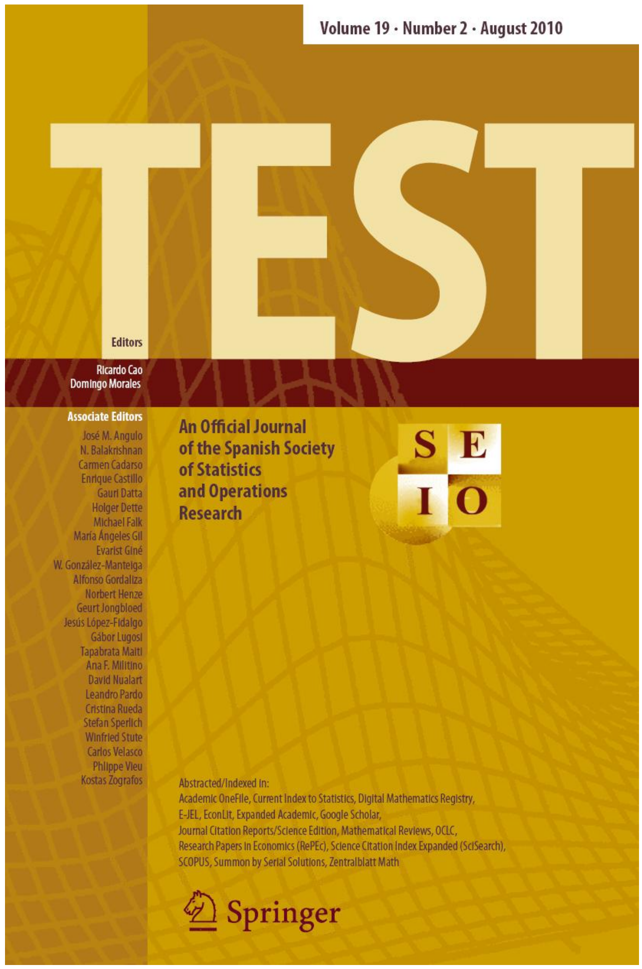
ISSN 1133-0686, Volume 19, Number 2



Test (2010) 19: 257–258
DOI 10.1007/s11749-010-0198-y

DISCUSSION

Comments on: ℓ_1-penalization for mixture regression models

Anestis Antoniadis

Received: 21 March 2010 / Accepted: 23 May 2010 / Published online: 30 June 2010

I congratulate the authors for a very interesting and timely contribution to an important problem: regularizing and recovering a sparse set of covariates in a finite mixture of regressions model with an extremely large number of predictors. The authors work primary with a regularized maximum likelihood procedure in a well chosen parametrization. This problem is formalized as a penalized maximum likelihood problem in which sparsity is induced by an ℓ_1-norm penalty on the coefficients. However, the resulting functional to be minimized is non-convex, which introduces numerical challenges in designing optimization algorithms to fit these models. The authors develop a general efficient EM methodology for solving the resulting non-convex optimization problem. The power of their methodology is well demonstrated, and there are many opportunities to explore its possible extensions. Their paper stimulated me to ask a few questions and to address the following comments.

Question 1—FMR and VCM. The authors have chosen a finite mixture of regressions (FMR) model for explaining a regression type relation between a response and large set of covariates where at least a fraction of the covariates may exhibit a different influence on the response among various observations. Another class of models that may address similar problems is the class of varying-coefficient multivariate regression models originally suggested by Hastie and Tibshirani (1993) that arise from nonparametric regression, nonlinear time series modeling and forecasting, functional data analysis, longitudinal data analysis, and others. Again, the main idea of the VCM is to allow regression coefficients to interact with another variable which can be an

This comment refers to the invited paper available at: doi:10.1007/s11749-010-0197-z.

A. Antoniadis (✉)
Departement of Statistics, Laboratoire Jean Kuntzmann, University Joseph Fourier, Grenoble 38041, France
e-mail: Anestis.Antoniadis@imag.fr

indicator of a subgroup of observations. While many procedures have been developed for estimating the varying-coefficients, the problem of variable selection for such high-dimensional problems has not been often addressed especially with parameterizations similar to the ones adopted by the authors. Is it worth investigating the extent to which the methodology adopted by the authors can be applied for such models and which may yield insights into the applicability of non-convex penalized maximum likelihood procedures to real data contexts where the assumption of FMR model may be unwarranted?

Question 2—Reparameterization, equivariance and convexity. Consider the arguments given in Sect. 3 for reparameterizing a (non-mixture) linear model which justifies the parameterization chosen later by the authors for their mixture of regressions models. As noted by the authors, especially for mixtures models, it is crucial to properly estimate the variance parameter, and this justifies the fact that one has to take σ^2 into the definition and optimization of the penalized maximum likelihood estimator defined in expression (3.6). However, such a choice not only doesn't lead to an equivariant estimator, but further leads to a non-convex optimization. The authors address this drawback by using a penalty term equal to $\lambda\frac{\|\beta\|_1}{\sigma}$. Another possible way to address the concomitant scale estimation could be to extend one of the ways addressed by Huber (1981, page 179) to jointly estimate σ and β in a linear model. One of his ideas for robust regression is to minimize

$$\sigma + \sigma\|Y - X\beta\|^2/(2n\sigma^2) + \lambda\|\beta\|_1,$$

over β and σ. It is easy to see that the above criterion is now jointly convex as a function of (β, σ) and that the resulting estimators of the parameters are equivariant. Somehow, Huber's technique has convexified the penalized Gaussian log likelihood. Do the authors think that such a parametrization could lead to a more efficient optimization algorithm?

Question 3—EM like algorithm and mixture complexity. The penalized approach, the derived algorithm and the very nice theoretical and asymptotic properties both for the low- and high-dimensional cases via appropriate and nonstandard oracle inequalities are impressive. However, they rely upon the fact that the order k of the mixture of regressions is fixed and known. Do the authors believe or have any arguments that allow us to hope that similar asymptotic results can be expected when the order is unknown and eventually estimated by an appropriate penalization as the one suggested in their Sect. 3.4? This is crucial, especially when one uses EM like algorithms to estimate the true model.

I would like to conclude by congratulating again the authors for such a wonderful piece of work!

References

Hastie T, Tibshirani RJ (1993) Varying-coefficient model. J R Stat Soc Ser B 55:757–796
Huber PJ (1981) Robust statistics. Wiley, New York