mathbf{0}_{a\times b}$ to be the zero matrix of size $a \times b$. Let us denote $|A|$ the determinant of the matrix $\boldsymbol{A}$.