Test
DOI 10.1007/s11749-010-0203-5

DISCUSSION

Rejoinder: ℓ_1-penalization for mixture regression models

Nicolas Städler · Peter Bühlmann · Sara van de Geer

Received: 13 May 2010 / Accepted: 23 May 2010

We are very grateful to all discussants for their many insightful and inspiring comments. We also would like to thank the co-editors Ricardo Cao and Domingo Morales for having arranged this discussion.

1 Variance estimation in linear model

Antoniadis, Sun and Zhang, and Fan and Lv raise the issue about estimation of the noise variance σ^2 in a linear model

$$Y = \mathbf{X}\beta + \varepsilon,$$

where $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 I_{n\times n}$. Knowledge of the noise level σ^2 is useful for, e.g., a "rough" selection of the tuning parameter λ in the Lasso:

$$\hat{\beta}_\lambda = \arg\min_\beta n^{-1}\|Y - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

This rejoinder refers to the comments available at: doi:10.1007/s11749-010-0198-y, doi:10.1007/s11749-010-0199-x, doi:10.1007/s11749-010-0200-8, doi:10.1007/s11749-010-0201-7, doi:10.1007/s11749-010-0202-6.

N. Städler · P. Bühlmann (✉) · S. van de Geer
Seminar für Statistik, ETH Zürich, 8092 Zürich, Switzerland
e-mail: buehlmann@stat.math.ethz.ch

N. Städler
e-mail: staedler@stat.math.ethz.ch

S. van de Geer
e-mail: geer@stat.math.ethz.ch

Published online: 30 June 2010

A reasonable value for the tuning parameter is then

$$\lambda = 2\sigma\sqrt{2\log(p)/n},$$

as used by Sun and Zhang (corresponding to their value $\sqrt{2\log(p)/n}$ since they scale the squared error with the factor $1/2$) which depends on the unknown noise level σ. As emphasized by Sun and Zhang, our estimator in (3.8) circumvents this problem by looking at a scaled Lasso estimator minimizing

$$\log\sigma + \frac{\|Y - \mathbf{X}\beta\|^2}{2n\sigma^2} + \lambda\|\beta\|_1/\sigma \tag{1.1}$$

with respect to β and σ. Now, a "universal" choice for the tuning parameter is $\lambda = \sqrt{2\log(p)/n}$ and, in particular, we do not need to specify σ. By the reparameterization as in (3.8), the problem is now convex in the new parameters (ρ, ϕ) and problem (1.1) can be efficiently minimized using coordinate descent (with coordinates $\rho, \phi_1, \ldots, \phi_p$). This idea is clearly described in Sun and Zhang, and their further detailed derivations are very interesting. They present some consistency and simulation results. Essentially, the naive estimator, based on residual sum of squares and some iteration procedure, performs better than the theoretically understood alternatives. We are pleased to see that Sun and Zhang have further advanced the issue of estimating the noise variance σ^2. A theoretical analysis of the naive estimator seems challenging due to the recursive nature of its definition.

Fan and Lv take the more "classical" approach of estimating the noise level σ^2:

$$\hat{\sigma}^2_\lambda = \left\|Y - \mathbf{X}\hat{\beta}_{\text{Lasso};\lambda}\right\|^2/n,$$

where $\hat{\beta}_{\text{Lasso};\lambda}$ is the ordinary Lasso estimator using the penalty parameter λ. (Instead of the factor $1/n$, we could use a modification using the factor $1/(n - \text{df})$ where df are the degrees of freedom of the Lasso (Zou et al. 2007).) Fan and Lv illustrate, using an example with pure noise, that $\hat{\sigma}^2_{\hat{\lambda}_{\text{CV}}}$ is under-estimating the true σ^2; here $\hat{\lambda}_{\text{CV}}$ denotes an estimate of λ using cross-validation. Fan and Lv call the reason "spurious correlation", and this happens since cross-validation typically selects too many variables. The refitted cross-validation method by Fan et al. (2010) is addressing some of these problems.

It is worth pointing out that estimation of σ^2 using the "classical" approach as in Fan and Lv or using the scaled Lasso procedure in (3.8) discussed also by Sun and Zhang, with the objective function corresponding to (1.1) above, are not equivalent at all. In fact, Sun and Zhang contribute additional insights by rewriting our Proposition 1 with their formula (10):

$$\hat{\sigma}^2_\lambda = \left\|Y - \mathbf{X}\hat{\beta}_\lambda\right\|^2/n + \hat{\sigma}_\lambda\lambda\left\|\hat{\beta}_\lambda\right\|_1,$$

where $\hat{\beta}_\lambda$ is the estimator from the scaled criterion function in (1.1). We clearly see that there is an additional term $\hat{\sigma}_\lambda\lambda\|\hat{\beta}_\lambda\|_1$ which causes some upward bias. Sun and Zhang propose in their formula (5) that the nuisance parameter should be "re-fitted" without penalty term. Even more generally, it seems to be a good strategy to re-fit all

parameters which enter in an unpenalized way into the criterion to be optimized. The conceptual differences between the approaches is again exploited in the empirical results: the "classical" approach is under-estimating the true variance (as discussed by Fan and Lv) while the other approach is over-estimating the true variance (as discussed in Sun and Zhang). We remark that a similar scaled Lasso procedure has been also considered by Barron and Luo (2008) (Theorem 3.2) where they propose a universal regularization parameter λ, based on the minimum description length principle.

Antoniadis makes an interesting connection to Huber's proposal for variance estimation in a linear model. Our proposal in (3.8) is likelihood-based and is changing the penalty function only. This is crucial for the formulation of an EM algorithm. Furthermore, the scaled Lasso in (3.8), i.e., using the criterion function from (1.1), allows using a universal penalty parameter λ which does not depend on the noise level anymore. We do not see why the latter property with Huber's proposal (which has not been intended by Huber for penalized estimation) should hold.

2 Non-convex penalty functions and iterative Lasso

Fan and Lv discuss why the SCAD is more favorable than the ℓ_1-penalty. Since the loss function in FMR models is not convex, they have a valid point since we have given up on convex optimization anyway. As they comment, the adaptive Lasso and its multi-iterated version can be thought as computational approximations for the non-convex SCAD penalty function. We agree with Fan and Lv that iteratively weighted ℓ_1-penalization performs often better than a single ℓ_1-penalization, particularly if the focus is on variable selection. Fan and Lv suggest that the iterative reweighting arising from the SCAD penalty, i.e., their formulae (4) and (5), is more desirable than the adaptive Lasso iteration we used in the paper which has zero as an absorbing state. We agree that the implementation would be straightforward. We do not know how big the differences would be in comparison to the adaptive Lasso we used (and we do not see a comparison between adaptive Lasso and SCAD in Fan and Lv's discussion).

The SCAD penalty for Gaussian mixture FMR models has been studied in Khalili and Chen (2007). However, they did not use the "trick" with taking the scale into the penalty function as in (1.1). For fixed p, Khalili and Chen (2007) derived an asymptotic oracle result for FMR models using SCAD; the result is exactly the same as when using the adaptive ℓ_1-penalization as described in our Theorem 2. Thus, in the fixed p asymptotics, there is no difference between SCAD and adaptive ℓ_1-penalization. In the high-dimensional framework, an oracle inequality could be established using the SCAD penalty. Since the penalty function is non-convex, the arguments would follow along the lines for analyzing ℓ_q-penalization with $q < 1$, and we refer to Birgé and Massart (2001) for the case with $q = 0$.

We agree that there are many positive aspects of SCAD and iteratively weighted ℓ_1-penalization. In many practical problems, such methods often perform better than plain ℓ_1-penalized estimators. However, as pointed out by del Barrio, the phenomenon of super-efficiency cannot be wiped away, and the scenario where the problem of super-efficiency is relevant occurs when the regression coefficients in a linear or

FMR model are small. In some of these settings, SCAD- or adaptive ℓ_1-penalized methods will not perform well in terms of prediction. We will make some further comments regarding this issue below in the next section.

3 Some theoretical issues

Del Barrio has made some comparisons between an ℓ_0-norm penalized estimator (called the Birgé–Massart (B–M) estimator) and the Lasso in linear models. We should emphasize that the additional log-factors in our Theorem 4 are due to technicalities for dealing with non-convex but smooth loss functions. In a linear model with the quadratic loss, our Theorem 4 can be sharpened as follows: with high probability

$$\left\|\mathbf{X}\hat{\beta}_\lambda - \mathbf{X}\beta_0\right\|^2/n \leq Cs_0 \log(p)/n, \tag{3.1}$$

where $s_0 = |\{j;\ (\beta_0)_j \neq 0\}|$, $C > 0$ is a constant depending on the restricted eigenvalue or the compatibility constant of the design and choosing $\lambda = C'\sqrt{\log(p)/n}$. Such oracle results are derived in, e.g., van de Geer (2008), Bickel et al. (2009), van de Geer and Bühlmann (2009). Thus, when assuming a compatibility condition on the design, the Lasso exhibits the same convergence rate as the B–M estimator. We remark that del Barrio mentions in his discussion our Condition 6 (about the maximal eigenvalue of the design): motivated by his comment, we managed to get rid of this assumption and there is no Condition 6 anymore appearing in our paper.

Del Barrio mentions the issue with small coefficients and connects to some problems with super-efficiency (as mentioned above). From a theoretical point of view, there are some results for high-dimensional settings addressing these points. First, when looking at prediction only, there is no requirement on the size of the coefficients: see our Theorem 4 in the paper and also formula (3.1) which is again derived without any assumption on the size of the coefficients. Furthermore, without any assumption on the size of the coefficients, we obtain in a linear model

$$\left\|\hat{\beta}_\lambda - \beta_0\right\|_1 \leq Cs_0 \log(p)/n,$$

where s_0 and $C > 0$ are as in (3.1). This result also holds for FMR models by extending Theorem 4 in a straightforward way to

$$\left\|\hat{\phi}_\lambda - \phi_0\right\|_1 \leq Cs_0 \log^3(n) \log(p \vee n)/n.$$

Thus, small coefficients do not really affect the ℓ_1-penalized estimator. However, when selecting variables in a more "aggressive" way than via ℓ_1-penalization, the situation may change. For the adaptive Lasso in a high-dimensional linear model, when aiming for not too many false positive selections of variables, the prediction error can become worse if, roughly speaking, there are too many small coefficients. This has been worked out in van de Geer et al. (2010). And thus we agree with del Barrio that small coefficients and good model selection can imply a loss in prediction accuracy.

Antoniadis asks whether our oracle inequality would still hold when the number k of mixture components would be unknown and estimated using the BIC criterion.

A rigorous derivation is not a simple consequence of our results: however, when the true number k_0 of mixture components is in the range $1 \le k_0 \le K_n$ with K_n potentially growing, but with maximal growth bounded by an expression involving the true unknown sparsity as well, an oracle inequality should carry over.

4 Other models

Lugosi presents some fascinating results and directions on combinatorial sparsity. As he points out, there are immense computational problems as the dimension N is very large (e.g., N equals all subsets of $\{1, \ldots, n\}$) and algorithms with linear complexity in N are not feasible anymore. His example with the multi-graph describes a beautiful trick how to come up with a simple algorithm.

Del Barrio mentions Gaussian mixture modeling with an ℓ_0-penalty method which has desirable statistical properties. Related to the issue above, the challenge is the computation. Maybe some greedy algorithm could be used: for the different problem of estimating the equivalence class of directed acyclic graphs from n observations, Chickering (2003) proves that a greedy search algorithm has the same asymptotic statistical properties as the BIC-optimal ℓ_0-regularized maximum likelihood estimator (when the dimension of the graph is fixed).

Antoniadis asks whether a varying coefficient model would be an alternative for addressing the issue of inhomogeneity in the data. Such a model is of the form

$$Y = \sum_{j=1}^{p} \beta_j\big(R^{(j)}\big) X^{(j)} + \varepsilon,$$

where $\beta_j(\cdot)$ are univariate smooth functions. The mixture (FMR) model we are considering in the paper is using a latent variable $Z \in \{1, \ldots, k\}$ to model different regression coefficient vectors which are totally unrelated to each other. On the other hand, a varying coefficient model uses one- or multi-dimensional observed variables $R^{(j)}$ to model different regression parameters which are related to each other via the smoothness of the functions $\beta_j(\cdot)$. Thus, models distinguish themselves whether the variable which causes different regression coefficients is observed or not and whether the different regression coefficient vectors are related to each other or not. The mixture modeling approach, using the blind approach for determining differences of the regression coefficients, is very "automatic" and "flexible".

References

Barron A, Luo X (2008) MDL procedures with ℓ_1 penalty and their statistical risk. In: Proceedings of the 2008 workshop on information theoretic methods in science and engineering

Bickel P, Ritov Y, Tsybakov A (2009) Simultaneous analysis of Lasso and Dantzig selector. Ann Stat 37:1705–1732

Birgé L, Massart P (2001) Gaussian model selection. J Eur Math Soc 3:203–268

Chickering D (2003) Optimal structure identification with greedy search. J Mach Learn Res 3:507–554

Fan J, Guo S, Hao N (2010) Variance estimation using refitted cross-validation in ultrahigh dimensional regression. Manuscript

Khalili A, Chen J (2007) Variable selection in finite mixture of regression models. J Am Stat Assoc 102:1025–1038
van de Geer S (2008) High-dimensional generalized linear models and the Lasso. Ann Stat 36:614–645
van de Geer S, Bühlmann P (2009) On the conditions used to prove oracle results for the Lasso. Electron J Stat 3:1360–1392
van de Geer S, Zhou S, Bühlmann P (2010) Prediction and variable selection with the Adaptive Lasso. Arxiv preprint 1001.5176 [mathST]
Zou H, Hastie T, Tibshirani R (2007) On the "degrees of freedom" of the Lasso. Ann Stat 35:2173–2192