Accordingly we will have hundreds of competitive explanatory variables. Under such a circumstance, when we incorporate competitive information, we face the problem of too many explanatory variables ([Martin and Kolassa 2009](#_ENREF_3)). Time series models can easily get over-fitted and generate poor forecasts and in an extreme case cannot even be estimated because of more explanatory variables than observations. Therefore a mechanism is needed to identify, select, and refine the most relevant competitive explanatory variables ([Castle, Doornik et al. 2008](#_ENREF_1)). In this paper we propose a forecasting method which incorporates competitive information in forecasting retailer product sales at the UPC level. Methodologically our research propose an effective forecasting method which solve the problem of too many explanatory variables, an issue of theoretical and practical significance in a world of ‘big data’. More importantly our research offer an operational guidance to the retail forecaster as to how to produce more accurate forecasts as simply as possible.

In this study, we first develop conventional forecasting models following the two-stage modelling strategy in Huang, Fildes et al. (2014). The model consists of two stages since it incorporates the promotional information for both the focal product and other competitive products within the same product category. The first stage is to refine the competitive promotional information. In practice, retailers may have hundreds of items (thus hundreds of promotional variables for them) at the SKU level for each product category. It is not possible to incorporate the promotional variables of all these products. Thus we need to identify the most relevant competitive explanatory variables. in this study, we implement the Least Absolute Shrinkage and Selection Operator (LASSO) following Huang, Fildes et al. (2014) and [Ma, Fildes et al. (2016)](#_ENREF_2). The LASSO algorithm was developed by [Tibshirani (1996)](#_ENREF_5) as an alternative to traditional selection procedure such as the most popular stepwise selection. It estimates a regression model including all the potential explanatory variables but put a constraint, usually determined by information criterion, on the sum of the absolute values of all the parameter coefficients. As the constraint is employed, some of the parameter coefficients will be pushed to zero, and the corresponding explanatory variables are removed from the regression model. We also implement the other strategy in Huang, Fildes et al. (2014) and [Ma, Fildes et al. (2016)](#_ENREF_2). We implement the principle component analysis to pool information across all the competitive explanatory variables and condense them into a small number of diffusion factors ([Stock and Watson 2002](#_ENREF_4)).

Castle, J. L., et al. (2008). "Model Selection when there are Multiple Breaks." Working paper No. 407, Economics Department, University of Oxford.

Ma, S., et al. (2016). "Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information." European Journal of Operational Research **249**(1): 245–257

Martin, R. and S. Kolassa (2009). Challenges of Automated Forecasting in Retail. International Symposium on Forecasting, Hong Kong.

Stock, J. H. and M. W. Watson (2002). "Macroeconomic Forecasting Using Diffusion Indexes." Journal of Business and Economic Statistics **20**.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological) **58**(1): 267-288.