Please cite this paper as:

Ma, S., Fildes, R. and Huang, T. (2014) (LUMS Working Paper 2014:9). Lancaster University: The Department of Management Science. Demand forecasting with high dimensional data: the case of SKU retail sales forecasting with intra- and inter-category promotional information

Lancaster University Management School

Working Paper 2014:9

Demand forecasting with high dimensional data: the case of SKU retail sales forecasting with intra- and inter-category promotional information

Shaohui Ma[a,1]

Robert Fildes[b]

Tao Huang[c]

[a] School of Economics and Management, Jiangsu University of Science and Technology, China, 212003

[b] Lancaster Centre for Forecasting, Lancaster University, UK, LA1 4YX

[c] Kent Business School, Kent University, UK, CT2 7PE

[1] Corresponding author. Tel.: +86 138 15179032. E-mail address: msh@tju.edu.cn; r.fildes@lancaster.ac.uk (R. Fildes); t.huang@kent.ac.uk (T. Huang)

Demand forecasting with high dimensional data: the case of SKU retail sales forecasting with intra- and inter-category promotional information

Abstract

In marketing analytics applications in OR, the modeler often faces the problem of selecting key variables from a large number of possibilities. For example, SKU level retail store sales are affected by inter and intra category effects which potentially need to be considered when deciding on promotional strategy and producing operational forecasts, but no research has put this well accepted concept into forecasting practice: an obvious obstacle is the ultra-high dimensionality of the variable space. This paper develops a four steps methodological framework to overcome the problem. It is illustrated by investigating the value of both intra- and inter-category SKU level promotional information in improving forecast accuracy. The method consists of the identification of potentially influential categories, the building of the explanatory variable space, variable selection and model estimation by a multistage LASSO regression, and the use of a rolling scheme to generate forecasts. The success of this new method for dealing with high dimensionality is demonstrated by improvements in forecasting accuracy compared to alternative methods of simplifying the variable space. The empirical results show that models integrating more information perform significantly better than the baseline model when using the proposed methodology framework. In general, we can improve the forecasting accuracy by 14.3 percent over the model using only the SKU's own predictors. But of the improvements achieved, 88.1 percent of it comes from the intra-category information, and only 11.9 percent from the inter-category information. The substantive marketing results also have implications for promotional category management.

Keywords: Analytics; OR in marketing; Forecasting; Retailing; Promotions

1. Introduction

Many marketing problems require the analyst to understand the interactions of a large number of potentially inter-related variables. For example, grocery retailers rely heavily on accurate sales forecasts at SKU level when making business decisions in a wide range of areas including marketing, production, inventory, and finance etc. Sales and promotional effects in any one SKU are potentially affected by marketing and sales activities in a large number of other categories – in other words, there are intra and inter-category variables that may affect the target variable(s). However, identifying important variables from such a large set of possibilities poses a serious modelling challenge- it is the subject of this paper.

In a retail forecasting system, product sales history, intra-category promotional schedules, and inter-category promotion schedules are all potential rich sources of information which may influence forecasting accuracy. When building product sales forecasting models, a series of related but fundamental questions must be answered: which sources of information should be inputted into the forecasting model? To what extent do different sources of information contribute to forecasting accuracy improvements? And critically, how to manipulate the high dimensional information to generate better forecasts?

The main challenge to be faced is that the dimensionality of promotional explanatory variables grows very rapidly when cross-product promotional information is considered, potentially much larger than the length of SKU sales history. The model may be easily over-fitted or even cannot be estimated. To build a forecasting model for a SKU, when considering both intra- and inter-category promotional interactions, the number of candidate explanation variables is usually in the order of tens of thousands. With high dimensionality, important predictors can be highly correlated with some unimportant ones, and the maximum spurious correlation also grows with dimensionality (Fan and Lv, 2008).

Traditional methods which deal with the problem of high dimensionality include the subset selection method, the penalized L-1 likelihood method, and the information summary approach. The subset selection method and the penalized L-1 likelihood method, which are of distinct mechanisms, both try to find out the most influential variables affecting the dependent variable. However, in the retail context, store managers may promote similar

products simultaneously (e.g. different SKUs under the same brand), which makes the price and promotional variables of different SKUs highly correlated to each other. As a result, these two methods may select some unimportant predictors which are highly correlated with the important predictors but fail to select the really important predictors. The information summary approach condenses the information of the vast number of variables (which we cannot directly use due to high dimensionality) into a small number of factors but at the cost of (potentially high) information loss.

To overcome the problem, a four step methodological framework is proposed in this paper which consists of the identification of potentially influential categories, the building of explanatory variable space, variable selection and model estimation by a multistage LASSO (Least Absolute Shrinkage and Selection Operator) regression, followed by a scheme to generate forecasts. The method breaks down the process of variables selection into three stages: 1) to select variables related to promotional history of the focal product; 2) to select intra-category variables and 3) finally, to select inter-category variables.

The development of a successful modelling system, necessarily automatic in order to deal with the large number of SKUs, would also allow retailers to simulate the expected results based on different promotional plans so that they can then optimize their promotional schedules (Levy et al. 2004, Zhang et al. 2008). The need for an effective modelling and forecasting system is therefore transparent. Existing studies in the literature have overlooked the inter-categorical variables because the available methods are incapable of effectively integrating the useful information contained by these variables, as we discuss in the next section.

In this paper, we focus on developing an automatic modelling approach which we validate by applying it to the problem of forecasting many thousands of retail SKUs in order to produce improved short term forecasts. Through a series of empirical data experiments, we show that the proposed method of variable selection is an effective approach to simplifying the dimensionality of the promotional marketing space: it improves forecasting accuracy significantly by simplifying and integrating more retail information. But generally, the inter-category information contributes limited accuracy improvements comparing to that of intra-category information.

The outline of the paper is as follows. In section two, we review existing related studies and address their limitations. In Section three we discuss methodological issues. Section four describes the data, introduces the experimental design and forecasting accuracy measures, and presents the empirical results. Section five discusses the findings, offering conclusions as to forecasting practice and further academic research.

2. Literature review

2.1 Model building with a large number of explanatory variables

We are in the era of massive automatic data collection, systematically obtaining many measurements, not knowing which ones will be relevant to the phenomenon of interest (Donoho, 2000). Traditional statistical methodology assumed many observations and a few well-chosen variables are not designed to cope with this kind of explosive growth of dimensionality of the observation vector. The increasing availability of data is thus creating new challenges for the market modeller. There are, essentially, three different approaches to address this problem. The first approach is concerned with finding the most influential subset of predictors; the second approach builds predictive models based on summaries of the predictor variables and the third approach is penalized (L-1) likelihood method which automatically selects significant variable via continuous shrinkage.

Best subset selection is a popular class of the dimension reduction methods concerned with finding the most influential subset of predictors in predictive modeling from a much larger set of potential predictors. The best subset problem belongs to the class of NP-hard problems known as induction of minimal structures (John et al., 1994). When the number of potential predictors is large, the selection process cannot be solved exactly with an acceptable amount of computation time. Consequently, heuristic optimization algorithms have evolved, including iterative improvement algorithms (e.g., stepwise regression, forward and backward feature selection algorithms) and stochastic search methods (e.g., Genetic algorithms (Melab et al., 2002), simulated annealing (Meiri and Zahavi 2006)), to solve larger scale combinatorial problems. However, the expensive computational cost still makes best subset selection procedures infeasible for high-dimensional data analysis.

The information summary approach to forecasting with high dimensional data is based on the assumption that the relevant information is captured by a small number of factors common to the predictor variables. A popular technique that combines the potentially relevant predictors into new predictors is principal components. For example, basing the forecast model on data summaries in the form of principal components, as in Stock and Watson (2002), allows information from all the predictors to enter into the forecasts. Stock and Watson (1999), Stock, Watson, and Marcellino (2003), and Forni et al. (2000, 2003), among others, all find that diffusion factors based forecasts have smaller mean-squared errors than forecasts based upon simple autoregressions and more elaborate structural models. A criticism of factor augmented regressions is that the factors are estimated without taking into account the dependent variable. Thus, when only a few factors are retained to represent the variations of whole explanatory variable space, they might not have any predictive power of the dependent variable whereas the discarded factors might be useful (Stock and Watson, 2002).

Penalized L-1 likelihood methods have been successfully developed over the last decades to cope with high dimensionality. They have been widely applied for simultaneously selecting important variables and estimating their effects in high dimensional statistical inference. Penalized L-1 regression is called the LASSO by Tibshirani (1996) in the ordinary regression setting which has received much attention due to its convexity and encouraging sparsity solutions. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. There has been much work in recent years, applying and generalizing the LASSO and L1-like penalties to a variety of problems (Tibshirani 2011). Efron et al. (2004) propose a fast and efficient least angle regression (LARS) algorithm for variable selection, a simple modification of which produces the entire LASSO solution path. A linear combination of L-1 and L-2 penalties is called an elastic net by Zou and Hastie (2005), which encourages some grouping effects. Zou (2006) introduced an adaptive lasso in a finite parameter setting. Meier et al. (2008) proposed a fast implementation for the group LASSO.

L-1 type regularization does not eliminate the conflict between consistent model selection and prediction. With high dimensionality, important predictors can be highly correlated with some unimportant ones, and the maximum spurious correlation also grows with dimensionality (Fan and Lv, 2008). But LASSO tends to arbitrarily select only one

variable among a group of predictors with high pairwise correlations. This may results in some unimportant predictors that are highly correlated with the important predictors being selected by LASSO while important predictors are missed.

2.2 Intra- and inter-category promotional effects

The idea that demand in one product category can be affected by marketing efforts in another is not new. In economics, products are considered complements (substitutes) if lowering (raising) the price of one product leads to an increase in sales of another (Nicholson, 1998). Product substitutability and complementarity have long been natural ways to perceive inter-category relationships.

Within one category, products of different brands, even the same brand in different flavors or different pack sizes are usually regarded as substitute for each other. A large body of research supports the view that brands within a product category are substitutes for one another (Frank and Massy, 1967; Kumar and Leone, 1988; Moriarty, 1985; Mulhern and Leone, 1991; Walters, 1988, 1991). Researchers have found that the majority of the promotional response stems from brand switching, the percentage of own-brand sales elasticity with respect to a particular promotion that is due to brand-switching elasticity is about 75%-84% (Gupta, 1988; Chiang, 1991; Chintagunta, 1993; Bucklin et al., 1998).

For many categories, consumer purchasing patterns are also affected by stimulating purchases of nonpromoted complements to the promoted products (Berman and Evans, 1989; Walters, 1988). For example, the promotion of a pie filling may stimulate sales of full-margin pie shells, or the promotion of taco shells may increase sales of nonpromoted taco sauce. In such cases, one promotion can increase the sales of products in two different categories. In addition, inter- rather than intra-product substitution may also be the predominant influence in certain product groups. Walters (1991), using store level SKU sales data, tested a conceptual framework for retail promotion effects that includes brand substitution effects, inter-store sales displacements and the purchase of complementary goods. He selected four product categories in his study (spaghetti, spaghetti sauce, cake mix and cake frosting) and found that both the complementary effects of promotion and the substitution effects of promotion on brand sales are significant. Bandyopadhyay (2009) proposed a dynamic model based on vector autoregression (VAR), and empirically studied intra- and inter-category

promotional effects with four brands of ice cream, two brands of topping, and three brands of frozen yogurt. He found that a multiple-category model that includes brands from substitute and complementary categories returns more accurate sales forecasts than does a single-category model that includes brands from only a single category. Hruschka (2013) analyzed multi-category buying decisions of households by a finite mixture of multivariate Tobit-2 models. He found 18% of all pairwise category correlations are significant. Studies also showed that the cross-category impact of national brands on store brands appears to be substantially greater than that of store brands on national brands (Wedel and Zhang 2004). This means that the promotional effects are asymmetrical which are not only within but also across categories.

Though existing research has provided evidences that the promotions of one product can influence the sales of another from both intra- and inter-categories, most of the existing literature has focused on developing explanatory models, using a set of ad hoc assumed product relationships to test the significance of the cross brand/category promotional affects. Whether these theoretical findings can be applied in a real forecasting system to help retailers improving the decision accuracy at SKU level is the question we concern ourselves with in this research. This is a very different problem than those only concerned with explanation and hypothesis testing. When we build forecasting models for tens thousands of SKUs in a store, a problem size many retailers face, most of these existing theoretical models lose their feasibility. For example, in a VAR model, the number of free parameters increases quadratically with the number of variables in a system, and for even moderately-sized systems the model becomes highly overparameterized relative to the number of available observations. Even basic least square regression will not be applicable because the dimensionality of cross category promotion explanatory variables is potentially much larger than the sample size. In practice, we also cannot easily identify which product complements/substitutes another. For example, beer and carbonated beverages could be either substitutive or complementary, for people could drink them at different times in a day. And even if we can specify a group of product categories within which possibly exists promotional interactive effects (no matter whether complementary or substitutive), we still cannot easily specify which products in these categories interact with each other.

2.3 SKU sales forecasting

The basic SKU sales methods are univariate forecasting models which are based on time series techniques that analyze past sales history in order to extract a demand pattern that is projected into the future (Raju, 1995; Ord and Fildes, 2013). The techniques range from the simpler moving averages and exponential smoothing family to the more complicated Box–Jenkins ARIMA approach, or the Exponential smoothing state space class of model (Hyndman et al. 2002; Taylor, 2007). The methods do not take external factors such as price changes and promotions into account (Alon, Qi & Sadowski, 2001). Gür Ali et al. (2009) found that the simple time series techniques perform well for periods without promotions. However, for periods with promotions, models with more inputs improve accuracy substantially. Therefore, univariate forecasting methods are usually adopted as a benchmark model in many studies (Gür Ali et al., 2009; Huang et al., 2014).

In order to improve SKU sales forecasting in the presence of promotions, many studies have integrated the focal product's promotional variables into their forecasting models. In practice, many retailers use a base-times-lift approach to forecast product sales at the SKU level (Cooper et al., 1999; Huang et al., 2014). The approach is a two-step procedure which initially generates a baseline forecast from a simple time series models and then makes adjustments for any incoming promotional events. The adjustments are estimated based on the lift effect of the most recent price reduction and/or promotion, and also the judgements made by brand managers (Fildes et al., 2008; Fildes et al., 2009). Another stream of studies uses a model-based forecasting system to forecast product sales by directly taking into account the promotional information. These methods are usually based on multiple regression models or data mining technologies whose exogenous inputs correspond to the focus product's own promotion features (Rinne and Geurts, 1988; Preston and Mercer, 1990; Cooper et al., 1999; Kuo, 2001; Aburto and Weber, 2007; Gür Ali et al., 2009). For example, in Cooper et al. (1999), a promotion-event forecasting system called PromoCast is reported, which uses a static cross-sectional regression analysis of SKU-store sales under a variety of promotion conditions, with store and chain specific historical performance information. The limitation of these studies is they overlook the potential importance of price reductions and promotions of other influential products.

Forecasting product sales integrating influential products' promotional information has

also been explored by previous researchers. A well know example is the SCAN*pro model and its extensions which decompose sales for a brand into own- and cross -brand effects of price, feature advertising, aisle displays, week effects, and store effects (Wittink et al., 1988; Foekens et al., 1994; Van Heerde et al., 2000, 2001, 2002; Andrews et al., 2008). CHAN4CAST was another well-known forecasting model which was developed by Divakar et al. (2005). They also employed a regression model capturing the effects of such variables as past sales, trend, own and competitor prices and promotional variables, and seasonality. In recent research, Huang, Fildes and Soopramanien (2014) proposed effective methods to forecast retail SKU sales by incorporating competitive information including prices and promotions. They found that the proposed methods generate substantially more accurate forecasts across a range of product categories.

These research studies have made significant contributions to a burgeoning literature on improving product sales forecasting by integrating more information. However, these studies have limitations. First, though models such as SCAN*pro, theoretically considered both the substitutive and complementary effects, very little past research has empirically considered the promotional interactive effects in a grocery forecasting system that can work in practice.. In CHAN4CAST (Divakar et al., 2005), the forecasting system they built is for consumer packaged goods companies like PepsiCo and Kraft Foods whose goods are sold through multiple channels in multiple geographic regions. They only empirically considered the promotional interaction among two beverage brands (Coke and Pepsi). As Cooper et al. (1999) pointed out "*the planning test for retailers is very different from that of manufacturers. A broad line for a manufacturer may have hundreds of SKUs that could be promoted. This is small compared to planning for the 30,000 items that are in stock at any given time for a retailer."* An exception is Huang, Fildes and Soopramanien (2014). Using the weekly data from a large U.S. retail chain, they included within category competitive (substitutive) promotional information into their Autoregressive Distributed Lag (ADL) model and empirically checked the forecasting improvements compared to the model without competitive information. The key similarities of this study and that by Huang et al. (2014) are that both studies aim to improve the forecasting accuracy for retailers at SKU level by integrating extra promotional information from other products. At the same time, there are some important differences. First, this paper considers both intra- and inter-category promotional interactions, while Huang et al. (2014) only considered intra-category competition. Second, Huang et al. (2014) used a "general to specific" approach to manually

select explanatory variables for every SKU one by one. Though theoretically showing that integrating intra-category competitive information could improve SKU forecasting accuracy, the approach is in fact inapplicable: in a real grocery forecasting system, it is impossible to manually manipulate individual forecasting models for tens thousands of items in a store. Instead, we propose to use a multi-stage variable selection and model estimation strategy based on LASSO regression; the total process is fully automatic and therefore can be easily integrated into a forecasting system. Third, Huang et al. (2014) pooled the SKU sales from 83 stores to simulate a chain level forecasting situation. This does not help a chain manager allocate SKU stocks at the store level, because of the heterogeneity among stores. Furthermore, the price and promotional indexes are both aggregated across multiple stores; this may weaken the explanatory power of these variables. In our research, we focus on store level sales forecasting, using the raw SKU level information to build a forecasting model without any aggregation. This is the forecasting situation directly links to a chain or store manager's weekly stocking allocation decisions. But this is a more challenging problem, for the data at the disaggregate level contains more noise than at the aggregate level. Fourth, Huang et al. (2014) considered 122 SKUs from 6 categories in their empirical study. It is a large scale empirical study compared to previous existing researches; most of them usually consider only tens of items in empirical study. This research empirically examines the forecasts on 926 SKUs in 15 categories for 80 weeks out of sample forecasting. At such a scale, we need to weigh the complexity of the model and the corresponding computing efficiency. Therefore, our results will be more realistic, robust and useful in SKU level decisions.

To summarize, this research is innovative in four respects:

i. The development of a novel fully-automatic algorithm that is capable of selecting key explanatory variables from a very large data set.

ii. The focus on retail store level modelling and forecasting at SKU level for thousands of products in order to capture dynamic promotional effects.

iii. The inclusion of both intra- and inter-category information.

iv. The examination of comparative results for a large number of SKUs over a large number of categories.

3. Methodology

When cross-category promotional information is considered, the dimensionality of the promotional explanatory variables grows very rapidly. For example, if the sales of a product is potentially affected by promotions of items in *c* categories, each category includes *i* items, and each item has *j* promotion tools, then there are approximately *c*i*j* potential variables in explanatory variable space (e.g. for the 10 category this may leads to 2000 predictor variables, for peanut butter the number of variables considered is 3222). A typical retailer usually has tens thousands of items stocked at any given time which are usually classified into hundreds of product categories. Obviously it is unreasonable and infeasible to assume a product is affected by all the products in all the categories in the store. The method proposed includes four steps which are illustrated in Figure 1. At step 1, we identify the promotional interactive relationships at category level by statistical tests. We propose to use a LASSO Granger causality test for such a purpose. At step 2, we prepare the explanatory variable set for every SKU based on the interactive categories we identified in the first step. Then we consider three separate approaches dealing with this high dimensional information: (i) extract only the five top sales products and use them as the representatives of the category, (ii) preprocess the information to lower the dimensionality by extracting diffusion factors, and (iii) input all the raw SKU level promotional information directly into the subsequent LASSO regression. At step 3, to deal with the high dimensionality remaining in variable space, we propose a three-stage LASSO strategy to select important predictors and estimate the model parameters: these break down the variables into predictors from the SKU itself, intra category predictors and finally, predictors from other categories. At step 4, we generate forecasts for every SKU with the estimated models.

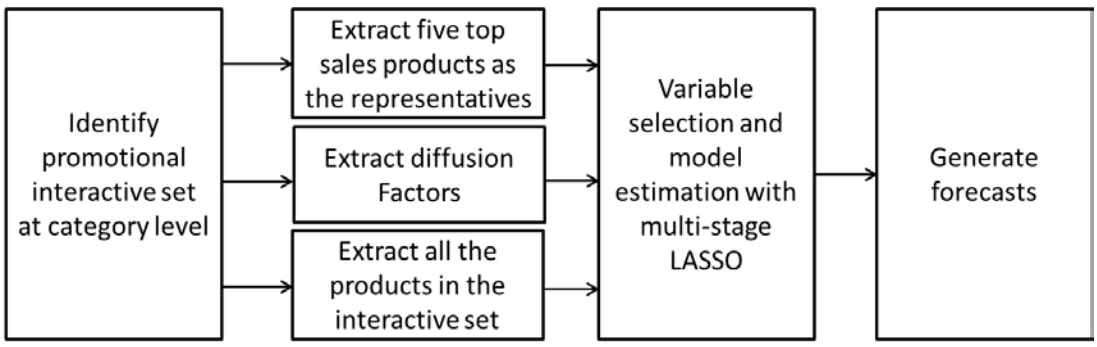

Figure 1 Methodology framework

3.1 Identifying the promotional interactions at category level

To identify which categories are promotional interactive with each other, a simple way is to resort to expertise by conducting a survey on retailing experts. But the approach is subjective and subject to the usual biases arising in judgmental decision making. Here we propose to use a LASSO Granger causality test directly to identity category level promotional interactions from product sales data.

Granger Causality testing is one of the earliest methods developed to quantify the causal effect from time series observations. It has gained success across many domains due to its simplicity, robustness, and extendibility (for example, Hiemstra and Jones, 1994). Ashley, Granger and Schmalensee (1980) gave the definition of causation as follows: Let Ω_t, represent all the information available in the universe at time t. Suppose that at time t optimum forecasts are made of Y_{t+1} using all of the information in Ω_t, and also using all of this information apart from the past and present values X_{t-j}, j>=0, of the series X. If the first forecast, using all the information, is superior to the second, than the series X has some special information about Y, not available elsewhere, and X is said to cause Y.

The main challenge in discovering causal relationship among product categories is the high dimensional time series need to be analyzed in this project. As the number of time series grows, the statistical significant tests become inefficient, leading to higher chance of spurious correlations. The LASSO Granger method we considered is one effective way to address such issue.

Specifically, we first build a set of promotional intensity indexes for every category, including price indexes, display intensity indexes, feature advertising intensity indexes. All of these indexes are calculated by weighted averaging the corresponding values across SKUs in a category. The weight is the weekly average sales of the SKU. That is, the larger the market share a SKU occupies in a category, the larger the weight it has in the calculation of promotional intensities. Second, we identify the promotional interactive set for every category one by one using LASSO-Granger algorithm (Arnold et al. 2007). In particular, this can be achieved by solving the following optimization problem:

$$\min_{\mathbf{a}} \sum_{t=1}^{T} \left\| Y_k(t) - \sum_{i=1}^{C} \mathbf{a}_i \mathbf{X}_i(t) \right\|_2^2 + \lambda \left\| \mathbf{a} \right\|_1 \quad (1)$$

where $\mathbf{X}_i(t)$ is the set of promotional intensity indexes in category i at time t; Y_k is the average sales in category k; T is the time length used for the test and C is the total number of categories considered; $\mathbf{a}$ is a coefficient vector to be minimized; λ is a nonnegative penalty parameter which determines the sparseness of $\mathbf{a}$. The optimal value of λ is determined by leave-one-out cross-validation in our empirical study. Finally, we determine that the promotions in category i cause the sales in category k if and only if $\mathbf{a}_i$ is a non-zero vector.

3.2 Building the explanatory variable space for forecasting

To build a forecasting model for jth SKU in product category k, three sets of information make up the potential explanatory variable space $\Omega_{kj} = \{S_{kj}^{own}, S_{kj}^{\text{intra}}, S_{kj}^{\text{inter}}\}$. The information set S_{kj}^{own} includes all the SKU$_{kj}$'s own promotional information, its sales history and time events (e.g., holidays). The information set S_{kj}^{intra} includes all the promotional information as well as sales history of SKUs in the category k. Similarly, S_{kj}^{inter} includes all the information of SKUs from identified interactive categories.

Considering the high dimensionality of the potential explanatory variable space, to utilize the information effectively, we test three approaches in this research. The first is to extract the information from the five best sale products in the category and use the information to represent that of the whole category. While the merit of this approach is that it is easy to implement and less computationally complex, it neglects a large part of the potentially useful information from other SKUs in the category. The second approach is to perform a Principle Component Analysis (PCA) on promotional variables to extract a few "factors" as representative of the whole category sale (Harrell, 2001; Stock and Watson, 2004; Huang et al. 2014). The method utilizes the variance-covariance structure of the predictors with the goal of finding a few linear combinations of the predictors to explain the covariance structure. In the empirical study, each explanatory variable, i.e. sales lag, price, display and feature, across SKUs in the same category is regarded as a cluster. For each cluster, we conduct PCA dynamically and extract m Principle Components (PCs). The PCA is an effective approach to lower the variable dimensionality, but it has a drawback in forecasting applications. Eigen-vectors corresponding to large eigenvalues are retained whereas those associated with

small eigenvalues are discarded. Thus, the retained factors might not have any predictive power of the dependent variable whereas the discarded factors might be useful (Stock and Watson, 2002). Here we conduct PCA dynamically as the inputs to the proposed multistage LASSO. This combines the merit of PCA which is effective in dealing with collinearity and LASSO which is good at variable selection in high dimensional space while make up their drawbacks. The final approach we considered is to input all the raw information as potential explanatory variables without any preprocessing. Obviously, this approach keeps all the potential useful information without any loss, but the high dimensionality in variable space leads to a high computational burden in the steps that follow.

3.3 Variable selection and model estimation with multistage LASSO regression

The main challenge to be faced is that the dimensionality of the promotional explanatory variables space grows very rapidly when cross-product promotional information is considered, potentially much larger than the length of the SKU sales' time series. In order to reduce the dimensionality from a huge scale effectively and efficiently, a multistage penalized likelihood method based on the LASSO penalty is applied to perform the variable selection and parameter estimation simultaneously. The LASSO is a regularization technique for simultaneous estimation and variable selection (Tibshirani, 1996) which continuously shrinks the coefficients toward 0 as the penalty increases, and some coefficients are shrunk to exact 0 if the penalty is sufficiently large. Moreover, continuous shrinkage often improves the prediction accuracy due to the bias variance trade-off.

With high dimensionality, important predictors can be highly correlated with some unimportant ones, and the maximum spurious correlation also grows with dimensionality (Fan and Lv, 2008). But LASSO tends to arbitrarily select only one variable among a group of predictors with high pairwise correlations. This may results in some unimportant predictors that are highly correlated with the important predictors being selected by LASSO while important predictors are missed. In a retailing store, it is very common to promote a set of products during the same period of time, especially during some special events. This results in the promotion explanatory variables from different SKUs being highly correlated which makes it difficult to distinguish their individual effects on the dependent variable. But from existing researches (Bucklin et al., 1998; Huang et al., 2014), we can know that a SKU's own promotion explanatory variable are more important than that of other SKUs, and the

promotions of SKUs in the same category as the focus SKU are more important than that of SKUs in other categories. If we input all the candidate explanatory variables simultaneously into a LASSO selector, it is likely to select poor variables, i.e., LASSO may select correlated products' promotion variables instead of the focal SKU's own predictors. We solve this problem by proposing a multistage LASSO regression strategy which is illustrated in Figure 2.

Specifically, in order to generate an h weeks ahead forecasting for SKU j in category k, the variable selection and parameter estimation process is divided into three stages. At the first stage, only the focal SKU j's own predictors are inputted into a LASSO regression, including sales lag, price, display, feature advertising and their lags, calendar events and week indicators, which can be modeled as an Autoregressive Distributed Lag (ADL) model (Huang et al., 2014),

$$\ln(Y_{kj,t+h}) = \eta_{kj} + \sum_{l=0}^{L}\left[\alpha_{kjl}\ln(Y_{kj,t-l}) + \beta_{kjl}\ln(P_{kj,t+h-l}) + \sum_{r=1}^{2}\gamma_{kjrl}D_{kjr,t+h-l} + \sum_{r=1}^{2}\rho_{kjr}F_{kjr,t+h-l}\right] + \sum_{d=1}^{12}\theta_{kjd}W_{t+h}^{d} + \sum_{c=1}^{9}\sum_{v=0}^{1}\delta_{kjc}C_{t+h-v}^{c} + y_{1kj,t+h} \quad (2)$$

where

$\ln(Y_{kj,t+h})$ is the log sales of the focal product j in category k in week $t+h$;

η_{kj} is the product j's specific constant;

$\ln(P_{kj,t+h})$ is the log price of the product j in category k in week $t+h$;

$D_{kj1,t+h}$ is an indicator variable for minor display: 1 if product j is minor displayed, in week $t+h$; 0 otherwise;

$D_{kj2,t+h}$ is an indicator variable for major display (including codes lobby and end-aisle in our empirical data): 1 if product j is major displayed in week $t+h$; 0 otherwise;

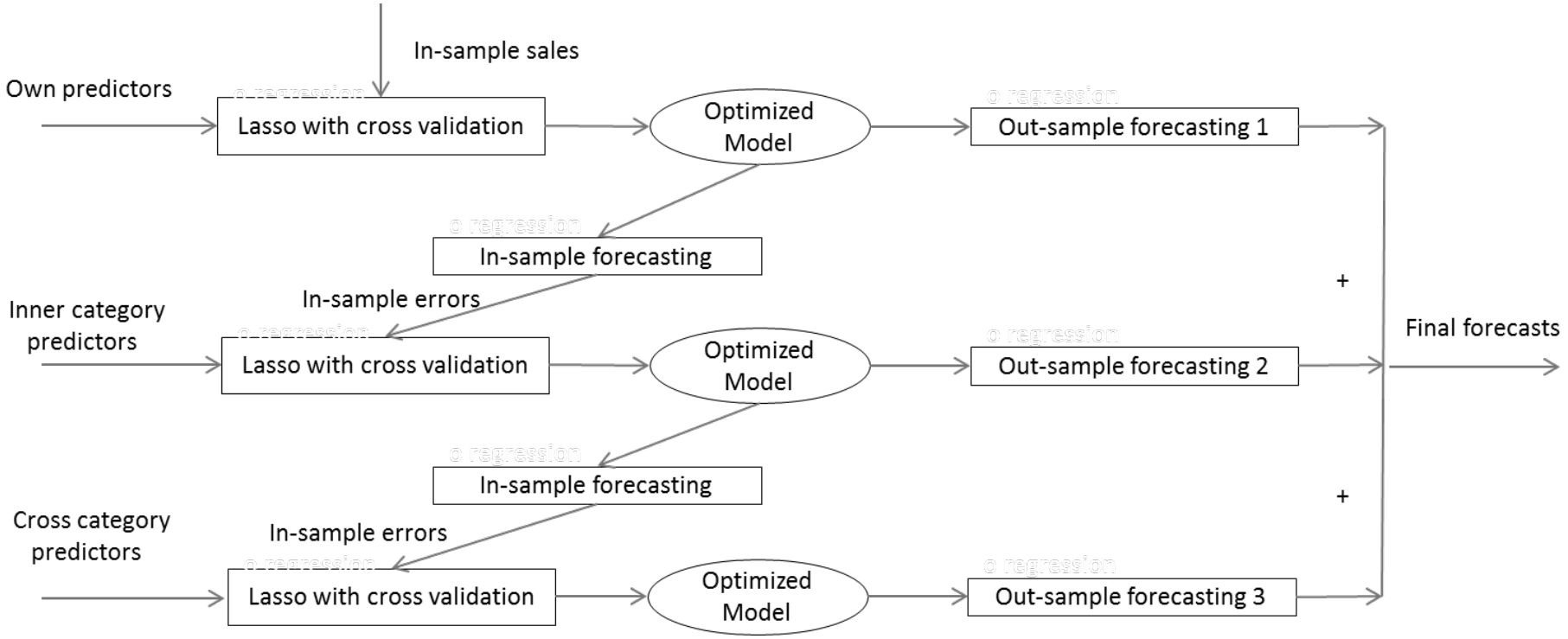

Figure 2 Multistage LASSO process

$F_{kj1,t+h}$ is an indicator variable for minor feature (small and medium size advertisement): 1 if product j is minor featured, in week $t+h$; 0 otherwise;

$F_{kj2,t+h}$ is an indicator variable for major feature (large size advertisement and retailer coupon or rebate): 1 if product j is major featured, in week $t+h$; 0 otherwise;

W_{t+h}^{d} is the d^{th} four-week-dummy variable: 1 if $t+h$ is in four-week d of the year; 0 otherwise;

C_{t+h-v}^{c} is the dummy variable for the c^{th} calendar event at week $t+h-v$. When v=0, the dummy variable represents the week of the calendar event, and the week before the event if v=1; c take the values from 1 to 9 representing all the calendar events including Halloween, Thanksgiving, Christmas, New Year's Day, President's Day, Easter, Memorial Day, 4th of July, and Labour Day.

L is the order of lags to be included which is assumed to be one in our empirical study. The α_{kjl} is the multiplier for sale lag of product j in category k, the β_{kjl} is the price elasticity, the γ_{kjrl} is the display multiplier, the ρ_{kjrl} is the feature multiplier, θ_{dj} is the four-week indicator multiplier, δ_{cj} is the calendar multiplier for event c, and the disturbance term is represented by $y_{1kj,t+h}$. It is worth noting that the promotional variables are assumed known to the retailer at $t+h$ in our model, as they usually form part of an agreed promotional plan with suppliers.

Assuming the data in time window [1, t] is used for model estimation, after variable selection and parameter estimation by LASSO regression, we calculate the in-sample forecasts error $\hat{y}_{1kj,1:t}$ and then generate out-sample forecasts in this first stage.

At the second stage, we use the in-sample forecasts error $\hat{y}_{kj,1:t}$ from the first stage as the dependent variable, and use the variables from other SKUs in the same category with the focal SKU as the explanatory variables and model the second stage forecasts by

$$y_{1kj,t+h} = \sum_{i=1,i\neq j}^{n_k}\left[\alpha_{kj}^{i}\ln(Y_{kit}) + \beta_{kj}^{i}\ln(P_{ki,t+h}) + \sum_{r=1}^{2}\gamma_{kjr}^{i}D_{kir,t+h} + \sum_{r=1}^{2}\lambda_{kjr}^{i}F_{kir,t+h}\right] + y_{2kj,t+h}\,, \quad (3)$$

where n_k is the number of SKUs (or factors extracted from PCA) in category k and the disturbance term is represented by $y_{2kj,t+h}$. In the model, if the inputs are factors extracted from PCA, then the variables Y, P, D and F in the model represent the corresponding factors. At this stage, variable selection and parameters estimation are again done by LASSO

regression. The in-sample forecasts error $\hat{y}_{2kj,1:t}$ and out-sample forecasts can then be calculated for the second stage.

At the third stage, the in-sample forecasts error $\hat{y}_{2kj,1:t}$ from the second stage are used as the dependent variable, and the variables from SKUs in the identified influential categories are used as the explanatory variables,

$$y_{2kj,t+h} = \sum_{s\in S_k}\sum_{i=1}^{n_s}\left[\alpha_{kj}^{si}\ln(Y_{si,t}) + \beta_{kj}^{si}\ln(p_{si,t+h}) + \sum_{r=1}^{2}\gamma_{kjr}^{si}D_{sir,t+h} + \sum_{r=1}^{2}\lambda_{kjr}^{si}F_{sir,t+h}\right] + \zeta_{kj,t+h} \qquad (4)$$

where S_k is the influential category set of category k and the disturbance term is represented by $\zeta_{kj,t+h}$ .We calculate the out-of-sample forecasts for the third stage. The final out-sample forecasts are the sum of the forecasts in the three stages.

4. Empirical study

4.1 Data

The empirical data comes from the IRI dataset (Bronnenberg et al., 2008)[2]. The IRI dataset includes grocery and drug chain data from a sample of stores in 50 markets and 30 categories, involving approximately 25%-30% of the consumer packaged goods sales in a grocery store. This is weekly data by SKU and includes information on sales, price, features and displays. Based on the objectives of this research, the records from a medium size grocery store in Chicago as the empirical sample were selected for fifteen product categories concerned with food and drink. Low-movement SKUs or SKUs which may have been introduced or discontinued were excluded. Our criterion was that at least 80% of the weeks must have positive movement for the SKU. The empirical dataset includes the weekly units sold, prices, displays and features of 926 SKUs in 15 food categories for 320 weeks.

Table 1 presents the means and medians of units sold per week and percentages of weeks concerning promotional activities, including price reductions (more than 5 percent), displays and features across fifteen categories. It is clear that the price reduction is the most frequent

[2] All estimates and analyses in this paper based on Information Resources, Inc. data are by the author and not by Information Resources, Inc.

type of promotion across all the categories. Feature advertising is also frequently used in many categories, such as frozen pizza and carbonated beverages. Display is only used occasionally for most of the categories except beer.

Table 1 Description statistics of the data sample

No.	Category	Num of SKUs	Mean units sold per week	Median units sold per week	Percentages of weeks concerning promotional activities		
					Price reductions	Displays	Features
1	Beer	98	12.80	7	0.30	0.27	0.13
2	Carbonated beverages	76	38.25	16	0.42	0.09	0.18
3	Coffee	46	5.90	5	0.34	0.02	0.10
4	Cold cereal	119	15.60	9	0.20	0.05	0.13
5	Frozen dinners	79	18.50	13	0.43	0.04	0.17
6	Frozen pizza	62	21.05	14	0.47	0.10	0.31
7	Frankfurters	21	22.95	10	0.35	0.08	0.16
8	Margarine/Butter	21	29.20	13	0.37	0.05	0.13
9	Mayonnaise	17	15.70	12	0.21	0.03	0.08
10	Milk	40	59.60	24	0.19	0.01	0.06
11	Peanut butter	16	14.30	10	0.22	0.01	0.07
12	Salty snacks	80	17.95	11	0.31	0.12	0.12
13	Soup	129	15.05	9	0.23	0.03	0.10
14	Spaghetti sauce	70	9.40	7	0.38	0.03	0.11
15	Yogurt	52	49.45	37	0.29	0.01	0.08

As an initial analysis, we can use prior experience to suggest some potential relationships among categories. For example, substitution might exist in beer and carbonated beverages, while carbonated beverages and salty snacks, milk and coffee, frozen pizza and beer etc., might be complementary. But for some categories, e.g., milk and yogurt, frozen pizza and coffee, it is difficult to identify by prior experience alone whether an interactive relationship between them is likely. We therefore resort to proposed LASSO Granger to empirically identify interactions among categories.

4.2 Empirical models

We estimate, for each SKU in the sample, eight alternative models which are explained in detail as the following.

(1) ETS. ExponenTial Smoothing state space model with seasonality and non-damped trend (Hyndman et al., 2002).

(2) ADL-own. ADL Model based on Eq. (2) with only the focal SKU's own predictors.

(3) ADL-intra-top5. ADL Model based on Eqs. (2) and (3) including the focal SKU's own predictors and predictors from the top five sales products in the same category.

(4) ADL-inter-top5. Similar to model (3) but also including the predictors of the top sales products from identified interactive categories.

(5) ADL-intra-all. ADL Model based on Eqs. (2) and (3) including the focal SKU's own predictors and predictors from all the products in the same category

(6) ADL-inter-all. ADL Model with predictors from all the SKUs in both intra- and inter-categories.

(7) ADL-intra-PCA(x). ADL Model including the focal SKU's own predictors and x principle components extracted by PCA from the same category. For example, if x=5, then for each set of promotional variables in the category we select 5 principle components by PCA.

(8) ADL-inter-PCA(x). Similar to model (7), but includes both intra- and inter-category principle components as explanatory variables in the model.

4.3 Forecasting evaluation

We use both scale-dependent and scaled error measures to compare the forecasting performance of the models. The first two criteria are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) which are traditional and popular scale-dependent error measures. They are easy to calculate, easy to understand and widely applied. They also have practical meanings to retailing managers, for they naturally place more weight on fast moving SKUs which usually contribute more revenues than slow moving items in a store. The third criterion is the Mean Absolute Scaled Error (MASE) which was proposed by Hyndman and Koehler (2006). It can be considered as a "weighted" arithmetic mean of the MAE based on the variations of the sales data in the estimation period (Davydenko and Fildes, 2013). It is

defined as

$$MASE = Mean\left(\frac{|e_t|}{\frac{1}{m-1}\sum_{i=2}^{m}|Y_i - Y_{i-1}|}\right) \quad (5)$$

where e_t is the forecast error at week t; m is the number of weeks in the estimation period, Y_i is the sale in week i. MASE is clearly independent of the scale of the data and very suitable for comparing the forecasts across multiple time series. The drawback of MASE is that it puts more weights to the data series which are comparatively stable, which makes it vulnerable to outliers. The last criterion we used is based on relative errors. The Average Relative Mean Absolute Error (AvgRelMAE) is proposed by Davydenko and Fildes (2013) for measuring forecasting accuracy at SKU-level demand. It is a geometric mean of the ratio of the MAE between the candidate model and the benchmark model.

$$AvgRelMAE = \left(\prod_{i=1}^{N}\frac{MAE_i^f}{MAE_i^b}\right)^{\frac{1}{N}} \quad (6)$$

where N is the number of SKUs in the sample, MAE_i^b is the MAE of the baseline statistical forecast for series i, MAE_i^f is the MAE of the candidate model f evaluated for series i. The AvgRelMAE has the advantages of being scale independent and robust to outliers, with a straightforward interpretation: a value smaller than one indicates an improvement by the candidate model over the benchmark.

4.4 Forecasting scheme for evaluation

All models are estimated for each SKU separately. We generate the forecasts with both a fixed forecasting scheme and a rolling scheme. For the fixed scheme, estimation of the models is based on the data of the first 240 weeks, and the remaining 80 weeks of data are used for forecasting evaluation. Although this is not likely to be used in practice, it helps us to evaluate a model's forecast performance over all observations of the validation sample. For the rolling scheme, we estimate the models with a moving window of 200 weeks and the forecast for one to four-week ahead horizons. The forecasting horizons are chosen to take into account typical ordering and planning periods. We move the estimation window forward week by week throughout the remaining sample period and we re-select variables and re-estimate the models based on the updated data sets. This differs from Huang, Fildes and

Soopramanien (2014) who used a fixed time window for manually variable selection and rolling windows for model estimation; here the models in this research are automatically re-specified for each rolling event based on each new moving time window. Thus, our forecasting procedure is an iterative one consisting of variable selection, model estimation, and forecasting throughout the forecasting subsample.

4.5 Results

4.5.1 Category level interactions

The 240 weeks calibration data is used to analyze the category level interactions. In Figure 3, a path diagram is presented to represent the Granger relationships among 15 selected product categories.

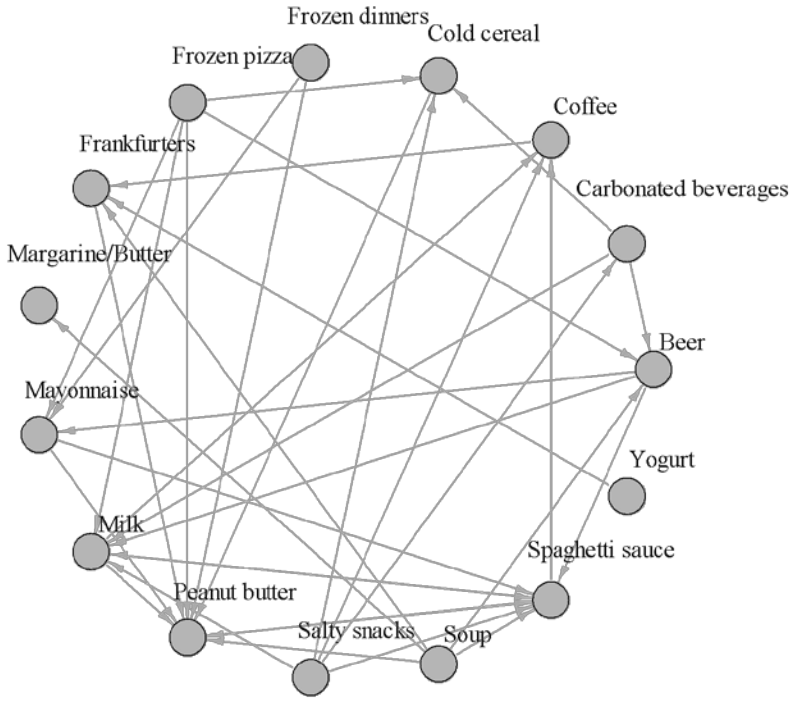

Figure 3 Promotional interactions at category level

Every category is represented by a node in the graphand there is a directed line from category X to Y if and only if X Granger causes Y. Following existing research on consumer cross category purchasing (Wedel and Zhang, 2004; Walters, 1991; Lee et al., 2013; Hruschka, 2013), we also find that the interactions between pair of categories are asymmetric

(as shown in Fig.3). For example, Carbonated beverage is affected by the promotion of Salty snack, but not vice versa. We can also find that some categories are easily affected by promotions in many other categories, such as Peanut butter and Spaghetti sauce, while some of them are more isolated, such as Soup and Frozen dinner.

4.5.2 Fixed scheme forecasts

The fixed scheme forecasting results are shown in the left panel of Table 2. The first row in Table 2 reports the results for the ETS model. This time series model delivers the worst forecasts among all the empirical models. The ADL-own model is used as a baseline model to calculate AvgRelMAE which is shown in the second row. All the forecasting measures for this model are substantially lower than the pure time series model which indicates the usefulness of the extra promotional information. The third row reports the results for the ADL-intra-top5 model which includes extra promotional information from the 5 top sales SKUs intra-category. The inclusion of extra information does not improve the forecasting accuracy. The ADL-intra-all model improves the baseline model slightly, while ADL-inter-all model fails to achieve better forecasts over the baseline. These results indicate that integrating more information does not necessarily improve the SKU sale forecasts under the fixed origin scheme. One possible reason is that the extent of promotional interactive effects among products are time varying and weak at individual SKU level, and the large amount of extra noisy information increases the risk of overfitting and therefore worsens the forecasts.

In the rows 7 to 12 of the Table 2, however, all the models based on principle components can significant improve the forecasts over the baseline model. But integrating more principle components in the model falls to improve the forecasts further. Model ADL-inter-PCA(3) which includes just three principle components for each promotion variable in a category provides the best forecasts for all the evaluation measures.

4.5.3 Rolling scheme forecasts

In the middle and right panel of Table 2 we report the results for one week ahead and four weeks ahead rolling forecasts. In contrast to the results from fixed scheme forecasts,

Table 2 The overall models' forecasting accuracy with different forecasting scheme and horizons

	Fixed scheme				Rolling scheme with horizon=1				Rolling scheme with horizon=4			
	MAE	RMSE	MASE	AvgRelMAE	MAE	RMSE	MASE	AvgRelMAE	MAE	RMSE	MASE	AvgRelMAE
ETS	8.027	19.095	0.811	1.116	8.136	19.395	0.822	1.195	8.141	19.397	0.822	1.195
ADL-own*	6.895	14.588	0.761	1	6.316	13.663	0.711	1	6.476	13.851	0.725	1
ADL-intra-top5	6.901	14.432	0.766	1.005	6.214	13.272	0.702	0.989	6.306	13.501	0.710	0.984
ADL-inter-top5	7.142	15.062	0.793	1.027	6.165	13.201	0.697	0.983	6.232	13.354	0.704	0.976
ADL-intra-all	6.816	14.522	0.756	0.997	6.120	13.121	0.692	0.978	6.159	13.318	0.697	0.969
ADL-inter-all	6.902	14.689	0.764	1.005	**6.092**	**13.015**	**0.690**	**0.975**	**6.146**	**13.245**	**0.696**	**0.967**
ADL-intra-PCA(3)	6.790	14.414	0.752	0.988	6.132	13.178	0.695	0.981	6.224	13.349	0.704	0.975
ADL-inter-PCA(3)	**6.752**	**14.255**	**0.750**	**0.986**	6.117	13.196	0.693	0.979	6.226	14.543	0.723	0.975
ADL-intra-PCA(4)	6.798	14.485	0.754	0.993	6.123	13.072	0.694	0.981	6.208	13.170	0.702	0.974
ADL-inter-PCA(4)	6.798	14.534	0.754	0.992	6.108	13.087	0.692	0.978	6.183	13.114	0.699	0.971
ADL-intra-PCA(5)	6.786	14.476	0.751	0.989	6.133	13.158	0.695	0.981	6.206	13.214	0.702	0.974
ADL-inter-PCA(5)	6.776	14.432	0.751	0.989	6.114	13.104	0.693	0.978	6.183	13.162	0.700	0.971

*ADL-own is the benchmark model used to calculate AvgRelMAE; bold text in the table means the best result in the column

first, all the models under the rolling scheme deliver substantially better forecasts than that of with fixed scheme. And more importantly, all the models integrating extra information, even only including extra information of the five top sales products, perform better than the baseline model. Second, we can see that models combining raw SKU level information benefit more from the rolling scheme than the models integrating promotional factors. The ADL-inter-all model outperforms all the other models across all measures. This is an astonishing result compared to its poor performance in the fixed scheme. The results confirm that the extent of promotional interactions among individual SKUs are unstable and dynamic across time periods. Third, the factor based models also perform pretty well though they are not the best. Considering they also perform well in the fixed schemes, we conclude that they are more robust models than the models without information pre-extraction. Fourth, by comparing the forecasting improvements between different horizons, we find the improvements over baseline model become substantially larger as the forecast horizon increases, e.g. the AvgRelMAE is 0.967 in horizon 4 weeks while it is 0.975 in horizon 1 week for the best performance model. This is in consistent with the results from Huang et al. (2014).

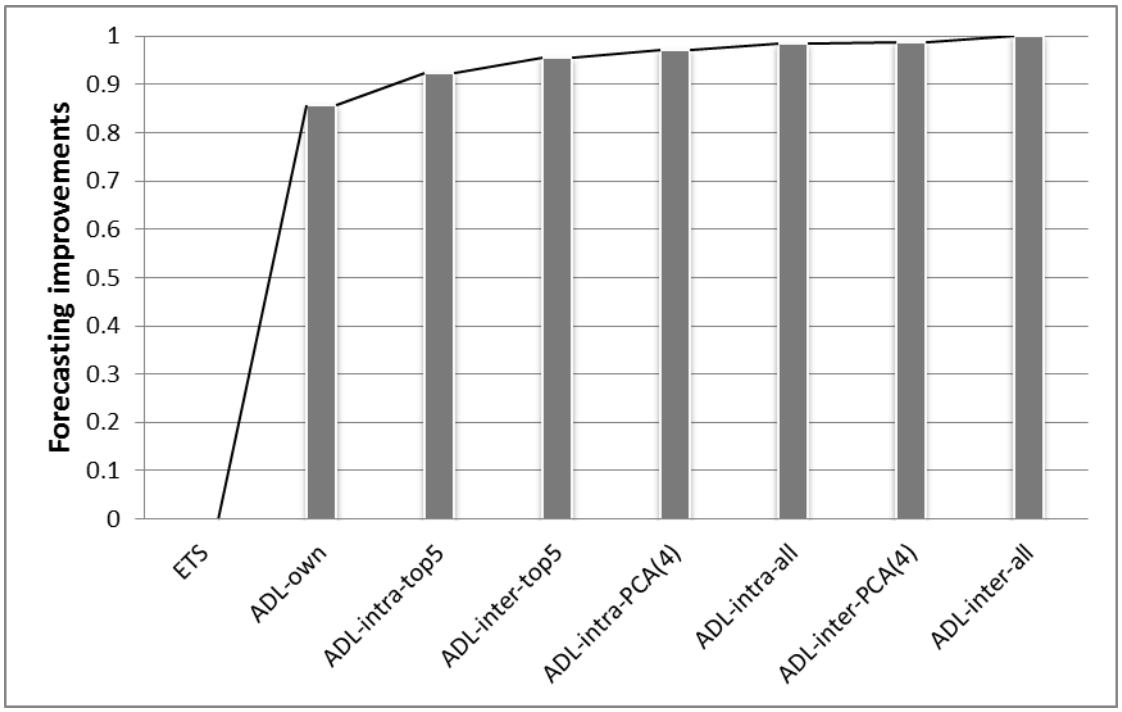

Figure 4 MAE improvements of the models with different information sets

Figure 4 shows the MAE improvements of the models with different information sets over ETS model. We only compare the MAEs in this figure and the following figures because the results from four measures are consistent with each other. The incorporation of the focal product's own predictors contributes 85.7% of improvements over ETS model. The extra information from the intra-category five top sales products contributes an additional 6.7%. The following extra information sets contribute less and less. The ADL-inter-PCA(4) can only improve over ADL-intra-all 0.5%, and the ADL-inter-all model improves over ADL-inter-PCA(4) only 1.3%. All the intra-category information at most contributes about 12.6% extra accuracy improvements over the own predictors model, while all the inter-category information contributes only 1.6% additional improvements.

In Table 3, we compare the forecasting results of three representative models, including ADL-own, ADL-intra-all and ADL-inter-all, for different categories individually. Those models are selected because they are the best preforming models with the three different information sets under the rolling scheme. The forecasts are averaged over forecasting horizon from one to four weeks in the table. In general, both ADL-intra-all and ADL-inter-all models consistently outperform the baseline model across all categories. But the extent of the improvements varies among different categories. Categories such as Cold cereal and Soup achieve limited forecasting improvements from extra information. In the category Frankfurters and Milk, however, both models improve the forecasts over the ADL-own model significantly.

In order to show the value of intra- and inter-category information at category level, in Figure 5, we illustrate the MAE improvements of ADL-intra-all and ADL-inter-all over ADL-own among different categories. In categories, such as Frankfurters, Margarine/Butter, Carbonated beverages, and Milk, the contribution from intercategory information was relatively large, ranging from 12% to 44%, compared with that in other categories. For Mayonnaise and Peanut butter, including the intercategory information in the model could even worsen the forecasts. An explanation is that the useful predictors from other intercategory may be too weak to compensate for the loss by including the extra volume of noisy information for these categories.

Table 3 The models' forecasting accuracy in various categories with weekly rolling scheme and 1-4 week ahead forecasting horizon

No.	Category	influential categories	ADL-own*			ADL-intra-all				ADL-inter-all			
			MAE	RMSE	MASE	MAE	RMSE	MASE	AvgRelMAE	MAE	RMSE	MASE	AvgRelMAE
1	Beer	2,6,13	4.431	9.305	0.904	4.196	8.720	0.853	0.951	4.189	8.678	0.852	0.950
2	Carbonated beverages	12	8.383	14.853	0.648	7.991	14.361	0.629	0.976	7.928	14.229	0.625	0.971
3	Coffee	10,12,14	2.211	3.685	0.693	2.173	3.571	0.687	0.989	2.170	3.556	0.687	0.988
4	Cold cereal	2,6,12	5.268	12.003	0.529	5.193	11.919	0.520	0.985	5.192	11.894	0.519	0.985
5	Frozen dinners	--	6.409	9.958	0.720	6.215	9.647	0.702	0.979	6.215	9.647	0.702	0.979
6	Frozen pizza	--	6.703	11.648	0.736	6.618	11.479	0.724	0.985	6.618	11.479	0.724	0.985
7	Frankfurters	3,13,15	9.554	26.206	0.452	9.364	25.954	0.446	0.983	9.214	25.050	0.442	0.971
8	Margarine/Butter	13	8.388	26.002	0.697	8.184	25.940	0.678	0.977	8.163	25.932	0.674	0.974
9	Mayonnaise	1,5,6	4.049	7.120	0.728	3.833	6.714	0.691	0.952	3.866	6.773	0.697	0.962
10	Milk	1,2,6,12,14	10.042	16.761	1.022	8.271	14.260	0.874	0.868	8.137	13.832	0.867	0.863
11	Peanut butter	4-7,9,10,13,14	4.205	7.712	0.711	4.106	7.467	0.696	0.979	4.131	7.453	0.698	0.983
12	Salty snacks	--	6.846	12.563	0.777	6.755	12.647	0.761	0.983	6.755	12.647	0.761	0.983
13	Soup	--	5.425	10.800	0.720	5.380	11.164	0.714	0.990	5.380	11.164	0.714	0.990
14	Spaghetti sauce	1,9-13	3.551	5.869	0.668	3.502	5.839	0.656	0.984	3.497	5.852	0.654	0.981
15	Yogurt	--	15.193	28.762	0.749	14.099	26.314	0.714	0.952	14.099	26.314	0.714	0.952

*ADL-own is the benchmark model used to calculate AvgRelMAE

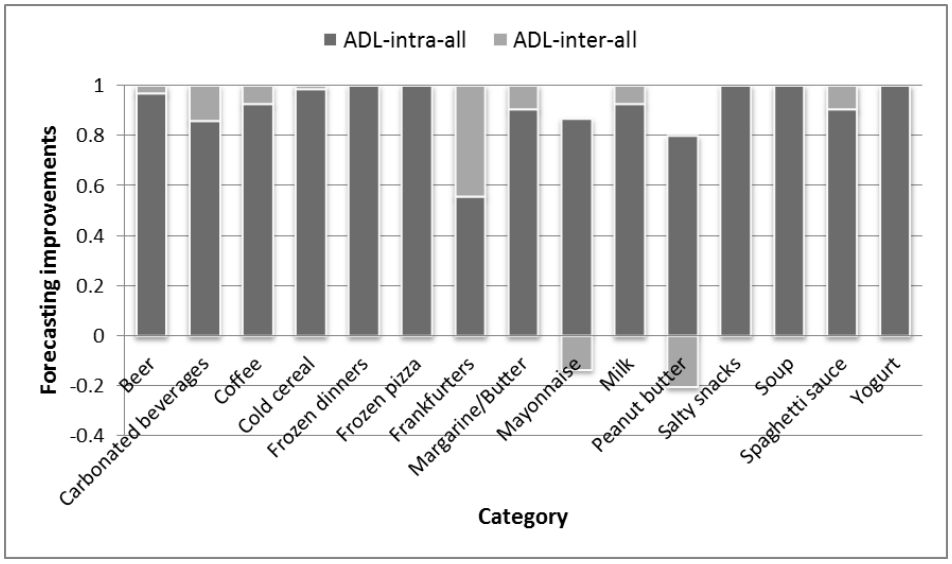

Figure 5 MAE improvements of ADL-intra-all and ADL-inter-all over ADL-own

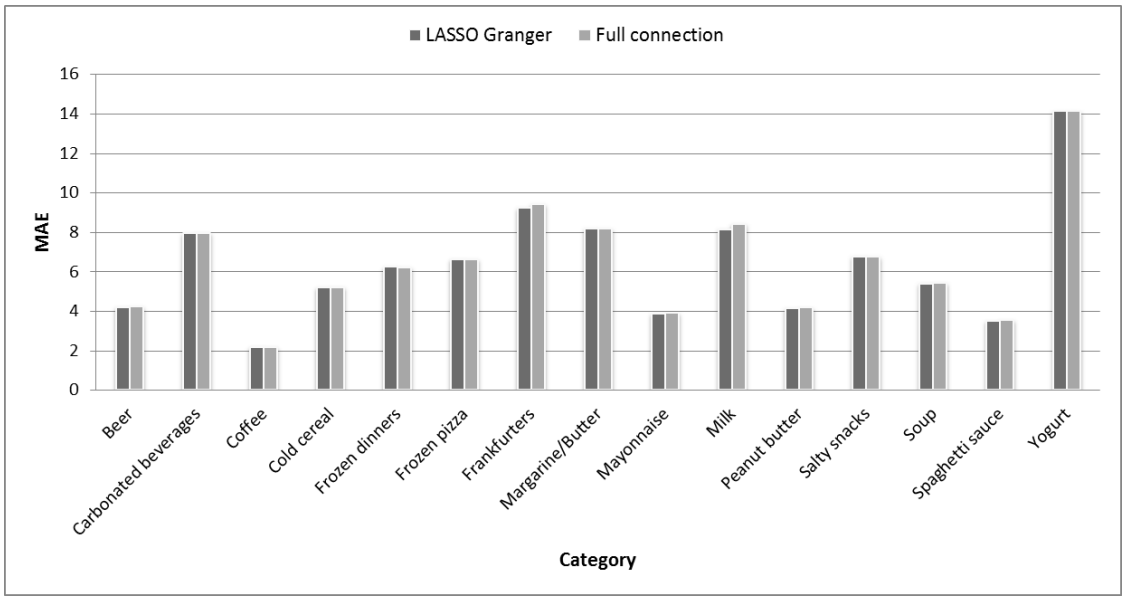

Figure 6 Forecasting comparisons between Full connection and LASSO Granger

In order to investigate whether the proposed LASSO Granger is an effective way to identify the category level interactive structure, we compare the forecasting results of the proposed LASSO Granger with the results of a fully connected structure based on ADL-inter-all under rolling scheme. The full connection means that when forecasting the sales of SKUs in one category, all other 14 categories are considered as influential categories. The comparison results are illustrated in Figure 6. All the MAEs of LASSO Granger across categories are smaller or equal to that of using structure of full connections. This means that connections (Figure 3) identified by LASSO Granger enhance the model's forecasting

abilities by reducing the redundant noisy data for some categories.

To show the necessity of the multistage LASSO, we compare the results from both one-stage and three stage LASSO regression in Table 4. For all models and both fixed and rolling forecasting schemes, three stage LASSOs are much more accurate than the forecasts produced with one stage LASSO whatever the measure.

Table 4 one week ahead forecasting comparison between one-stage and three-stage LASSO

Model	Estimation	Scheme	MAE	RMSE	MASE	AvgRelMAE
ADL-inter-top5	one-stage	fixed	8.284	27.844	0.929	1.087
ADL-inter-top5	three-stage	fixed	7.142	15.062	0.793	1.027
ADL-inter-all	one-stage	fixed	7.268	15.853	0.802	1.043
ADL-inter-all	three-stage	fixed	6.902	14.689	0.764	1.005
ADL-inter-PCA(3)	one stage	fixed	6.865	15.011	0.760	0.996
ADL-inter-PCA(3)	three-stage	fixed	6.752	14.255	0.750	0.986
ADL-inter-top5	one-stage	Rolling	6.224	13.408	0.701	0.989
ADL-inter-top5	three-stage	Rolling	6.165	13.201	0.697	0.983
ADL-inter-all	one-stage	Rolling	6.168	13.366	0.694	0.984
ADL-inter-all	three-stage	Rolling	6.092	13.015	0.690	0.975
ADL-inter-PCA(3)	one stage	Rolling	6.219	13.891	0.711	0.989
ADL-inter-PCA(3)	three-stage	Rolling	6.117	13.196	0.693	0.979

5. Discussion and Conclusion

In analyzing high-dimensional marketing data, the problem faced is that valuable predictors of consumer behaviour are often hidden in large number of useless noisy variables. When the dimensionality increases with the integration of intra- and inter- categorical information, the number of unreliable predictors which are correlated with valuable ones also increases rapidly. This makes the model difficult or even impossible to estimate. It is also difficult to select the 'correct' best specified model because the corresponding candidate models are many. Various methods have been proposed for selecting important variables from within the space. A key contribution of this paper is to propose a novel sequential selection method building on an approach, LASSO, well-known in statistics but rarely if ever used in marketing where the underperforming stepwise selection method is most often applied. This

new method meets one of the key requirements when analyzing ‘big data’ of being fully automatic. It is therefore suitable for application in the important marketing problem of SKU/ store level sales forecasting and promotional planning, when considering intra- and inter-category promotional information leads to high-dimensionality, which is this paper’s concern. The second substantive contribution of this paper is that it develops guidelines to practitioners on whether and how they can improve sales forecasting accuracy at SKU level by integrating intra- and inter-category promotional information when they are building a forecasting system for grocery retailers.

Specifically, on the methodological side, we propose a four steps framework to overcome the high dimensionality of the retail data set that results from integrating the intra- and inter-category promotional information. Our results show that the scheme of how one generates the sequence of regression estimates necessary to make forecasts is very important when integrating extra information. The multi-stage LASSO strategy is the key to improving the forecasts. This contributes to avoiding the selection of misleading variables among correlated variables by separating different sources of information into several layers. When considering inter-category information, the first stage in simplifying the problem and lessening the computational burden is to limit the number of categories to be considered: LASSO Granger is an effective way to identify the promotional interactions among categories. Then, various simplification schemes have been evaluated but a key element is to break down the process of variables selection into three stages: models that include just the target variables promotional history, those that also include the intra-category variables and finally, inter-category variables are included. In addition to selecting from amongst these variable sets, diffusion indices were also developed (based on principal components) that reduced the dimensionality of these sets. Differing from existing approaches (e.g. Scott and Watson), we combine diffusion factor with LASSO selection. We first cluster the massive number of explanatory variables into hundreds of subsets according to their common attributes (i.e. sales lag, price, display and feature), then for each subset, we conduct PCA dynamically and extract principle components as the inputs to the proposed multistage LASSO. This combines the merit of PCA which is effective in dealing with collinearity and LASSO which is good at variable selection in high dimensional space while make up for their drawbacks. Finally, a rolling forecasting scheme was shown to effectively utilize extra information by capturing complex dynamic relationships among products. The total selection process is fully automatic and therefore can be easily integrated into a forecasting system.

Our substantive results demonstrate which of the methods of variable selection work best in SKU level retail forecasting. Those models that integrate extra information, even if including extra information only from the intra-category five top sales products, perform significantly better than the baseline model when using a rolling forecasting scheme. Considering various measures of performance, the diffusion approach proved the most robust. In general, we can improve forecasting accuracy by about 14% over the baseline model that includes only the focal SKU's own predictors. But among the improvements, about 89% comes from the intra-category information, and only 11% from the inter-category information. However, the forecasting results at category level show that the accuracy improvements are spread unevenly among different categories. Though intra-category information still consistently contributes the main part of the forecasting improvements across categories, inter-category information can also contribute up to 40% in some categories. But integrating more information increases the computational complexity substantially: from data processing, model selection and estimation. In return, better forecasting accuracy can consistently be achieved. In practice, we need to weigh the benefit from increasing forecast accuracy and the cost and practicality of increasing computational complexity. Because of the rapidly decreasing of the cost on data storage, processing and computation, integrating more information to improve the grocery retailer's forecasting is a promising option.

When faced with large numbers of potentially explanatory variables it is all too easy for researchers to identify misleading relationships. In the existing marketing analytics literature, association-rule discovery or cross category choice models are popular methods to analyze the correlations between sets of products. These methods are often promoted as a means to obtain product associations on which to base a retailer's promotion strategy. Based on this approach, researchers have argued that associated products with a high lift/interest can be promoted effectively by only discounting just one of the two products (e.g. Song and Chintagunta, 2007; Mehta, 2007; Wang & Shao, 2004; Van den Poel et al., 2004). But Vindevogel et al. (2005) empirically show that this implicit assumption does not hold. A simple reason is that while associated products are often purchased together, this does not necessary imply that promotion of one product stimulates the other. The methods proposed in this paper directly capture this promotional interaction to form a correlation set for every product to improve their forecasts. They have the advantage of being rigorously validated through a rolling origin forecasting scheme. Based on the results the methods proposed could

also be used to build a promotional optimization expert system for retailers. This opens a very interesting direction for further exploration.

Acknowledgments The first author acknowledges the ongoing support of the National Natural Science Foundation of China under grant nos. 70871057, 71171100, and the support of State Scholarship Fund for overseas studies.

References

Aburto, L., & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. Applied Soft Computing, 7, 136-144.

Alon I., Qi M., & Sadowsik, R. J.(2001). Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods. Journal of Retailing Consumer Services, 8(3), 147–156.

Andrews, R. L., Currim, I. S., Leeflang, P., & Lim, J. (2008). Estimating the SCAN*PRO model of store sales: HB, FM or just OLS? International Journal of Research in Marketing, 25, 22-33.

Arnold A.,Liu Y., & Abe, N. (2007). Temporal causal modeling with graphical granger methods. KDD '07, New York, USA, pp. 66-75.

Ashley R., Granger, C. W. J. & Schmalensee, R.(1980). Advertising and aggregate consumption: an analysis of causality. Econometrica, 48(5): 1149-1167

Bandyopadhyay, S. (2009). A dynamic model of cross-category competition: theory, tests and applications. Journal of Retailing, 85(4), 468-479.

Berman B., & Evans, J.R. (1989) Retail management: a strategic approach. New York: Macmillian.

Bronnenberg B.J., Kruger, M.W., & Carl, F. M. (2008). The IRI Academic Dataset, Marketing Science, 27(4), 745-748.

Brovelli A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., & Bressler, S. L.(2004). Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. Proceedings of the National Academy of Sciences of the United States of America, 101(26):9849-54.

Bucklin, R.E., Gupta, S., & Siddarth, S. (1998). Determining segmentation in sales response across consumer purchase behaviors. Journal of Marketing Research, 35, 189–97.

Cooper, L. G., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). “Promocast”: a new forecasting method for promotion planning. Marketing Science, 18, 301-316.

Chiang J. (1991). A simultaneous approach to the whether, what, and how much to buy questions, Marketing Science, 10 (4), 297–315.

Chintagunta, Pradeep, K. (1993). Investigating purchase incidence, brand choice, and purchase quantity Decisions of households. Marketing Science, 12 (2), 184–208.

Curry, D., Divakar, S., Mathur, S. K., & Whiteman, C. H. (1995). BVAR as a category management tool: An illustration and comparison with alternative techniques. Journal of Forecasting, 14, 181-199.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting, 29(3), 510–522.

Divakar, S., Ratchford, B. T., & Shankar, V. (2005). CHAN4CAST: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. Marketing Science, 24, 334-350.

Donoho, D. L. (2000) High-dimensional data analysis: the curses and blessings of dimensionality. Aide-Memoire of the lecture in AMS conference, Math challenges of 21st Century. Available at http: // www-stat.stanford.edu/~donoho/Lectures.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. Annals of Statistics, 32, 407- 451 .

Erdem, T. (1998). An empirical analysis of umbrella branding. Journal of Marketing Research, 35, 339–51.

Fan J, Lv J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of Royal Statistical Society, Series B. 70, 849–911

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. International Journal of Forecasting, 25, 3-23.

Fildes, R., Nikolopoulos, K., Crone, S., & Syntetos, A. A. (2008). Forecasting and operational research: A review. Journal of the Operational Research Society, 59, 1150–1172.

Foekens, E. W., Leeflang, P. S. H., & Wittink, D. R. (1994). A comparison and an exploration of the forecasting accuracy of a loglinear model at different levels of aggregation. International Journal of Forecasting, 10, 245–261.

Forni, M., Hallin M., Lippi M. & Reichlin L. (2000). The generalized factor model: identification and estimati on. Review of Economics and Statistics, 82, 540–554.

Forni, M., Hallin M., Lippi M. & Reichlin L. (2003). Do financial variables help forecasting inflation and real activity in the EURO area? Journal of Monetary Economics, 50, 1243-1255.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. Journal of Marketing Research, 25, 322-355.

Gür Ali, Ö., SayIn, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. Expert Systems with Applications, 36(10), 12340–12348.

Harrell, F. E. (2001). Regression Modeling Strategies: with applications to linear models, logistic regression, and survival Analysis. New York: Springer.

Heerde, H.J., Leeflang, V., Peter, S. H. & Wittink, D.R. (2000). The estimation of pre-and postpromotion dips with store-Level scanner data. Journal of Marketing Research, 37,383 – 395.

Heerde, H.J., Leeflang, V., Peter, S. H. & Wittink, D.R. (2001). Semiparametric analysis to estimate the deal effect curve. Journal of Marketing Research, 38, 197 – 215.

Heerde, H.J., Gupta, S. & Wittink, D. R. (2003). Is 75% of the sales promotion bump due to brand switching? No, only 33% is. Journal of Marketing Research, XL, 481-491.

Hiemstra, C. & Jones, J. D.(1994). Testing for linear and nonlinear Granger causality in the stock price-volume Relation. Journal of Finance, 49(5):1639-1664.

Hruschka, H. (2013). Comparing small- and large-scale models of multicategory buying behavior. Journal of Forecasting, 32(5): 423-434.

Huang, T., Fildes, R. & Soopramanien, D.(2014). The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. European Journal of Operational Research, 237(2): 738-748.

Hyndman, R.J., Koehler, A.B., Snyder, R.D., and Grose, S. (2002) A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting, 18(3), 439–454.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22, 679-688.

John, G.H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In: Proceedings of the Eleventh International Conference on Machine Learning, 121–129.

Kumar, V. & Leone, R. (1988). Measuring the effect of retail store promotions on brand and store substitution. Journal of Marketing Research, 25 (May), 178-85.

Kuo, R. J. (2001). A sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. European Journal of Operational Research, 129, 496–517.

Kalyanam, K., Borle S., Boatwright P. (2007). Deconstructing each item's category contribution. Marketing Science, 26(3): 327-341.

Lee S., Kim, J., & Allenby, G.M. (2013). A direct utility model for asymmetric complements. Marketing Science, 32(3), 454-470.

Levy, M., Grewal, D., Kopalle, P.K. & Hess, J.D. (2004). Emerging trends in retail pricing practice: implications for research. Journal of Retailing, 80(3): xiii-xxi.

Mehta, N. (2007). Investigating consumers purchase incidence and brand choice decisions across multiple product categories: A theoretical and empirical analysis. Marketing Science, 26(2), 196-217.

Meier, L., van de Geer, S. & Buhlmann, P. (2008). The group lasso for logistic regression. Journal of the Royal Statistical Society, Series B, 70, 53-71 .

Meiri R. & Zahavi J. (2006) Using simulated annealing to optimize the feature selection problem in marketing applications. European Journal of Operational Research, 171(3):842-858
Melab, N., Cahon, S., Talbi, E.-G., & Duponchel, L. (2002). Parallel GA-based wrapper feature selection for spectroscopic data mining. International Parallel and Distributed Processing Symposium: IPDPS 2002 Workshops.
Moriarty, M. (1985). Retail promotional effects on intra and interbrand sales performance. Journal of Retailing, 61 (Fall), 27-47.
Mulhern, F.J., and Leone, R.P. (1991). Implicit price bundling of retail products: a multiproduct approach to maximizing store profitability. Journal of Marketing, 55, 63-76.
Nicholson, Walter (1998). Microeconomic theory: basic principles and extensions. South-Western Cengage Learning, Mason, Ohio.
Ord, J. K., Fildes, R.(2013). Principles of business forecasting. South-Western Cengage Learning, Mason, Ohio.
Preston, J., & Mercer, A. (1990). The evaluation and analysis of retail sales promotions. European Journal of Operational Research, 47, 330- 338.
Raju, J. S. (1995). Theoretical models of sales promotions: Contributions, limitations, and a future research agenda. European Journal of Operational Research, 85(1), 1-17.
Rinne, H., & Geurts, M. (1988). A forecasting model to evaluate the profitability of price promotions. European Journal of Operational Research, 33, 279-289.
Song, I., & Chintagunta, P.K. (2007). A discrete-continuous model for multicategory purchase behavior of households. Journal of Marketing Research, 44(4), 595-612.
Stock, J. & Watson M. (1999). Forecasting inflation, Journal of Monetary Economics, 44, 293-335.
Stock, J. & Watson, M. (2002). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association, 97, 1167-1179.
Stock, J.., & Watson M. (2003). Forecasting output and inflation: The role of asset prices. Journal of Economic Literature, 41, 788-829.
Stock, J. & Watson, M. (2004). Forecasting with many predictors. In Handbook of Economic Forecasting. North Holland, Elsevier. Taylor, J. W. (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. European Journal of Operational Research, 178, 154-167.
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of Royal Statistical Society, Series B, 58, 267–288.
Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B, 73(3): 273-282.
Van den Poel, Schamphelaere, D.D., Wets, J.G. (2004). Direct and indirect effects of retail promotions. Expert Systems with Applications, 27(1), 53–62.
Vindevogel, B., Van den Poel D., et al. (2005). Why promotion strategies based on market basket analysis do not work. Expert Systems with Applications 28(3): 583-590.
Walters, R,G. (1988). Retail promotions and retail store performance: a test of some key hypotheses, Journal of Retailing, 64 (Summer), 153-180.
Walters, R.G. (1991). Assessing the impact of retail price promotions on product substitution, complementary purchase, and inter-store sales displacement. Journal of Marketing, 55 (April), 17-28.
Wang, F.S., Shao,H.M. (2004). Effective personalized recommendation based on time-framed navigation clustering and association mining. Expert Systems with Applications, 27(3), 365–377.
Wedel, M. Zhang, J. (2004). Analyzing brand competition across subcategories. Journal of Marketing Research. 41(4), 448-456
Wittink, D., Addona, M., Hawkes, W., & Porter, J. (1988). SCAN*PRO: the estimation, validation and use of promotional effects based on scanner data. In Internal paper: Cornell University.
Zhang, J.L., Chen J. & Lee, C.Y. (2008). Joint optimization on pricing, promotion and inventory control with stochastic demand. International Journal of Production Economics 116(2): 190-198.
Zou, H. & Hastie, T. (2005) Regularization and variable selection via the elasticnet. Journal of the Royal Statistical Society, Series B, 67, 301-320 .
Zou, H., Hastie, T. & Tibshirani, R.(2006). Sparse principal component analysis. Journal of Computational

and Graphical Statistics, 15, 265-286 .