A Discussion on "Model Selection for Generalized Linear Models with Factor-Augmented Predictors"

Hansheng Wang and Chih-Ling Tsai

January 5, 2009

Professors Ando and Tsay should be congratulated for such nice work, which provides an effective statistical method to handle high dimensional datasets with generalized linear models. In this discussion, we supply and review additional references from the literature of both sufficient dimension reduction and variable selection. Thus, the authors' contribution could be further appreciated under an even larger context, both theoretically and practically.

The fast development of information technology enables us to collect an enormous amount of information at very low cost. Statistically, this means that we might face datasets with ultra-high dimensional predictors. As a result, analyzing high dimensional data effectively has become a fundamental and important subject in the last decade; see Fan and Li (2006) for excellent discussions. To effectively conduct dimension reduction in regression, one could construct "factors" so that they attain the optimal predictability. As noted by the authors, the traditional principal component analysis fails to be the mostly effective approach because the role of the response is completely ignored, and the resulting principal components might not be sufficiently predictive to the response. In fact, the same problem was also noted by Li (1991), which leads to the seminal theory of *effective dimension reduction* and an innovative method of *sliced inverse regression* (SIR). Later, such a theory was further developed by Cook (1998), which resulted in the theory of *sufficient dimension reduction*. Subsequently, various inverse regression methods have been proposed. For a comprehensive review, we refer to Cook (2007) and Luo et al. (2009). A close look at the computational details of most inverse regression methods (e.g., SIR) reveals that they are indeed a supervised version of dimension reduction, where the effect of the response has been carefully taken into consideration. Similar philosophy was also shared in the partial least square (PLS) literature (Helland, 1988, 1990; Naik and Tsai, 2000; Nguyen and Rocke, 2002; Ding and Gentleman, 2005; Li et al., 2007).

[†]Hansheng Wang is an Associate Professor of Statistics at Guanghua School of Management, Peking University, Beijing, 100871, P. R. China (E-mail: *hansheng@gsm.pku.edu.cn*). Chih-Ling Tsai is a Professor of Statistics at Graduate School of Management, University of California-Davis, California, 95616-8609, U.S.A (E-mail: *cltucd@gmail.com*). Wang's research is supported in part by a NSFC grant (No. 10771006) and a grant from Microsoft Research Asia.

While inverse regression is a useful method for dimension reduction, Bair et al. (2006) considered the supervised principal component analysis, which can be considered as a two-stage procedure. In the first stage, a set of relevant variables are selected by simple univariate regressions, which leads to a reduced (or supervised) dataset. In the second stage, principal components are extracted from the reduced dataset and are used for predictions. Recently, such a method was further improved by the method of pre-conditioning (Paul et al., 2008). In addition to the inverse regression and supervised principal component approaches, Fan and Lv (2008) proposed an innovative and stimulating method to analyze ultrahigh dimensional data. It is namely SIS (Sure Independence Screening), which is based on correlation learning to reduce dimensionality. The theoretical and empirical performances of the above methods have been well documented in the literature.

In this article, Professors Ando and Tsay obtain an information selection criterion for generalized linear models by employing factor analysis to cope with the high-dimensional variable problem. We believe exploring its relationships with aforementioned dimension reduction methods would not only enrich the authors' results but also broaden the applicability of those methods to generalized linear, dynamic, and time series models. Lastly, we want to congratulate the authors again for such interesting work. We look forward to seeing its important applications in the foreseeable future.

References

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by supervised principal components," *Journal of the American Statistical Association*, 101, 119–137.

Cook, R. D. (2007), "Fisher Lecture: Dimension reduction in regression," *Statistical Science*, 22, 1–26.

Cook, R. D. (1998), *Regression Graphics*, John Wiley, New York.

Ding, B. and Gentleman, R. (2005), "Classification using generalized partial least squares," *Journal of Computational and Graphical Statistics*, 14, 280–298.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space (with discussions)," *Journal of Royal Statistical Society, Series B.*, 70, 849-911.

Fan, J. and Li, R. (2006), "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery," *Proceedings of the International Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III, European Mathematical Society, Zurich*, 595–622.

Helland, I. S. (1988), "On the structure of partial least squares regression," *Communications in Statistics – Simulation and Computation*, 17, 581–607.

— (1990), "Partial least squares regression and statistical method," *Scandinavian Journal of Statistics*, 17, 97–114.

Li, K.-C. (1991), "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, 86, 316–327.

Li, L., Cook, R. D., and Tsai, C. L. (2007), "Partial inverse regression," *Biometrika*, 94, 615–625.

Luo, R., Wang, H., and Tsai, C. L. (2009), "Contour projected dimension reduction," *The Annals of Statistics*, To appear.

Naik, P. and Tsai, C. L. (2000), "Partial least squares estimator for single-index models," *Journal of Royal Statistical Society, Series B.*, 62, 763–771.

Nguyen, D. V. and Rocke, D. M. (2002), "Tumor classification by partial least squares using microarrray gene expression data," *Bioinformatics*, 18, 39–50.

Paul, P., Bair, E., Hastie, T., and Tibshirani, R. (2008), "Preconditioning for feature selection and regression in high-dimensional problems," *The Annals of Statistics*, 36, 1595–1618.