Journal of Econometrics 186 (2015) 277–279

Contents lists available at ScienceDirect

Journal of Econometrics

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

Editorial

High dimensional problems in econometrics

The technological innovations in information processing and the increased storage capability have made possible to collect very large data sets in various fields of economics and finance. Researchers, companies, and governments look for ways to exploit this rich information. This special issue collects 11 papers who present state-of-the-art techniques to deal with many predictors, many regressors or many instruments. It grew out of the CIREQ conference on high dimensional models in econometrics organized by Marine Carrasco and Sílvia Gonçalves in Montreal, Canada, on May 4–5, 2012.

Many predictors or regressors

In this issue, five papers consider models with many predictors or regressors. Two of them use a factor-augmented regression model as underlying model where the variable to be forecast depends on a few latent factors. Two others use Lasso as dimension reduction device and one uses series least-squares estimation.

The factor model has become extremely popular in empirical work since Stock and Watson's (2002) seminal paper. By assuming that the data are generated by an approximate factor model, factor-augmented regression models effectively achieve dimension reduction by condensing information on a large number of predictors into a few common factors (the so-called diffusion indices). The standard estimation approach is a principal components regression, where we first estimate the common factors by a few principal components and then use these as predictors for the target variable of interest. The papers by Cheng and Hansen and Kelly and Pruitt contribute to this literature by proposing forecast methods that account for the selection of the factors, an important issue when estimating factor-augmented regression models.

Cheng and Hansen ("Forecasting with factor-augmented regression: a frequentist model averaging approach") consider forecast combination based on model averaging criteria. The main idea is that rather than forecasting with a single factor-augmented regression model, a forecaster can do better by averaging forecasts obtained from many different models whose predictors vary according to which factors and which lags of these factors are included; forecast models may also depend on lagged values of the dependent variable. The weights on these individual forecasts are chosen to minimize estimates of the mean squared forecast error (MSFE). For one-step ahead forecasts, the estimate of the MSFE is obtained by a Mallows criterion whereas for h-step ahead forecasts leave-h-out cross validation is used. These model averaging criteria have been studied in the context of regression models with observed regressors. Cheng and Hansen extend their applicability to the context of factor-augmented regressions where some of the predictors are latent and need to be estimated in a first step. Their main result shows that Mallows and leave-h-out cross validation criteria remain asymptotically unbiased estimates of the MSFE despite the presence of generated regressors. Thus, no adjustment is needed to account for factors estimation uncertainty when using these two model averaging criteria. This is in contrast with inference on the factor-augmented regression coefficients, where Bai and Ng (2006) showed that factors estimation uncertainty can be safely ignored only when the number of predictors from which we extract the factors is sufficiently large compared to the time series dimension. The simulations show that model averaging based on Mallows and leave-h-out cross validation yields lower MSFE than alternative methods.

Kelly and Pruitt ("The three-pass regression filter: a new approach to forecasting using many predictors") propose a new estimation method for factor-augmented regression models. The main difference with the standard principal components regression approach is that only factors that are relevant for the target variable are used when forecasting this variable. Instead, principal components regressions select the factors that are most correlated with the predictors and these are not necessarily those that are most correlated with the target variable. To identify the target-relevant factors, Kelly and Pruitt propose a three-step procedure: in the first step, one runs time series regressions of each predictor on a set of factor proxies and obtain a cross section of estimated slope parameters. In a second step, one runs cross section regressions of predictors on the estimate slope parameters from step 1 and obtain a time series of estimated factors. Finally, one runs a time series regression of the target variable on the estimated factors from step 2. This three-pass-regression filter relies on observed variables that act as proxies for the relevant factors. The factor proxies can be obtained from economic theory. To deal with the case where such proxies are not available, the authors propose a way to construct automatic proxies from the available data on the target variable and the panel predictors. Interestingly, the three-pass-regression filter amounts to a form of partial least squares when automatic proxies are used. The paper derives the asymptotic properties for the new method and shows its finite sample superiority over alternative methods. The main conclusion that emerges is that the three-pass-regression filter often leads to more accurate forecasts, especially when the relevant factors are among the weakest principal components.

An alternative method for dealing with many predictors/regressors in high dimensional regression models is to seek estimators that minimize a penalized version of the least squares

http://dx.doi.org/10.1016/j.jeconom.2015.02.009
0304-4076/

residuals function. Popular methods are the LASSO and the adaptive LASSO (or ALASSO), which are based on ℓ_1-penalized least squares criterion functions. Under appropriate sparsity conditions, these methods perform simultaneous variable selection and estimation of high dimensional regression models, where the number of regressors is potentially much larger than the number of observations. For simple data structures such as linear regression models with fixed regressors, it is well known that the LASSO/ALASSO correctly identify the non-zero components of the regression coefficients vector with probability tending to one and, at the same time, estimate these non-zero components accurately, with the same asymptotic precision as that of the OLS method applied to the true model (this is the so-called oracle property).

Two papers in the special issue rely on the LASSO and ALASSO to estimate high dimensional linear models subject to a sparsity condition: the paper of Chatterjee, Gupta, and Lahiri and the paper of Kock and Callot. The first of these considers the setup of a linear regression model with i.i.d. errors and a large number of fixed regressors whereas the second considers the case of a high dimensional linear vector autoregression model.

Chatterjee, Gupta and Lahiri ("On the residual empirical process based on the ALASSO in high dimensions and its functional oracle property") extend the oracle property of ALASSO for the regression parameters to the residual empirical process. The main theoretical result is an asymptotic uniform linearity (AUL) property according to which the ALASSO-based residual empirical process can be uniformly approximated by the empirical process of the unobserved regression error plus a linear term that depends on the normalized ALASSO regression estimator. This result is established under a sparsity condition and holds when the number of estimated coefficients p is much larger than the sample size n (in particular, p is allowed to increase with n at an arbitrarily large polynomial function of n). Based on this result, the authors show that the distribution function of the ALASSO-based residuals converges to the same Gaussian process as that of the OLS residuals based on the oracle. This functional oracle property allows for the construction of confidence bands for the error distribution function. Another application of the AUL property is the construction of prediction intervals for new observations of the dependent variable. The authors consider both applications in the context of a simulation study and show that ALASSO-based inference is superior in finite samples to alternative methods of inference, including those without penalization.

Kock and Callot ("Oracle Inequalities for High Dimensional Vector Autoregressions") study the properties of Lasso and adaptive Lasso for estimating the coefficient of a stationary vector autoregressive (VAR) model with Gaussian error. The number of variables as well as the lag length in the VAR model are allowed to increase as the sample size increases. So that the number of unknown variables may be much larger than the time series length. However, the number of nonzero coefficients is much smaller. The authors extend oracle inequalities and asymptotic results that were known in the i.i.d. case to their dynamic setting. In particular, they establish that the estimators of the non-zero coefficients are asymptotically equivalent to the oracle assisted least squares estimators and the rate of convergence of these estimators is identical to the one of least-squares including only the relevant covariates.

The paper by Belloni, Chernozhukov, Chetverikov and Kato ("On the asymptotic theory for least squares series: pointwise and uniform results") considers nonparametric estimation of the conditional mean (and linear functionals of it, such as the average partial derivative function) using series least squares estimators. The function of interest is approximated by means of a few basis functions whose number k grows with the sample size n at a certain rate. Belloni et al. contribute to the literature on series estimators by providing a number of new asymptotic results, including pointwise and uniform convergence rates as well as an asymptotic distribution theory (in the form of central and functional central limit theorems). Results are obtained under weaker conditions on the rate of divergence of k with n than previously obtained in the literature and misspecification of the model is allowed i.e. the approximation error may be non-vanishing. One important result is that the series estimator of the conditional mean is shown to attain the optimal uniform convergence rate under rather general conditions.

Approximate factor models and large panels

Two other papers in the special issue that also rely on approximate factor models as a means of dimension reduction are Fan, Liao, and Shi and Onatski. The paper by Gonçalves and Kaffo considers a large panel.

The paper by Fan, Liao, and Shi ("Risks of Large Portfolio") is about assessing the estimation error of the risk of a large portfolio. In finance, the risk of portfolio of asset returns is usually measured by its variance. Unfortunately, the covariance matrix of the returns is unknown and needs to be estimated yielding an estimation error, which is likely to be more important when the number of assets is large. In this paper, the number of assets can be much larger than the sample size. Three estimators of the covariance are considered: (a) the sample covariance, (b) the estimator based on a factor structure with known factors, (c) the estimator based on a factor model with unknown factors. For these three estimators, the paper proposes a method to compute an upperbound (named H-club) of the risk estimator. The approach is similar to the construction of a standard confidence interval. Under the maintained assumption that the excess returns have an approximate factor structure, the limit distribution of these three risk estimators are derived, namely a standard normal after proper rescaling and then an asymptotic upper bound is computed.

The paper by Onatski ("Asymptotic analysis of the squared estimation error in misspecified models") derives asymptotic approximations of the squared estimation error of the common component in an approximate factor model when the number of factors is potentially misspecified. Both strong and weak factor asymptotics are considered. When factors are strong (i.e. their explanatory power dominates the explanatory power of the idiosyncratic error term), choosing the number of factors as to minimize the squared loss function leads to choosing the true number of factors. Nevertheless, when factors are weak (i.e. they are not pervasive and no clear separation exists between the largest eigenvalues and the rest of the eigenvalues of the sample panel covariance matrix), the minimizer of the squared loss function may be smaller than the number of true factors, reflecting the fact that it might not be optimal to include factors that are too weak to be accurately estimated. Onatski proposes consistent estimators of the asymptotic squared loss function under both strong and weak factors and shows that the minimizers of these loss estimates are asymptotically loss efficient. As it turns out, many of the estimators of the number of factors in the existing literature, including the popular Bai and Ng (2002) estimator, are asymptotically loss efficient when factors are strong but fail to be so when factors are weak. One additional use of the loss estimates proposed by Onatski is for model comparison: the difference in the loss estimate due to increasing the number of factors can be used to quantify the gain/loss from additional factors.

The paper by Gonçalves and Kaffo ("Bootstrap inference for linear dynamic panel data models with individual fixed effects") considers bootstrap inference in the context of a linear autoregressive panel data model with n individual fixed effects and T time series observations. When both n and T grow at the same rate, the ordinary least squares estimator is asymptotically biased (see Hahn and Kuersteiner (2002)). Gonçalves and Kaffo discuss the complications that this asymptotic bias introduce for bootstrap inference. Specifically, they show that the recursive-design bootstrap (whereby the

bootstrap observations are generated recursively using the estimated structure of the model, resampling from the residuals) is capable of replicating the entire distribution of the fixed effects estimator, including its asymptotic bias. In contrast, the pairs bootstrap as well as a fixed-design version of the bootstrap (where bootstrap observations on the dependent variable are generated by adding the resampled residuals to the estimated conditional mean, evaluated at the original observations) are not able to mimic the incidental parameter bias and are therefore invalid when used to approximate the distribution of the standard fixed effects estimator. Interestingly, the pairs bootstrap becomes valid when applied to the bias-corrected estimator of the common autoregressive coefficient, whereas the fixed-design bootstrap remains invalid.

Many Moments

The last three papers, namely those of Carrasco and Tchuente, Cheng and Liao, and Florens and Van Bellegem, are concerned with inference in the presence of many moment conditions. The dimension of the parameter of interest is finite in the first two papers, whereas it is infinite dimensional in the third paper.

Carrasco and Tchuente ("Regularized LIML for many instruments") consider the estimation of a finite dimensional parameter in a linear instrumental variable (IV) model using many instruments. The number of instruments may exceed the sample size. They consider regularized versions of the limited information maximum likelihood estimator (LIML) based on three different regularizations: Tikhonov, Landweber Fridman, and principal components. They show that the estimators are consistent and asymptotically normal in the heteroskedastic case and reach the semiparametric efficiency bound in the homoskedastic case. They show that the regularized LIML possesses finite moments when the sample size is large enough. The higher order expansion of the mean square error (MSE) reveals that the regularized LIML estimator has a smaller MSE than the regularized two-stage least squares estimator proposed by Carrasco (2012). A data driven selection of the regularized parameter based on the approximate MSE is proposed.

Cheng and Liao ("Select the Valid and Relevant Moments: An Information-Based LASSO for GMM with Many Moments") consider estimating a finite dimensional parameter starting from an increasing sequence of candidate moment conditions, some of which may be invalid and some of which may be valid but redundant. The estimation method simultaneously estimates the parameter and purges both invalid and redundant moment conditions. The method applies a Lasso type penalization where the penalty term depends on a preliminary consistent estimator which accounts for the strength and validity of moments. To obtain such preliminary estimator, a maintained assumption is that a fixed number of valid and relevant moment conditions are known a priori so that the identification is guaranteed without additional moment conditions. The main goal is, however, to choose additional moment conditions to improve the asymptotic efficiency of the initial GMM estimator.

The paper of Florens and Van Bellegem ("Instrumental Variable Estimation in Functional Linear Models") considers a regression model where the dependent variable is real but the regressor is a function, element of a Hilbert space. The object of interest is the estimation of the regression coefficient which is itself a function. The regressor is assumed to be endogenous and identification is achieved via an instrumental variable which is also a function belonging to some Hilbert space. This model can be thought of as a generalization of the usual instrumental variable regression where the numbers of regressors and instruments are infinite. In this case, the classical IV estimator is not consistent. The authors modify the classical IV by adding a Tikhonov regularization. They derive the rate of convergence of the estimator and the asymptotic normality of linear functionals of the estimator. An empirical application completes the paper. The dependent variable is the annual growth rate of gross domestic product per capita in the United Kingdom. The dependent variable is the age specific fertility rate which is a discretized curve for mother's age from 15 to 44 years old. The instrument used is the rate of twin (or multiple) birth. The framework of functional regression allows them to study the marginal impact of the age of fertility onto the economic growth.

Acknowledgements

We would like to thank the 2012 conference sponsors for their generous support. We are especially indebted to the main sponsor: Centre interuniversitaire de recherche en économie quantitative (CIREQ). We are also grateful to the Canada Research Chair in Economics (held by Russell Davidson), the William Dow Chair in Economics, McGill University (held by Jean-Marie Dufour), Journal of Applied Econometrics, and Centre interuniversitaire de recherche en analyse des organisations (CIRANO) for partial financial supports. We thank the program committee: Marine Carrasco, Jean-Marie Dufour, Jean-Pierre Florens, Sílvia Gonçalves, and Lynda Khalaf. The conference would not have been possible without the skills of CIREQ staff: Sharon Brewer and Mélanie Chartrand. Finally, special thanks go to the referees who devoted time and energy to the improvement of the papers in this issue.

References

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.
Bai, J., Ng, S., 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Econometrica 74, 1133–1150.
Carrasco, M., 2012. A regularization approach to the many instruments problem. J. Econometrics 170, 383–398.
Hahn, J., Kuersteiner, G., 2002. Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. Econometrica 70, 1639–1657.
Stock, J.H., Watson, M.W., 2002. Forecasting using principle components from a large number of predictors. J. Amer. Statist. Assoc. 97, 1167–1179.

Marine Carrasco *
Université de Montréal, Departement de Sciences Economiques, CP 6128, succ Centre Ville, Montreal, QC H3C3J7, Canada
E-mail address: marine.carrasco@umontreal.ca.

Victor Chernozhukov
Department of Economics, MIT, 50 Memorial Drive, E52-361B, Cambridge, MA 02142, USA
E-mail address: vchern@mit.edu.

Silvia Gonçalves
Université de Montréal, Departement de Sciences Economiques, CP 6128, succ Centre Ville, Montreal, QC H3C3J7, Canada
E-mail address: silvia.goncalves@umontreal.ca.

Eric Renault
Department of Economics, Box B, Brown University, Providence, RI 02912, USA
E-mail address: eric_renault@brown.edu.

Available online 2 March 2015

* Corresponding editor.