Journal de la Société Française de Statistique
Vol. 155 No. 2 (2014)

Statistique en grande dimension : problématiques et enjeux

Title: Statistic in high dimension: problems and challenges

Charles Bouveyron[1]

La statistique a dû faire face récemment à une modification majeure de la nature des données. L'accroissement de la masse et de la taille des données a en effet nécessité la proposition de nouvelles approches statistiques adaptées aux caractéristiques des données modernes. En particulier, la grande dimension des données (nombre important de variables) pose un ensemble de problèmes à la statistique multivariée classique que l'on résume usuellement par le terme "fléau de la dimension", dû à Richard Bellman. Parmi les problèmes que posent la grande dimension des données ont peut citer les problème numériques, les problèmes d'inférence ou les problèmes de biais des estimateurs. Il a donc été nécessaire de développer ces dernières années des méthodes capables de pallier ces problèmes. Ce numéro spécial se propose de présenter quelques-unes de ces nouvelles méthodes de statistique en grande dimension.

L'article de Loic Yengo, Julien Jacques et Christophe Biernacki considère le problème de la classification des variables en régression linéaire sur données de grande dimension. Le regroupement des variables est en effet une manière de faire de la réduction de dimension et donc d'introduire de la parcimonie dans le modèle de régression. Pour ce faire, les auteurs considèrent les regresseurs comme aléatoires et distribués selon une loi mélange. Cela permet de contenir le nombre de paramètres du modèle tout en permettant la détection des variables non utiles (associées à une composante du mélange centrée en zéro).

La contribution de Gilles Celeux, Marie-Laure Martin-Magniette, Cathy Maugis-Rabusseau et Adrian E. Raftery porte également sur la sélection de variables mais dans le contexte de la classification non-supervisée (*clustering*) en grande dimension. L'approche retenue dans cet article est celle de la sélection de modèles : le choix des variables utiles est vu comme un problème de sélection entre des modèles où les variables jouent alternativement des rôles différents. L'approche est en particulier comparée à une proposition récente de Witten et Tibshirani (sparse K-means) et est appliquée à la classification de données de transcriptome.

Enfin, l'article de Raphaël Coudret, Benoit Liquet et Jérôme Saracco considère la régression inverse par tranches pour l'analyse de données de grande dimension. La régression inverse par tranches (sliced inverse regression, SIR) est en effet particulièrement intéressante dans le cas où des relations non-linéaires existent entre la variable à expliquer et les covariables. Cependant, la méthode SIR s'avère instable quand le nombre de prédicteurs est plus grand que le nombre

[1] Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes & Sorbonne Paris Cité
E-mail : charles.bouveyron@parisdescartes.fr

 ISSN: 2102-6238

d'observations et il est alors nécessaire d'avoir recours à la régularisation. Cet article introduit deux nouvelles méthodes de régulation pour la méthode SIR et propose une comparaison avec d'autres méthodes de régression, notamment sur des données issues de la génomique.

Nous espérons que ce numéro spécial permettra aux lecteurs du Journal de la SFdS de mesurer la diversité des recherches menées dans le domaine de la statistique en grande dimension. La lecture de ce numéro spécial donnera peut-être à certains l'envie de contribuer à ce domaine passionnant et très actif de la statistique. Pour conclure, nous tenons à remercier à Gilles Celeux, éditeur en chef du journal de la SFdS, qui est à l'initiative de ce numéro spécial. Nous remercions enfin les auteurs pour leurs contributions ainsi que les relecteurs anonymes pour la qualité de leur travail.

In the past decade, Statistics is facing a dramatic shift in the nature of data. The size and dimensionality of modern data are indeed significantly different from the data of the past century and require new statistical techniques to be managed properly. In particular, the high dimensionality of the data (large number of variables) is a well-known problem in multivariate statistics which can be subsumed under the heading "the curse of dimensionality" (R. Bellman, 1957). Among the problems linked to the high dimensionality of data, one can cite the numerical and inference issues and the estimators' bias. Fortunately, researchers have proposed in the last years a large range of methodologies able to circumvent these problems. This special issue presents some of those new statistical methodologies.

The article of Loic Yengo, Julien Jacques and Christophe Biernacki focuses on the problem of variable classification in high-dimensional linear regression. Their approach considers a mixture model over the regressors which allows to keep the the regression model sparse. This strategy allows in addition to identify the irrelevant variables associated to the mixture component centered in zero.

The work of Gilles Celeux, Marie-Laure Martin-Magniette, Cathy Maugis-Rabusseau and Adrian E. Raftery concerns also variable selection but in the context of high-dimensional clustering. Their approach recasts the problem of variable selection as a model selection problem. The proposed methodology is in particular compared with a recent work of Witten and Tibshirani (sparse k-means) on transcriptomic data.

Finally, Raphaël Coudret, Benoit Liquet and Jérôme Saracco introduce two new regularization techniques for the sliced inverse regression (SIR) method. Regularization is indeed necessary when using SIR on high-dimensional problem with small sample size. The authors compare their proposal with existing regularization on genomic data.

We hope that this issue of the Journal de la SFDS will allow the readers to evaluate the diversity of the works in high-dimensional statistics. We want to greatly thank Gilles Celeux, editor-in-chief of the Journal de la SFDS, for soliciting such a special issue. We are also very grateful to the authors for contributing to this issue and the anonymous reviewers for the quality of their work.

 ISSN: 2102-6238