Business and Economics Journal
ISSN: 2151-6219

Sapra, Bus Eco J 2015, 6:3
DOI: 10.4172/2151-6219.1000168

Opinion Open Access

High-Dimensional Statistical Analysis in Business and Economics

Sunil Sapra*

Department of Economics and Statistics, California State University, USA

High-dimensional models have become increasingly common in business and economics in recent years. Modern data in business involve thousands to millions of records on individuals. Over the past few decades, with the availability of new technologies, it has become common in economics, finance, and marketing to collect data on a large number of features for a limited number of individuals. Fan et al. [1] review the literature on high-dimensional models and their applications in economics and finance. In vector autoregressive models, the number of parameters grows with the size of the model and the problem of estimating these models can quickly become computationally intractable. Panel data used widely in economics also offers an application of high-dimensional data analysis. In finance, volatility matrix estimation is an example of high-dimensional statistics. Scanner data on transactions by households on a large number of products is yet another example in marketing.

High-dimensional statistical analysis refers to aforementioned situations in which the number of parameters is much greater than the sample size. Unfortunately, common econometric techniques, which are suitable for low-dimensional data for which the number of records is greater than the number of covariates, are not suitable for high-dimensional data. So, what can go wrong if a technique suitable for low-dimensional data is applied in a high-dimensional setting? The main problem is that low-dimensional data analysis techniques, such as least squares or logistic regression will produce a perfect fit on the training data, but will produce a poor fit on an independent test data. James et al. [2] attributed this over-fitting for test data to excessive flexibility of the least squares and logistic regression in high-dimensional settings. Over-fitting means that the statistical procedure fits mostly noise to the data, instead of fitting the signal. The key therefore is to avoid over-fitting with high-dimensional data by fitting less flexible least squares or logistic regression models to the data, such as LASSO, principal components regression, ridge regression, forward stepwise selection, etc. The main idea underlying these techniques is regularization or shrinkage, which means reducing the number of non-zero coefficient estimates. Another key idea is that high-dimensional statistical problems are plagued by the curse of dimensionality: deterioration in the quality of the fitted model and predictions as new features are added to the model. As James et al. [2] explained, if additional signal features, which are truly associated with the response are included in the model, the quality of the fit indeed improves. However, if additional noise features, not associated with the response are included, the model fit and predictions deteriorate and the risk of over-fitting increases. High-dimensional data can be a blessing if the additional features are relevant and associated with the response resulting in a superior predictive model, but can also be a curse if these features constitute noise in which case the quality of predictions is poor. Furthermore, even for the relevant features, the additional variance that accompanies their inclusion may more than offset the reduction in bias. Finally, the results for high-dimensional techniques should be interpreted with caution due to extreme multicollinearity among features. Common measures of model fit, such as p-value, R-squared for low-dimensional settings can be misleading in high-dimensional settings since these measures will make the model fit appear almost perfect due to over-fitting.

High-dimensional statistical analysis is a fast evolving area of research and is highly promising for researchers in business and economics. It is likely to witness contributions of better techniques for prediction as well as inference in future years. Augmentation of penalized least squares and penalized likelihood methods for variable selection in sparse regression models discussed in Buhlmann and van de Geer [3] and Fan et al. [1] with data mining techniques as detailed in Belloni et al. [4] is a promising approach for improved prediction and inference.

References

1. Fan, Jianqing, Jinchi L and Lei Qi (2011) Sparse High-Dimensional Models in Economics. The Annual Review of Economics 3: 291-317.
2. James, Gareth, Danela Witten, Trevor Hastie, Robert Tibshirani (2013) An Introduction to Statistical Learning with Applications in R. Springer, New York.
3. Buhlmann, P. and Sara van de Geer (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, New York.
4. Belloni, A., Victor Chernozhukov, and Christian Hansen (2014) High-Dimensional Methods on Structural and Treatment Effects. Journal of Economic Perspectives 28: 29-50.

***Corresponding author:** Sunil Sapra, Professor, Department of Economics and Statistics, California StateUniversity, LosAngeles, CA90032, USA, Tel: (323)-343-2941; FAX: (323)-343-5462; E-mail: ssapra@exchange.calstatela.edu

Received July 13, 2015; **Accepted** July 15, 2015; **Published** July 22, 2015

Citation: Sapra S (2015) High-Dimensional Statistical Analysis in Business and Economics. Bus Eco J 6: 168. doi:10.4172/2151-6219.1000168