Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A globally convergent algorithm for lasso-penalized mixture of linear regression models

Luke R. Lloyd-Jones [a,*], Hien D. Nguyen [c], Geoffrey J. McLachlan [b]

[a] Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland, 4072, Australia
[b] School of Mathematics and Physics, University of Queensland, St Lucia, Queensland, 4072, Australia
[c] Department of Mathematics and Statistics, La Trobe University, Bundoora Victoria, 3086, Australia

ARTICLE INFO

Article history:
Received 2 November 2016
Received in revised form 28 August 2017
Accepted 3 September 2017
Available online xxxx

Keywords:
Lasso
Mixture of linear regressions model
MM algorithm
Major League Baseball

ABSTRACT

Variable selection is an old and pervasive problem in regression analysis. One solution is to impose a lasso penalty to shrink parameter estimates toward zero and perform continuous model selection. The lasso-penalized mixture of linear regressions model (L-MLR) is a class of regularization methods for the model selection problem in the fixed number of variables setting. A new algorithm is proposed for the maximum penalized-likelihood estimation of the L-MLR model. This algorithm is constructed via the minorization–maximization algorithm paradigm. Such a construction allows for coordinate-wise updates of the parameter components, and produces globally convergent sequences of estimates that generate monotonic sequences of penalized log-likelihood values. These three features are missing in the previously presented approximate expectation–maximization algorithms. The previous difficulty in producing a globally convergent algorithm for the maximum penalized-likelihood estimation of the L-MLR model is due to the intractability of finding exact updates for the mixture model mixing proportions in the maximization-step. This issue is resolved by showing that it can be converted into a simple numerical root finding problem that is proven to have a unique solution. The method is tested in simulation and with an application to Major League Baseball salary data from the 1990s and the present day, where the concept of whether player salaries are associated with batting performance is investigated.

1. Introduction

Variable selection is an old and pervasive problem in regression analysis and has been widely discussed because of this; see George (2000) and Greene (2003, Ch. 8) for classical introductions to the topic, and see Hastie et al. (2009, Ch. 3) and Izenman (2008, Ch. 5) for some modern perspectives. In recent years, regularization has become popular in the statistics and machine learning literature, stemming from the seminal paper of Tibshirani (1996) on the least absolute shrinkage and selection operator (lasso). A recent account of the literature regarding the lasso and related regularization methods can be found in Bühlmann and van de Geer (2011). The mixture of linear regressions (MLR) for modeling heterogeneous data was first considered in Quandt (1972). The introduction of the expectation–maximization (EM) algorithm by Dempster et

* Corresponding author.
E-mail address: l.lloydjones@uq.edu.au (L.R. Lloyd-Jones).

http://dx.doi.org/10.1016/j.csda.2017.09.003
0167-9473/

al. (1977) made such models simpler to estimate in a practical setting. Subsequently, MLR models became more popular; see DeSarbo and Cron (1988), De Veaux (1989), and Jones and McLachlan (1992) for example.

The lasso-penalized MLR model (L-MLR) was considered in Khalili and Chen (2007) among a class of other regularization methods for the model selection problem in the fixed number of variables setting. The L-MLR was then generalized to the divergent number of variables setting in Khalili and Lin (2013), and to the mixture of experts setting in Khalili (2010). Furthermore, Stadler et al. (2010) (see also Bühlmann and van de Geer (2011, Sec. 9.2)) considered an alternative parameterization of the L-MLR to Khalili and Chen (2007), and suggested a modified regularization expression. An alternative modified grouped lasso criterion (Yuan and Lin, 2006) was suggested for regularization of the MLR model in Hui et al. (2015). A recent review of the literature regarding the variable selection problem in MLR models can be found in Khalili (2011).

In this article, we propose a new algorithm for the maximum penalized-likelihood (MPL) estimation of L-MLR models. This algorithm is constructed via the MM (minorization–maximization) algorithm paradigm of Lange (2013, Ch. 8). Such a construction allows for some desirable features such as coordinate-wise updates of parameters, monotonicity of the penalized likelihood sequence, and global convergence of the estimates to a stationary point of the penalized log-likelihood function. These three features are missing in the approximate-EM algorithm presented in Khalili and Chen (2007). Previously, MM algorithms have been suggested for the regularization of regression models in Hunter and Li (2005), where they are noted to be numerically stable. Coordinate-wise updates of parameters in lasso-type problems were considered in Wu and Lange (2008), who also noted such updates to be fast and stable when compared to alternative algorithms. Furthermore, Stadler et al. (2010) also consider a coordinate-wise update scheme in their generalized EM algorithm, although the global convergence properties of the algorithm could only be established for the MPL estimation of a modified case of the L-MLR model with a simplified penalization function. Note that we use the phrase "global convergence" to mean that the algorithm generates a sequence of updates that converge to some limiting value, when initialized from any point in the parameter space. This is in opposition to the alternative interpretation that the algorithm generates a sequence of updates that converge to a globally optimal solution of the problem being solved. Our use of the phrase "global convergence" is in compliance with the terminology used in the optimization literature; see Sriperumbudur and Lanckriet (2012) for example.

The difficulty in producing a globally convergent algorithm for the MPL estimation of the L-MLR model, which led both Khalili and Chen (2007) and Stadler et al. (2010) to utilize approximation schemes, is due to the intractability of the problem of updating the mixture model mixing proportions in the maximization-step of their respective algorithms. In our algorithm, we solve this issue by showing that it can be converted into a simple numerical root finding problem and prove that a unique solution exists. Aside from the new algorithm, we also consider the use of the L-MLR as a screening mechanism in a two-step procedure, as suggested in Bühlmann and van de Geer (2011, Sec. 2.5). Here, the L-MLR model is used to select the variable subset (step one) to include in a subsequent estimation of a MLR model (step two). This procedure allows for the use of generic asymptotic results for quasi-maximum likelihood estimation, such as those of Amemiya (1985); see also White (1982). Optimization of the lasso tuning parameter vector $\boldsymbol{\lambda}$ via derivative free numerical methods is also explored as an alternative to exhaustive grid search.

To supplement the presented algorithm (see Appendix A) and procedures, we perform a set of simulation studies to demonstrate the capacity of our methodology. A user-friendly program that implements the proposed algorithm in C++ is also available at https://github.com/lukelloydjones/LMLR and is shown to be capable of handling reasonably large estimation problems. To compare the performance of the method on real data, we analyze the same data set on Major League Baseball (MLB) salaries presented in Khalili and Chen (2007). This allows for an initial comparison with the foundational work and an exploration of whether common measures of batting performance are good predictors of how much a batter is paid. This analysis is supplemented with a current data set from MLB seasons 2011–15, which allows for an investigation into how the distribution of salaries has changed and whether the same or new predictors are relevant. Baseball has always had a fascination with statistics, with baseball's link with statistics going back to the origins of the sport (Marchi and Albert, 2013). In particular, economists have long had great interest in the labor market and finances associated with MLB (Brown et al., 2015). Of notable fame is the Moneyball hypothesis (Lewis, 2004), which states that the ability of a player to get 'on base' was undervalued in the baseball labor market (before 2003) (Hakes and Sauer, 2006). Baseball statistics on player and team performance are some of the best kept of any sport especially in the modern era. Furthermore, baseball owners and players agree that playing performance is measurable and is associated with salary (Scully, 1974; Fullerton Jr. and Peach, 2016). In this article we emphasize the application of our new methodology to these data in so far as there exists a statistical association, and believe that the implications of the performance salary association, with respect to MLB as a whole, have been explored in more depth elsewhere.

The article proceeds as follows. In Section 2, we introduce the L-MLR model and present the MM algorithm for its MPL estimation. In Section 3, we discuss the use of L-MLR models for statistical inference, and present the two-stage screening and estimation procedure. Section 4 outlines the algorithm's implementation. Simulation studies are then presented in Section 5, and we apply our method to data from salaries of batters from Major League Baseball (MLB) from the 1990s and present day in Section 6. Conclusions are then drawn in Section 7.

Please cite this article in press as: Lloyd-Jones L.R., et al., A globally convergent algorithm for lasso-penalized mixture of linear regression models. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.09.003.