

Comments on: ℓ_1 -penalization for mixture regression models

Jianqing Fan · Jinchi Lv

Received: 7 April 2010 / Accepted: 23 May 2010 / Published online: 30 June 2010
© Sociedad de Estadística e Investigación Operativa 2010

We would like to wholeheartedly congratulate Professors Städler, Bühlmann and van de Geer for an interesting and important paper on developing the L_1 regularization theory and methodology in finite mixture regression (FMR) models. An innovated reparametrization scheme is introduced to ensure equivariance under affine transformations and enhance the performance. Some nonasymptotic oracle inequalities on the average excess risk of the Lasso-type estimator are established in high dimensions, where the number of covariates can be much larger than the sample size. The authors also introduce an efficient EM-type algorithm combined with an improved coordinate descent for implementation. We appreciate the opportunity to comment on several aspects of this paper.

This comment refers to the invited paper available at: doi:[10.1007/s11749-010-0197-z](https://doi.org/10.1007/s11749-010-0197-z).

Fan's research was partially supported by NSF Grants DMS-0704337 and DMS-0714554 and NIH Grant R01-GM072611. Lv's research was partially supported by NSF Grant DMS-0806030. We sincerely thank the Co-Editor, Professor Ricardo Cao, for his kind invitation to comment on this discussion paper.

J. Fan (✉)

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA
e-mail: jqfan@princeton.edu

J. Lv

Information and Operations Management Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA
e-mail: jinchilv@marshall.usc.edu

1 Challenges of FMR

Classical theory and methods have been developed on various statistical models that have concave log-density functions, which naturally give convex negative log-likelihood loss functions. Properties of the resulting estimators and their implementations are rooted on the well-developed classical convex optimization theory and algorithms. Yet, the FMR models are examples in which the log-densities may no longer be concave and thus the corresponding losses can be nonconvex. This adds extra challenges to the scope of the study that the authors have undertaken.

When the dimensionality p of the parameter space is comparable to or exceeds the sample size n , it is desirable to regularize the estimation. This regularization hopes to achieve two goals: efficiently recovering the true underlying sparse model and its parameters and improving the risk profile of the estimators. The contribution of the paper mainly focuses on the latter. For a comprehensive overview, see Fan and Lv (2010) on the recent developments of theory, methods, and implementations in high dimensional statistical inference.

There are legitimate reasons for focusing on the risk profile of penalized likelihood method. The parameters in FMR models are hard to estimate precisely. The problem is certainly much harder than the case where the missing class labels are known (and hence the mixture probabilities $\{\pi_j\}_{j=1}^k$ are known). Even in the latter case, there are only about $n_j = n\pi_j$ data points available for estimating parameters (β_j, ϕ_j) . Therefore, for classes with small π_j 's, the parameters are hard to estimate accurately. Yet, those classes contribute very little to the overall risk profile.

2 Folded-concave penalty and iteratively reweighted LASSO

How important is accurately recovering the true underlying models? The main goals of high dimensional regression and classification, according to Bickel (2008), are:

- To construct as effective a method as possible to predict future observations;
- To gain insight into the relationship between features and response for scientific purposes, as well as, hopefully, to construct an improved prediction method.

The former appears in problems such as the text and document classification and portfolio optimization, whereas the latter appears naturally in many genomic studies and other scientific endeavors where it is important to know the possible cause-effect relationship between the responses and covariates, according to Fan and Lv (2008, 2010). These two goals are not necessarily always the same. For the penalized L_1 method, these two goals can sometimes be very different. An example of this was given by Fan and Lv (2008) in which the data generating process is

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + \varepsilon, \quad \varepsilon \sim N(0, 1), \quad (1)$$

where covariates X_1, \dots, X_p follow the standard normal distribution, equally correlated with correlation coefficient ρ . When $p = 1000$ and $n = 70$, it is impossible to recover the variable X_4 (note that $\text{cov}(X_4, Y) = 0$) using LASSO, i.e., the true model, whereas in terms of risk profile (the prediction error in this case), LASSO works well,

thanks to the persistency property of the LASSO estimator (Greenshtein and Ritov 2004).

The aforementioned problem is partially due to the bias of LASSO penalty, as pointed out by Fan and Li (2001) and formally shown by Zou (2006). To attenuate the problem, Fan and Li (2001) proposed the penalized likelihood method, which minimizes

$$n^{-1} \ell_n(\boldsymbol{\beta}) + \sum_{j=1}^p p_\lambda(|\beta_j|), \tag{2}$$

where $\ell_n(\boldsymbol{\beta})$ is the negative log-likelihood. The misconception of many researchers in the field is that one has to solve the nonconvex problem (2). But this was never the intention of Fan and Li (2001). In absence of algorithms to effectively implement the penalized L_1 regression at that time, Fan and Li (2001) used the local quadratic approximations (LQA): Given the estimate $\boldsymbol{\beta}^{(k)}$ at the k th iteration, one minimizes the convex function (assuming $\ell_n(\boldsymbol{\beta})$ is convex)

$$n^{-1} \ell_n(\boldsymbol{\beta}) + \sum_{j=1}^p \frac{p'_\lambda(|\beta_j^{(k)}|)}{2|\beta_j^{(k)}|} \beta_j^2, \tag{3}$$

after ignoring the constant term $\sum_{j=1}^p [p_\lambda(|\beta_j^{(k)}|) - p'_\lambda(|\beta_j^{(k)}|)|\beta_j^{(k)}|/2]$. With the advancement on the computation of penalized L_1 -regression, Zou and Li (2008) proposed to use the local linear approximations (LLA), resulting in the following convex optimization problem

$$n^{-1} \ell_n(\boldsymbol{\beta}) + \sum_{j=1}^p p'_\lambda(|\beta_j^{(k)}|)|\beta_j|. \tag{4}$$

This is really the iteratively reweighted LASSO, the original intention of Fan and Li (2001). The weighting schemes decrease with $|\beta_j^{(k)}|$ since the function $p'_\lambda(\cdot)$ is required by Fan and Li (2001) to decrease to zero. Both algorithms are specific members of minorization–maximization (MM) algorithms, according to Hunter and Li (2005) and Zou and Li (2008). Hence, they converge.

There are many possible penalty functions satisfying the above properties. Fan and Li (2001) advocated the SCAD penalty given by

$$p'_\lambda(|\theta|) = \lambda \left\{ I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a - 1)\lambda} I(|\theta| > \lambda) \right\} \quad \text{for some } a > 2. \tag{5}$$

The MCP penalty of Zhang (2010) is given by

$$p'_\lambda(|\theta|) = (\lambda - |\theta|/a)_+, \tag{6}$$

which is a generalization of the hard-thresholding penalty (with $a = 1$) in Antoniadis (1997). Note that $p'_\lambda(0+) = \lambda$ in the above two examples. This means that zero is not an absorbing state, using the terminology of Fan and Lv (2008): Even $\beta_j^{(k)} = 0$

at some iteration, there is still a chance for it to escape from zero. An example is $\beta^{(0)} = \mathbf{0}$ and $\beta^{(1)}$ is indeed the LASSO estimator. SCAD does not stop there, but continues to ameliorate the bias issue of LASSO.

The weighted LASSO is also used in the paper to attenuate the bias issue. It takes the adaptive LASSO

$$p'_\lambda(|\theta|) = \frac{1}{|\theta|}, \tag{7}$$

using the notation in (5) or more generally $p'_\lambda(|\theta|) = \frac{1}{|\theta|^\gamma}$ for some $\gamma > 0$ according to Zou (2006). Unfortunately, for this kind of adaptive LASSO weighting scheme, zero is an absorbing state since $p'_\lambda(0+) = \infty$. Once a component hits zero at some iteration, it will remain at zero throughout the iterations. Therefore, SCAD is a better weighting scheme and it can be implemented by the authors' algorithm without any extra difficulty. It can also work together with the EM-coordinate decent algorithm in Sect. 6.

The authors establish elegantly the consistency of the Lasso estimator and the oracle property of the adaptive Lasso for the finite mixture Gaussian regression models with fixed dimensionality p . It would be interesting to generalize the results along this line to higher dimensionality p and growing number of components k , e.g., when p grows in a polynomial or non-polynomial (NP) order of sample size n (see, e.g., Lv and Fan 2009). Of course, the condition of a (root- n) consistent initial estimator would be rather stringent in high dimensions. Using SCAD-like penalty functions, this requirement disappears. The initial value $\beta^{(0)} = \mathbf{0}$ suffices when LLA is used. The question is then whether the oracle properties with NP-dimensionality holds using the SCAD-like penalty. We conjecture that for nonconvex smooth loss as considered by the authors, the oracle property should hold for SCAD-like folded-concave penalties under some regularity conditions.

The nonasymptotic oracle inequalities on the average excess risk of the Lasso-type estimator in high dimensions by the authors are very interesting. A natural question is whether similar nonasymptotic oracle inequalities hold for other regularization methods such as SCAD.

3 One-component FMR models revisited

To gain further insights on the convex penalty and folded-concave penalty such as SCAD, let us consider a specific case in which $k = 1$, the one-component FMR models. For this specific case, penalized likelihood methods have been extensively studied in high dimensional problems. Let us use a simulated example to illustrate the relative performance. Consider the multiple linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{8}$$

where \mathbf{y} is an n -dimensional response vector, \mathbf{X} is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is a p -dimensional parameter vector, and $\boldsymbol{\varepsilon}$ is an n -dimensional noise vector. Set $(n, p) = (50, 1000)$ with the true regression coefficients vector $\boldsymbol{\beta}_0 = \mathbf{0}$, which means that the true underlying sparse model is the null model. Sample the rows of \mathbf{X} as i.i.d.

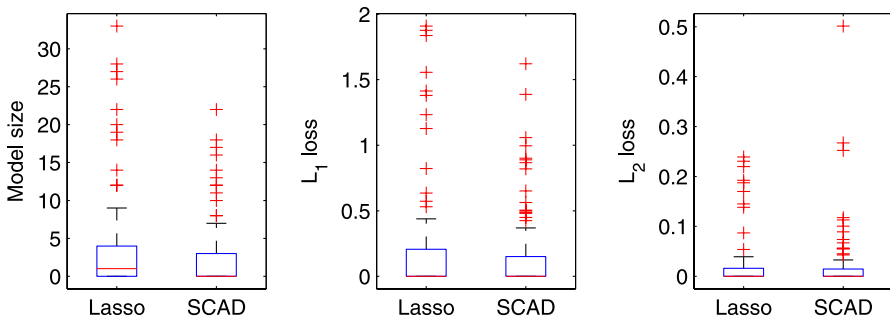


Fig. 1 Boxplots of selected model size, L_1 loss, and L_2 loss over 100 simulations for Lasso and SCAD in linear model (8), where $n = 50$ and $p = 1000$

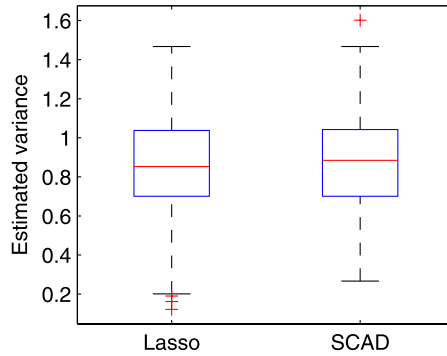
copies from $N(\mathbf{0}, \Sigma)$ with $\Sigma = (r_{ij})_{p \times p}$ and $r_{ij} = 0.8^{|i-j|}$, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ with $\sigma = 1$ independent of \mathbf{X} . We generated 100 data sets and applied Lasso and SCAD for variable selection, where the regularization parameter λ was tuned by using the five-fold cross-validation (CV). Both methods were implemented using the ICA algorithm. Figure 1 depicts the boxplots of selected model size, L_1 loss, and L_2 loss given by both methods. In particular, the medians of those three performance measures are 1, 0 and 0 for LASSO, and are all 0 for SCAD over 100 simulations.

4 Spurious correlations and variance components

From Fig. 1, there is over 50% chance that LASSO picks one or more variables to predict Y , which is really the realized noise in this case. This is called spurious correlation in Fan and Lv (2008). Let $\hat{\mathcal{S}}$ and \mathcal{S}_0 be the set of selected and true variables, respectively. Fan et al. (2010) argued that the variables in $\hat{\mathcal{S}} \cap \mathcal{S}_0^c$ are used to predict the realized noises. In our simulated null model, the variables in $\hat{\mathcal{S}}$ are chosen to best predict the noises. As a result, the residual sum of squares underestimates substantially the error variance. Therefore, when the selected model is inconsistent, the variance component σ^2 is substantially underestimated, as those spurious variables are selected to optimize the prediction of the realized noise. Figure 2 shows the estimated error variance corresponding to the selected models in Fig. 1.

The method described above is really the penalized likelihood estimator (3.6) or (3.8) when $k = 1$ for the variance component. It shows that the variance can be substantially underestimated when the selected model $\hat{\mathcal{S}}$ is inconsistent. On the other hand, model selection consistency is very hard to achieve in ultra-high dimensional problems. Therefore, the performance of penalized likelihood estimator (3.6) or (3.8), a direct plug-in method, will not perform robustly due to the spurious correlation inherent in ultra-high dimensional problems. Fan et al. (2010) proposed a refitted cross-validation to attenuate the problem of spurious correlation. It can be applicable to the FMR models for better estimation of the variance components.

Fig. 2 Boxplots of estimated error variance $\hat{\sigma}^2$ over 100 simulations for Lasso and SCAD in linear model (8), where $n = 50$ and $p = 1000$



References

- Antoniadis A (1997) Wavelets in statistics: A review. *J Ital Stat Soc* 6:97–144
- Bickel PJ (2008) Discussion of “Sure independence screening for ultrahigh dimensional feature space”. *J R Stat Soc Ser B* 70:883–884
- Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 96:1348–1360
- Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J R Stat Soc Ser B* 70:849–911
- Fan J, Lv J (2010) A selective overview of variable selection in high dimensional feature space (invited review article). *Stat Sin* 20:101–148
- Fan J, Guo S, Hao N (2010). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. Manuscript
- Greenshtein E, Ritov Y (2004) Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli* 10:971–988
- Hunter DR, Li R (2005) Variable selection using MM algorithms. *Ann Stat* 33:1617–1642
- Lv J, Fan Y (2009) A unified approach to model selection and sparse recovery using regularized least squares. *Ann Stat* 37:3498–3528
- Zhang C-H (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 38:894–942
- Zou H (2006) The adaptive Lasso and its oracle properties. *J Am Stat Assoc* 101:1418–1429
- Zou H, Li R (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann Stat* 36:1509–1566