**Forecasting Retail SKU Sales in The Presence of Structural Breaks**

Tao Huang[[1]](#footnote-1)

Surrey Business School, University of Surrey, GU2 7XH, UK

Robert Fildes

Centre for Marketing Analytics and Forecasting, Lancaster University, LA1 4YX, UK

Didier Soopramanien

Lancaster University management school, Lancaster University, UK, LA1 4YX

Abstract

Grocery retailers need accurate forecasts at SKU level for their inventory management decisions. Previous studies have developed forecasting models which incorporate the impact of various marketing activities. These models, however, do not consider that the effect of these marketing activities on sales may not be constant over time. Under such a circumstance, the models could be subject to the structural break problem, i.e., the models with constant parameters are unable to capture the varying effect of the marketing activities. As a result, the generated forecasts may potentially be biased and less accurate. In this study, we propose new forecasting methods for retail product sales by taking into account the problem of structural break. Our methods generate more accurate forecasts compared to conventional models which assume constant parameters for various marketing activities.

Keywords:

Forecasting, OR in marketing, Analytics

1. **Introduction**

Grocery retailers rely on accurate sales forecasts for their inventory management ([Petropoulos, Makridakis, Assimakopoulos, & Nikolopoulos, 2014](#_ENREF_67" \o "Petropoulos, 2014 #3)). Poor forecasts of product sales lead to out-of-stock conditions and overstocking conditions. When a specific item is out-of-stock, retailers directly lose the income and profit from the sale of the item. If the out of stock situation happens on a regular basis, it can lead to consumer dissatisfaction. In the long term, retailers may see customers switching to other retail chains ([Corsten & Gruen, 2003](#_ENREF_25" \o "Corsten, 2003 #624)). To avoid such situations, retailers may intentionally overstock to maintain a high customer satisfaction level but this significantly raises inventory costs (e.g., capital cost, warehousing, and deterioration etc.) and reduces profits ([L. G. Cooper, Baron, Levy, Swisher, & Gogos, 1999](#_ENREF_23" \o "Cooper, 1999 #662)). In 2014, retailers in North America had a loss of $634.1 billion due to out-of-stock and spent $471.9 billion on overstock ([OrderDynamics, 2015](#_ENREF_59" \o "OrderDynamics, 2015 #737)). One of the solutions to mitigate the dilemma is to generate more accurate sales forecasts at SKU level, which improves the effectiveness of the supply chain management by reducing the bullwhip effect and enabling Just-In-Time delivery ([Ouyang, 2007](#_ENREF_60); [Sodhi & Tang, 2011](#_ENREF_68)).

In practice, many retailers generate forecasts at SKU level using a two-stage ‘base-lift’ approach. The forecasts are generated separately depending on whether or not the focal product is being promoted. The ‘base’ forecasts are generated using simple univariate models, while the ‘lift’ effect, which is effectively caused by the marketing activities, is estimated by the brand/category manager based on his/her experience. In this context, some previous studies have proposed procedures to help managers improve the accuracy of their judgments (e.g., [Robert Fildes, Nikolopoulos, Crone, & Syntetos, 2008](#_ENREF_34" \o "Fildes, 2008 #159); [Goodwin, 2002](#_ENREF_37" \o "Goodwin, 2002 #652); [Nikolopoulos, 2010](#_ENREF_58" \o "Nikolopoulos, 2010 #605)). Others have developed models to estimate the ‘lift’ effect based on data ([L. G. Cooper et al., 1999](#_ENREF_23" \o "Cooper, 1999 #662); [L. G. Cooper & Giuffrida, 2000](#_ENREF_24" \o "Cooper, 2000 #24); [Trusov, Bodapati, & Cooper, 2006](#_ENREF_73" \o "Trusov, 2006 #25)). Some studies propose methods to directly generate the final forecast of the product sales. For example, [Gür Ali, et al. (2009](#_ENREF_30" \o "Gür Ali, 2009 #715)) proposed the regression tree method with a range of variables constructed from the sales, price, and promotion of the focal product. [Huang, Fildes, and Soopramanien (2014)](#_ENREF_41" \o "Huang, 2014 #732) proposed two-stage general-to-specific Autoregressive Distributed Lag (ADL) models which incorporate the promotional information of not only the focal product but also of the competitive products within the same product category. [Ma, Fildes, and Huang (2016)](#_ENREF_46" \o "Ma, 2016 #733) further integrate the promotional information of the products from related product categories.

However, these studies assume that the impact of marketing activities on product sales remains constant over time. In practice, the effect of prices and promotions may change due to many non-controllable factors which may include, for instance, the change of economic conditions, the change in consumer tastes, and the entry of new competitors etc. which are usually neither observable or measurable ([Wildt, 1976](#_ENREF_78" \o "Wildt, 1976 #635); [Wildt & Winer, 1983](#_ENREF_79" \o "Wildt, 1983 #218)). Customers may become more sensitive to price reductions and promotions during an economic crunch. They may change their tastes due to factors including cognitive bias, the change of their familiarity with the product, and the change of their lifestyle and social status ([Meeran, Jahanbin, Goodwin, & Quariguasi Frota Neto, 2017](#_ENREF_50" \o "Meeran, 2017 #2)). When a new competitor enters the market, the effect of prices and promotions of the focal product may be reduced not only because the new competitor launches their marketing activities but also because customers seek variety. In the year of 2014, the German low-price retail chain Aldi has opened more than 400 stores in the United States, which put pressures on existing retail chains ([Loeb, 2015](#_ENREF_45" \o "Loeb, 2015 #734)).

Under any of the circumstances described above, conventional models which use constant parameters to represent the effect of the price and promotions may potentially be subject to the structural break problem ([Allen & Fildes, 2001](#_ENREF_2" \o "Allen, 2001 #204); [Armstrong, 2001](#_ENREF_8" \o "Armstrong, 2001 #215)). The model which is subject to structural break may generate biased and less accurate forecasts. This has been historically addressed in the macroeconomics literature (see [M. B. Clements & Hendry, 1994](#_ENREF_20" \o "Clements, 1994 #647); [M. H. Pesaran & Timmermann, 2005](#_ENREF_66" \o "Pesaran, 2005 #622)). However, the problem of structural break has been totally overlooked in forecasting retailer product sales. In this study, we propose new forecasting methods which deal with the structural break proplem and produce more accurate forecasts.

Our research in the domain of retail forecasting in particular at the SKU level is significant for the following reasons. First, our methods have superior forecasting performance compared to conventional models which assume no change in the effect of product prices and promotions. Second, unlike any earlier studies which rely on incorporating additional information (which leads to additional cost), our methods rely on how promotional information could be effectively utilized. In practice, the change of the effect of the marketing activities may be caused by many influencing factors (as mention above) for which the data are difficult to collect or measure. Third, our research provides an evaluation of various forecasting methods which offers operational guidance to not only retailers but also to manufacturers when competitive promotional information is unavailable. Fourth, the methods we propose are fully automatic compared to [Huang et al. (2014)](#_ENREF_41" \o "Huang, 2014 #732) and easy to implement.

The remainder of the paper is arranged as follows: Section 2 summarizes previous studies in the literature. Section 3 explains the origins and the consequence of the structural break problem. In section 4, we introduce two techniques which may potentially improve the forecasting accuracy by mitigating the forecast bias due to structural breaks. Section 5 explore the data. in section 6, we propose our new forecasting methods. Section 7 describes the design of the model evaluation. Section 8 summarizes and discusses the evaluation results. In Section 9, we explore the determinants of improvement of the proposed models. In the last section, we make recommendations for both manufacturers and retailers, address research limitations, and highlight directions for future research.

1. **Literature review**

In practice, many retailers forecast their product sales at SKU level using a two-stage ‘base-lift’ method. The method entails dividing the data into promoted and non-promoted periods based on whether the SKU is being promoted. The method is a combination of simple univariate methods (for the non-promoted period) and human judgments by brand/category managers (for the promoted period) ([R. Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009](#_ENREF_33" \o "Fildes, 2009 #421); [Robert Fildes et al., 2008](#_ENREF_34" \o "Fildes, 2008 #159)). A stream of studies has been devoted to helping managers with better adjustment procedures overcome their cognitive bias ([Arenas, Pedregal, J., Fildes, & Kourentzes, 2013](#_ENREF_7" \o "Arenas, 2013 #738); [R Fildes & Goodwin, 2007](#_ENREF_32" \o "Fildes, 2007 #161)). Other studies try to improve the adjustment with model-based forecasting systems. e.g., they may estimate the ‘lift’ effect by the promotional event based on historical information related to previous promotions, store/category features, and manufacturers etc. ([L. G. Cooper et al., 1999](#_ENREF_23" \o "Cooper, 1999 #662); [L. G. Cooper & Giuffrida, 2000](#_ENREF_24" \o "Cooper, 2000 #24); [Trusov et al., 2006](#_ENREF_73" \o "Trusov, 2006 #25)). One limitation of these methods is that, as they split the data into two periods, they tend to overlook the information in the promoted period when forecasting the product sales in the non-promoted period, and vice versa.

other studies have proposed more holistic methods to generate the forecasts. Divakar et al. (2005) developed the CHAN4CAST system with models of dynamic regression structures to forecast brand volume sales for the manufacturer/channel. [Gür Ali et al. (2009)](#_ENREF_38" \o "Gür Ali, 2009 #715) evaluated the forecasting performance of support vector machine (SVM) models and regression tree models. [Huang et al. (2014)](#_ENREF_41" \o "Huang, 2014 #732) proposed two-stage general-to-specific ADL models which incorporate competitive promotional information within the same product category of the focal product. [Ma et al. (2016)](#_ENREF_46" \o "Ma, 2016 #733) further integrated the promotional information not only from the same category but also from other related categories. These studies tried to generate accurate sales forecast by capturing the effect of marketing activities. For example, the short-term effect of prices and promotions ([Blattberg, Briesch, & Fox, 1995](#_ENREF_11" \o "Blattberg, 1995 #36)), the (asymmetrical) competitive effect ([R. L. Andrews, Currim, Leeflang, & Lim, 2008](#_ENREF_5" \o "Andrews, 2008 #717); [Dekimpe, Hanssens, & Silva-Risso, 1999](#_ENREF_27" \o "Dekimpe, 1999 #92); [Wedel & Zhang, 2004](#_ENREF_74" \o "Wedel, 2004 #6); [Wittink, Addona, Hawkes, & Porter, 1988](#_ENREF_79" \o "Wittink, 1988 #163)), and the dynamic effects which lead to purchase acceleration and anticipation ([Mace & Neslin, 2004](#_ENREF_47" \o "Mace, 2004 #20); [Van Heerde, Gupta, & Wittink, 2003](#_ENREF_72" \o "Van Heerde, 2003 #49)).

However, evidence suggests that the effect of prices and promotions may change over time (e.g. [Houston & Weiss, 1975](#_ENREF_40" \o "Houston, 1975 #687); [Little, 1966](#_ENREF_44" \o "Little, 1966 #688); [Mahajan, Bretschneider, & Bradford, 1980](#_ENREF_48" \o "Mahajan, 1980 #220); [Moinpour, McCullough, & MacLachlan, 1976](#_ENREF_51" \o "Moinpour, 1976 #689); [Monroe & Guiltinan, 1975](#_ENREF_52" \o "Monroe, 1975 #690); [Morrison, 1966](#_ENREF_53" \o "Morrison, 1966 #691); [Myers, 1971](#_ENREF_55" \o "Myers, 1971 #692); [Myers & Nicosia, 1970](#_ENREF_56" \o "Myers, 1970 #693); [Wichern & Jones, 1977](#_ENREF_77" \o "Wichern, 1977 #694); [Wildt, 1976](#_ENREF_78" \o "Wildt, 1976 #635); [Winer, 1979](#_ENREF_80" \o "Winer, 1979 #221)). [Wildt (1976)](#_ENREF_78" \o "Wildt, 1976 #635) and [Wildt and Winer (1983)](#_ENREF_79" \o "Wildt, 1983 #218) attribute the change in the effect of the marketing activities to the change in economic conditions, consumer tastes, and competitive environment etc. Customers may, for instance, become more sensitive to price reductions and sales promotions because these are more attractive when there is an economic crunch compared to other time periods. Customers may also display a change in their tastes and preferences. This can occur as customers accumulate more knowledge of the product when they try to seek variety or they may also change their preferences when they reach a different social status or decide to adopt a different lifestyle ([Meeran et al., 2017](#_ENREF_50" \o "Meeran, 2017 #2)). Research at the store level also finds that the introduction of new brands in a product category (e.g., the store-owned brand) decreases the promotional elasticities of premium national brands and increase promotional elasticities of the second tier national brands ([Nijs, Dekimpe, Steenkamps, & Hanssens, 2001](#_ENREF_57" \o "Nijs, 2001 #45); [Van Heerde, Srinivasan, & Dekimpe, 2008](#_ENREF_75" \o "Van Heerde, 2008 #640)). Lastly, the effect of prices and promotions may change during the different stages of the product lifecycle ([Mahajan et al., 1980](#_ENREF_48" \o "Mahajan, 1980 #220)). The change in the effect of prices and promotions on sales, however, has been overlooked by previous studies in the forecasting literature.

3. **Structural break and potential forecast bias**

Previous models overlook the change in the effect of prices and promotions on product sales, and they may potentially be subject to the structrual break problem. The problem originates from the large changes in the model’s parameters ([Allen & Fildes, 2001](#_ENREF_2" \o "Allen, 2001 #204); [Armstrong, 2001](#_ENREF_8" \o "Armstrong, 2001 #215)). As a result, the forecasts of the sales become the weighted average of the true product sales before and after the change of the true parameter which are biased and less accurate ([M. P. Clements & Hendry, 1999](#_ENREF_21" \o "Clements, 1999 #199)). [Pesaran and Timmermann (2005)](#_ENREF_66" \o "Pesaran, 2005 #622) illustrate analytically how a structural break may lead to forecast bias using a simple regression model. In a retailer context, suppose that we have the sales and price information of the focal product from week 1 to week *T,* i.e.,, and we presume that the sales are only driven by the price here for exposition and that there is a structural break at week (where ). That is, the true parameter of the price variable changes from to after . The unobserved true demand can be represented as follows:

where, is an indicator which equals to 1 before week and 0 afterwards. and are the sales and the price of the product at week *t*. We consider to be strictly exogenous as we assume that retailers do not change product prices based on their short-term sales[[2]](#footnote-2). and are the true parameters of the price before and after the structural break at week . is the error term. when and when . We may estimate a model with a functional form which is congruent with the demand (e.g., ) using the data before the structural break, e.g., ,. The OLS estimate for the parameter is:

where and are the matrices for the sales and price for the time period from week *m* to week *T*. We assume that there is no structural break after week *T* and the true demand after week *T* remains as . Therefore, the *h*-step ahead forecast error at week *T*+*h* can be represented as:

where ,and is the matrix for the error term at week .

Therefore, the forecast at week is biased as , which is unequal to zero. In Appendix A, we illustrate the impact of the structural break on the forecasting performance using a simulation example.



The problem of the structural break has been mainly addressed in the macroeconomics literature (e.g. [Castle, Doornik, & Hendry, 2008](#_ENREF_14" \o "Castle, 2008 #241); [M. P. Clements & Hendry, 1999](#_ENREF_21" \o "Clements, 1999 #199); [J. P. Cooper & Nelson, 1975](#_ENREF_22" \o "Cooper, 1975 #698); [Hendry, 1995](#_ENREF_39" \o "Hendry, 1995 #259); [Muellbauer, 1994](#_ENREF_54" \o "Muellbauer, 1994 #699); [H. M. Pesaran & Timmermann, 2007](#_ENREF_62" \o "Pesaran, 2007 #254); [M. H. Pesaran & Pick, 2011](#_ENREF_63" \o "Pesaran, 2011 #749); [Stock & Watson, 1996](#_ENREF_70" \o "Stock, 1996 #747)). Previous studies suggest that the parameters of the forecasting models may change due to exogenous factors including a shift in the market sentiments, new regulations, and a change of debt management etc. These studies focus on the domain of financial interest rate and stock market return (e.g., [Ang & Bekaert, 2002](#_ENREF_6" \o "Ang, 2002 #749); [Perez-Quiros & Timmermann, 2000](#_ENREF_61" \o "Perez-Quiros, 2000 #750); [M. H. Pesaran & Timmermann, 2002](#_ENREF_65" \o "Pesaran, 2002 #748)).

4. **Dealing with structural breaks**

In this study, we propose new methods to forecast retail sales at SKU level by taking into account the impact of structural breaks. In this section, we introduce the Intercept Correction (IC) method and the Estimation Window Combining (EWC) method which will later be incorporated as one of the stages in our proposed methods.

4.1 The Intercept Correction method

When a forecasting model is subject to structural break, the IC model uses conventional models to generate forecasts but specifies non-zero values for the model’s errors in the forecasting period ([Clark & McCracken, 2007](#_ENREF_17" \o "Clark, 2007 #739); [M. B. Clements & Hendry, 1994](#_ENREF_20" \o "Clements, 1994 #647); [M. P. Clements & Hendry, 1999](#_ENREF_21" \o "Clements, 1999 #199)). Based on the same example in section 3, if we believe the model is subject to structural break and generate biased forecasts, we may estimate the bias as an average value of the most recent residuals, i.e., , where is the estimated bias for the IC method, is the number of residuals being used to calculate the forecast bias. When , the estimate reduces to , which is the residual at the forecast origin ([e.g., Chevillon, 2016](#_ENREF_15" \o "Chevillon, 2016 #123)). The estimated bias are added back to the out-of-sample forecasts, which may potentially improve the forecasting accuracy though at a cost of inflated forecasting error variance ([M. P. Clements & Hendry, 1999](#_ENREF_21" \o "Clements, 1999 #199)). However, in a retailer context, product sales at SKU level usually exhibit large variations, unexpected outliers, and missing values, which makes the bias estimation difficult. Therefore, whether we can rely on the IC method to further improve forecasting accuracy becomes an empirical question.

4.2 The Estimation Window Combining method

An alternative method to deal with the forecast bias due to structural break is the estimation window combining method ([M. H. Pesaran and Timmermann (2005)](#_ENREF_66" \o "Pesaran, 2005 #622). Unlike the IC method, the EWC method does not estimate the forecast bias: it takes a trade-off between the forecast bias and the forecast error variance by combining the forecasts generated by the same model but with different estimation windows ([M. H. Pesaran & Pick, 2011](#_ENREF_63" \o "Pesaran, 2011 #749)). In the example in section 3, if we know that there is a structural break and that we are aware of the location, it is possible to estimate the model with the post-break data and generate unbiased forecasts. Alternatively, if we suspect that there is a break but we do not know the location, we may use the most recent observations close to the forecast origin as long as there are enough observations to estimate the model. It becomes less likely for the model to be subject to structural break if we keep *m* as large as possible (so that we discard more old data). If *m* by chance becomes larger than , the model will be estimated with post-break data only and generate unbiased forecasts.

However, this does not imply that when we may suspect that there are structural breaks we must adopt estimation windows as small as possible to achieve higher forecasting accuracy (even there are enough observations to estimate the model). This is because that the reduction of the forecast bias comes with the cost of an inflated forecasting error variance as we will be using less information (e.g., the estimation sample becomes smaller). In the example in section 3, the Mean Squared Error (MSE) at week , which is a commonly used forecasting error measure, can be represented as , where , and can be interpreted as the squared forecast bias; , and can be interpreted as the efficiency term ( is the forecasting error variance), μ, , and ψ. [M. H. Pesaran and Timmermann (2005)](#_ENREF_66" \o "Pesaran, 2005 #622) show that the change of the MSE for week when we estimate the model with one less pre-break observation (e.g., with estimation window compared to) can be represented as:

where is the MSE at week based on the estimation window [m+1, T]. When the observation at week *m* is excluded in the estimation, the change of the squared bias term (e.g., ) will always be non-positive (i.e., the bias will decrease), but the change of the efficiency term (e.g., ) depends on the error variance before and after the structural break. If (e.g., there are more pre-break variations compared to post-break variations in the product sales which cannot be explained by the price variable), will be smaller than or equal to , and the MSE will decrease as the change for both the squared bias term and the efficiency term are non-positive. However, if , will be larger or equal to . Under this condition, the MSE may either increase or decrease depending on how the non-positive change of the squared bias term compares to the non-negative change of the efficiency term. Therefore, when we exclude pre-break data and adopt a smaller estimation window, we may have either better or worse forecasting performance depending on the trade-off between the reduced forecast bias and the potentially inflated forecasting error variance, and vice versa.

Instead of searching for the optimal estimation window, [M. H. Pesaran and Timmermann (2005)](#_ENREF_66" \o "Pesaran, 2005 #622) suggest combining the forecasts generated by the model of the same specification but estimated with different windows. The rationale is that the forecasts generated with large estimation windows may be subject to more bias (contains more pre-break data) but smaller error variance (with more observations), and vice versa. Thus a combination of the forecasts may potentially achieve an effective trade-off between forecast bias and forecasting error variance and lead to higher accuracy ([Clemen, 1989](#_ENREF_18" \o "Clemen, 1989 #745); [Jose & Winkler, 2008](#_ENREF_43" \o "Jose, 2008 #746)).

In this study, we combine the forecasts with equal weights as the equal weight combining scheme has been found effective and easy to implement.([M. Clements & Hendry, 1998](#_ENREF_19" \o "Clements, 1998 #608); [Dekker, van Donselaar, & Ouwehand, 2004](#_ENREF_28" \o "Dekker, 2004 #246); [R. Fildes & Stekler, 2002](#_ENREF_35" \o "Fildes, 2002 #522); [M. H. Pesaran, Schuermann, & Smith, 2009](#_ENREF_64" \o "Pesaran, 2009 #255)). For example, we may estimate the model using the most recent observations to generate the 1st set of the *h*-step-ahead forecast, e.g., , where represents the parameters estimated with the sample window . The value of is arbitrarily chosen given there are enough observations to estimate the model and there are enough variations for the explanatory variables. We then add more observations (e.g., one) to the estimation window and generate the 2nd set of the *h*-step-ahead forecast, e.g., and so forth. We may have the set of the *h*-step-ahead forecasts, e.g., . Finally, we may calculate the final forecasts as the average value of the () sets of *h*-step-ahead forecasts:

In Appendix B, we illustrate how we can achieve more accurate forecasts with the IC method and the EWC method with simple examples.

1. **The data**

We evaluate the forecasting performance of the models using the retail dataset made available by the Information Resources, Inc. (IRI) company. A description of the dataset can be found in [Bronnenberg, Kruger, and Mela (2008)](#_ENREF_12" \o "Bronnenberg, 2008 #741). The dataset contains weekly data at SKU level with variables including product unit sales, price, features, and displays etc. We conduct our evaluation based on 1831 SKU’s for 28 product categories from 28 stores[[3]](#footnote-9). Table 1 shows the basic statistics for the selected SKU’s for each of the categories. Some categories (e.g., Carbonated Beverages and Hotdog) exhibit much higher promotional intensity compared to others (e.g., Margarine/Butter and Mayonnaise). Figure 1 exhibits the data series for a typical SKU in the Beer category. For example, the product has occasional price reductions and feature/display events which are associated with sales spikes or calendar events (e.g., holidays such as Halloween, Thanksgiving, and Christmas etc.). Figure 1 also shows the calendar effects such as Christmas, New Year’s Day etc.

Table 1. Statistical description of different product categories

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Price mean | Sales mean | Display percentage | Feature percentage | Number of SKU's | Category | Price mean | Sales mean | Display percentage | Feature percentage | Number of SKU's |
| Beer | 8.3 | 20.6 | 13.9% | 4.0% | 169 | Mayonnaise | 3.0 | 79.7 | 3.0% | 0.4% | 22 |
| Blades | 8.1 | 14.6 | 7.4% | 2.2% | 22 | Milk | 2.5 | 222.3 | 2.1% | 1.8% | 30 |
| Carbonated Beverages | 2.1 | 113.6 | 26.8% | 15.6% | 82 | Mustard & Ketchup | 2.1 | 64.5 | 5.3% | 0.9% | 22 |
| Cigarette | 22.3 | 22.2 | 0.0% | 0.8% | 202 | Peanut butter | 3.7 | 34.2 | 3.2% | 0.6% | 15 |
| Coffee | 5.2 | 14.5 | 5.2% | 2.9% | 86 | Photo | 7.2 | 9.2 | 4.6% | 5.1% | 13 |
| Cold cereal | 3.5 | 70.7 | 4.0% | 18.1% | 125 | Salty snacks | 2.3 | 50.9 | 6.7% | 5.0% | 100 |
| Deodorant | 2.7 | 6.9 | 4.1% | 5.2% | 126 | Shampoo | 3.5 | 9.9 | 12.8% | 7.1% | 70 |
| Face Tissue | 2.1 | 75.8 | 0.3% | 11.7% | 6 | Soup | 1.5 | 61.6 | 1.2% | 9.7% | 139 |
| Frozen Dinner | 2.0 | 43.8 | 5.3% | 23.7% | 87 | Spaghetti sauce | 2.4 | 39.1 | 1.6% | 6.5% | 51 |
| Frozen pizza | 3.4 | 31.2 | 8.9% | 9.1% | 147 | Sugar substitutes | 2.8 | 14.5 | 0.1% | 1.4% | 20 |
| Household Cleaner | 2.5 | 29.9 | 0.3% | 3.6% | 18 | Toilet Tissue | 5.4 | 89.1 | 4.3% | 8.3% | 20 |
| Hotdog | 4.0 | 68.6 | 13.2% | 15.6% | 35 | Toothbrush | 2.6 | 8.7 | 3.1% | 6.3% | 27 |
| Laundry Detergent | 8.8 | 28.9 | 2.3% | 8.8% | 57 | Toothpaste | 2.8 | 35.5 | 11.0% | 12.5% | 25 |
| Margarine/Butter | 2.0 | 71.4 | 0.1% | 6.3% | 36 | Yogurt | 1.1 | 115.1 | 0.7% | 6.3% | 75 |

Figure 1. Store level data for an SKU in the Beer category[[4]](#footnote-10)



1. **The models**

We propose new forecasting methods which consist of three stages. At the first stage, we identify the most informative competitive explanatory variables for the focal product. Grocery retailers typically sell hundreds of SKU’s and this leads to hundreds of potential competitive explanatory variables for the focal product. Incorporating all the variables into the model would easily overfit the model and even make the estimation infeasible ([Martin & Kolassa, 2009](#_ENREF_49" \o "Martin, 2009 #623)). Therefore, we select the variables using the Least Absolute Shrinkage and Selection Operator (LASSO) ([Tibshirani, 1996](#_ENREF_72" \o "Tibshirani, 1996 #672)). For example, we have the following model:

where

represents log product sales of the focal product at week *t*  
 represents the matrix for the explanatory variables including the product price, feature, and display of all the products in the same product category

*u* represents the identically distributed error term

is the vector of the parameter coefficients  
*N* is the number of parameters which is the total number of SKUs for the category  
 is the shrinkage factor

The LASSO procedure imposes a constraint to the sum of the absolute values of all the parameter coefficients of the model. Therefore, the selection procedure removes less relevant explanatory variables by pushing their parameters towards zero. We control the model simplification process using the shrinkage factor based on 10-fold cross-validation (Ma et al., 2016)[[5]](#footnote-11). Alternative schemes including information criteria are also available (e.g., Huang et al., 2014)[[6]](#footnote-12).

At the second stage, we construct the General Autoregressive Distributive Lag (ADL) model based on the variables retained by the LASSO procedure with their dynamic terms (Huang et al., 2014). The general ADL model takes into account the dynamic effect of the (LASSO retained) marketing activities as well as potential trend, four-week seasonality, and calendar events. The general ADL model can be represented as:

where

is the log sales of the focal product at week

is the term for the deterministic trend which captures any potential steady change during the estimation period ([Song & Witt, 2003](#_ENREF_69" \o "Song, 2003 #267)).

and represent the log price of the focal product and the competitive product at week

and represents the Feature dummy for the focal product at week

is the four-week-dummy variable  
 is the dummy variable for the calendar event at week . The dummy variable represents the week of the calendar event when , , and the week before the event if . takes the values from 1 to 9 representing all the calendar events *[[7]](#footnote-13)*

are the parameters  
 is the error term and we assume

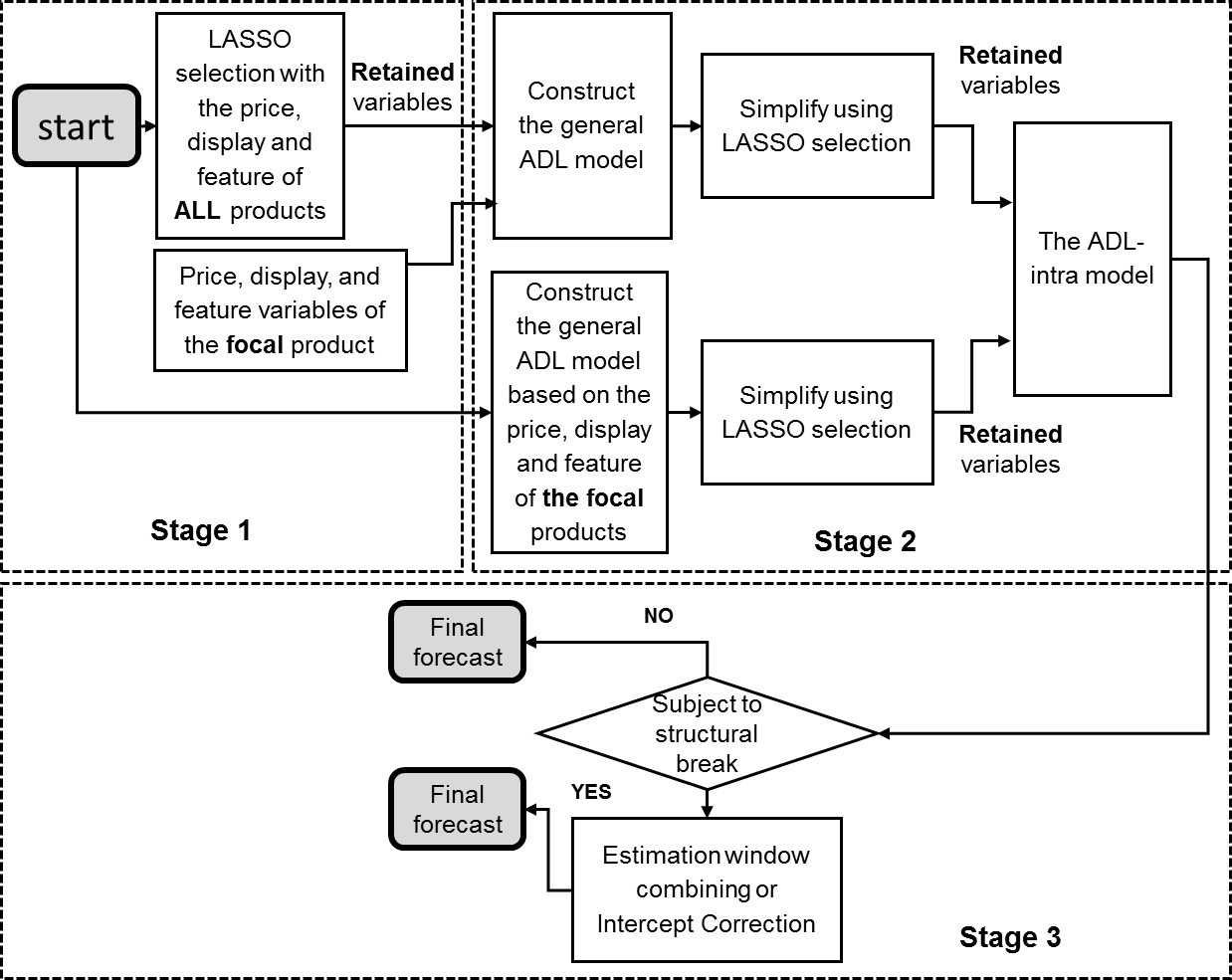
is the order of the lags and is set to as 2.

*, ,* and are the numbers of selected competitive price, Feature, and Display variables for the product category.

We further simplify the general ADL model using the LASSO procedure (we refer this simplified model as the ADL-raw model thereafter). Previous studies show that models simplified by the LASSO procedure have a good forecasting performance and outperform traditional models specified based on statistical significance ([Epprecht, Guegan, & Veiga, 2013](#_ENREF_29" \o "Epprecht, 2013 #755); [Ma et al., 2016](#_ENREF_46" \o "Ma, 2016 #733)). One limitation of the LASSO procedure is that it may potentially miss important variables under the condition of high multicollinearity ([Fan & Lv, 2008](#_ENREF_30" \o "Fan, 2008 #751); [Ma et al., 2016](#_ENREF_46" \o "Ma, 2016 #733)). In practice, retailers tend to promote relevant products at the same time, which may even increase the multicollinearity. To mitigate the issue, we construct a supplementary parallel model which only include the price and promotion variables of the focal product. For example,

We also simplify this model using the LASSO procedure (we refer this simplified model as the ADL-own model thereafter). We then combine the variables in the ADL-raw model and the variables in the ADL-own model, and we refer the model with all the combined variables as the ADL-intra model. We use the ADL-own model as a parallel model because previous studies suggest that own promotional variables are usually more important compared to variables of other products ([Bucklin, Gupta, & Siddarth, 1998](#_ENREF_13" \o "Bucklin, 1998 #752)). We, therefore, reduce the probability of (wrongfully) discard them at a cost of efficiency.

Figure 2. An illustration for the three-stage ADL-intra-EWC and the three-stage ADL-intra-IC models



At the final stage, we integrate the ADL-intra model with the EWC method and the IC method respectively to take into account the problem of structural break. We implement the EWC method and the IC method to the ADL-intra model based on the ADL-intra model if the sequential Chow test rejects the null hypothesis of no structural break. Otherwise, we keep the forecasts generated by the ADL-intra model as the final forecasts. respectivelyThe ADL-intra-EWC model and the ADL-intra-IC model may generate more accurate forecasts by taking into account the problem of structural break. 2provides a guide to implementing the ADL-intra-EWC model and the ADL-intra-IC model.

g ADLsimplified based on statistical significancetremendous amount of([L. G. Cooper et al., 1999](#_ENREF_23" \o "Cooper, 1999 #662))

1. **The experimental design**

In this study, we evaluate the forecasting performance of the following models:

1. The Base-lift method: the method is widely used by retailers in practice. It generates baseline forecasts using simple exponential smoothing method and makes adjustments for any incoming promotional event based on the lift effect of the most recent promotional event. More detailed descriptions can be found in [Gür Ali et al. (2009)](#_ENREF_38" \o "Gür Ali, 2009 #715) and Huang et al. (2014).
2. The ADL-own model: as introduced in section 6.
3. The ADL-intra model: as introduced in section 6.
4. The ADL-intra-EWC model: as introduced in section 6[[8]](#footnote-14).
5. The ADL-own-EWC model: similar to ADL-intra-EWC, except that the EWC method is integrated into the ADL-own model.
6. The ADL-intra-IC model: as introduced in section 6.
7. The ADL-own-IC model: similar to ADL-intra-IC, but the EWC method is integrated into the ADL-own model.

We evaluate the forecasting performance of these models with rolling origins for robustness ([Tashman, 2000](#_ENREF_71" \o "Tashman, 2000 #207)). For example, we specify the model with an estimation window of 160 weeks, move the estimation window forward every two weeks and have 18 rolling events. We re-specify the model using the updated estimation window and generate the forecasts. We presume the value of the price and promotion information to be known, as this is part of the retailer’s inventory plan, and we use the forecast value of the product sales when the forecast horizon is beyond one week. For each rolling event, we generate one to week-ahead forecasts, where is 1, 4, and 8, to approximate the situation retailers face in practice. For the EWC method, we estimate the ADL-intra model with ten estimation windows with different lengths (e.g., [1, 160], [3, 160], and [5 160] etc. until [19, 160]), and generate ten sets of forecasts. We then combine the ten sets of forecasts with equal weights. For the IC methods, we estimate the forecast bias as the average value of the sixteen most recent residuals and add the value equally to the forecasts for each forecast horizon.

The models are evaluated using four error measures: the Mean Absolute Error (MAE), the symmetric Mean Absolute Percentage Error (sMAPE), the Mean Absolute Scaled Error (MASE) proposed by [Hyndman and Koehler (2006)](#_ENREF_42" \o "Hyndman, 2006 #187), and the Relative Average Mean Absolute Error (RelAvgMAE) proposed by [Davydenko and Fildes (2013)](#_ENREF_26" \o "Davydenko, 2013 #744). These error measures approximate the loss function of the retailer from different aspects. The error measures for SKUs and rolling events based on forecast horizon of 1 to (i.e. , , and =1, 4 and 8) are as follows:

where and are respectively the actual value and forecast value of the forecast period for data series based on the rolling event. We add one-half mean squared error to the final forecasts before we transform the log values to levels (Cooper et al.,1999). We apply is the total number of observations in the full estimation window. and are the Mean Absolute Errors for the candidate model and the benchmark model for data series *s*, with forecast horizon of *H*, for the rolling event.

1. **Results and discussion**

In Table 2, we summarize the forecasting performance of the models across the 28 product categories. In Table 3, we also show the p-values of the Wilcoxon Sign Rank (WSR) test for the statistical significance of the difference between the models’ forecasting performance. The WSR test is the non-parametric version of the traditional t-test and does not assume the differences (i.e. the errors) are normally distributed. We find the following from the analysis of the comparisons of forecasts from the different models: (i) the Base-lift model generates the least accurate forecasts. (ii) The ADL-intra model outperforms the ADL-own model, which suggests the value of competitive promotional information [Huang et al. (2014)](#_ENREF_41" \o "Huang, 2014 #732). (iii) The ADL-own-EWC model significantly outperforms the ADL-own model. (iv) The ADL-own-IC model outperforms the ADL-own model for most of the scenarios expect for the MAE error measure. (v) The ADL-intra-EWC model significantly outperforms the ADL-intra model. (vi) The ADL-intra-IC model outperforms the ADL-intra model for all the scenarios expect for the MAE error measure. Overall, The ADL-intra-EWC model and the ADL-intra-IC model generate the most accurate forecasts[[9]](#footnote-15).

Table 2. The forecasting performance of the models for all forecast period

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| All forecast period, Forecast horizon= 8 | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank |
| Base-lift | 22.92 | 8 | 47.0% | 8 | 0.775 | 8 | 1.136 | 8 |
| ADL-own | 15.76 | 6 | 40.8% | 7 | 0.697 | 7 | 1.000 | 7 |
| ADL-intra | 15.44 | 3 | 40.5% | 4 | 0.695 | 5 | 0.991 | 4 |
| ADL-own-EWC | 15.67 | 5 | 40.7% | 5 | 0.696 | 6 | 0.996 | 6 |
| ADL-intra-EWC | 15.35 | 2 | 40.4% | 2 | 0.694 | 4 | 0.988 | 3 |
| ADL-own-IC | 16.23 | 7 | 40.8% | 6 | 0.694 | 3 | 0.997 | 5 |
| ADL-intra-IC | 15.57 | 4 | 40.4% | 3 | 0.692 | 2 | 0.988 | 2 |
| ADL-EWC-IC | 15.29 | 1 | 40.4% | 1 | 0.689 | 1 | 0.985 | 1 |
| All forecast period, Forecast horizon= 4 | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank |
| Base-lift | 22.67 | 8 | 46.2% | 8 | 0.762 | 8 | 1.106 | 8 |
| ADL-own | 15.63 | 6 | 40.5% | 7 | 0.690 | 7 | 1.000 | 7 |
| ADL-intra | 15.16 | 3 | 40.1% | 4 | 0.686 | 5 | 0.989 | 4 |
| ADL-own-EWC | 15.55 | 5 | 40.3% | 6 | 0.688 | 6 | 0.995 | 6 |
| ADL-intra-EWC | 15.08 | 2 | 40.0% | 3 | 0.685 | 4 | 0.985 | 3 |
| ADL-own-IC | 15.94 | 7 | 40.3% | 5 | 0.684 | 3 | 0.993 | 5 |
| ADL-intra-IC | 15.19 | 4 | 39.9% | 2 | 0.681 | 2 | 0.982 | 2 |
| ADL-EWC-IC | 15.00 | 1 | 39.9% | 1 | 0.679 | 1 | 0.980 | 1 |
| All forecast period, Forecast horizon= 1 | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank |
| Base-lift | 24.99 | 8 | 45.4% | 8 | 0.762 | 8 | 1.026 | 8 |
| ADL-own | 16.66 | 6 | 39.9% | 7 | 0.689 | 7 | 1.000 | 7 |
| ADL-intra | 15.66 | 4 | 39.4% | 4 | 0.686 | 5 | 0.980 | 4 |
| ADL-own-EWC | 16.58 | 5 | 39.7% | 6 | 0.686 | 6 | 0.994 | 6 |
| ADL-intra-EWC | 15.59 | 2 | 39.3% | 3 | 0.684 | 4 | 0.976 | 3 |
| ADL-own-IC | 17.01 | 7 | 39.6% | 5 | 0.681 | 3 | 0.982 | 5 |
| ADL-intra-IC | 15.60 | 3 | 39.2% | 1 | 0.678 | 1 | 0.966 | 2 |
| ADL-EWC-IC | 15.50 | 1 | 39.2% | 2 | 0.679 | 2 | 0.965 | 1 |

Table 3 . The p-values of the Wilcoxon Sign Rank (WSR) test

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Benchmark | Candidate model | MAE | | | SMAPE | | | MASE | | |
| h=1 | h=4 | h=8 | h=1 | h=4 | h=8 | h=1 | h=4 | h=8 |
| ADL-intra | ADL-intra-EWC | 0.005 | 0.001 | 0.002 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.001 |
| ADL-intra | ADL-intra-IC | 0.070 | 0.734 | 0.004 | 0.037 | 0.427 | 0.087 | 0.077 | 0.894 | 0.015 |
| ADL-own | ADL-own-EWC | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ADL-own | ADL-own-IC | 0.359 | 0.314 | 0.002 | 0.036 | 0.690 | 0.035 | 0.145 | 0.683 | 0.010 |
| ADL-own | ADL-intra | 0.025 | 0.114 | 0.259 | 0.003 | 0.009 | 0.043 | 0.021 | 0.142 | 0.286 |
| ADL-own | Base-lift | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

We also investigate the models’ forecasting performance for the time period depending on whether or not the focal product is being promoted as sales for the two periods tend to exhibit different levels of variations. Table 4 shows the forecasting performance of the models for the non-promoted period and the promoted forecast period respectively. The results are in similar compared to those shown in Table 2. For the non-promoted period, the Base-lift method generally has the worst performance except for the MASE and the AvgRelMAE when the forecast horizon is short (e.g., when h=1 and h=4). This indicates that the simple models can be difficult to beat when the focal product is not being promoted and the product sales are comparably stable ([Gür Ali et al., 2009](#_ENREF_38" \o "Gür Ali, 2009 #715); [Huang et al., 2014](#_ENREF_41" \o "Huang, 2014 #732)). The ADL-intra model generates more accurate forecasts compared to the ADL-own model. The ADL-own-EWC model and the ADL-own-IC model both outperform the ADL-own model. The ADL-intra-EWC model and the ADL-intra-IC model both outperform the ADL-intra model. For the promoted forecast period, the Base-lift method generates the least accurate forecasts. The ADL-intra model outperforms the ADL-own model. The ADL-intra-EWC model beats the ADL-intra model, and the ADL-own-EWC model beats the ADL-own-EWC model. However, the ADL-intra-IC model and the ADL-own-IC model cannot effectively outperform their counterparts (e.g., the ADL-intra-IC model and the ADL-own-IC model respectively). This may be due to the high volumes and high variations of the product sales when the focal product is being promoted. As a result, the value of the bias correction can get submerged by high variations in the data.

Table 4. The forecasting performance of the models for promoted and non-promoted period

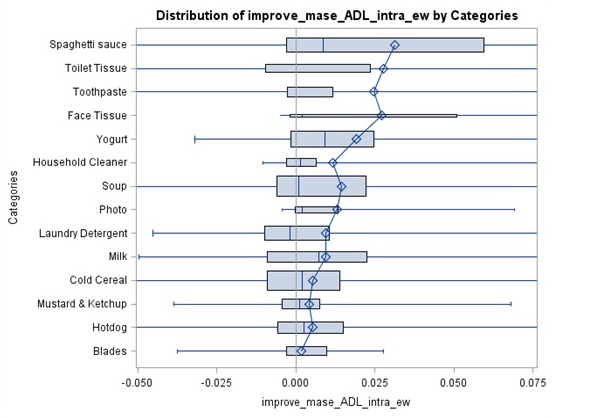
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Forecast horizon= 8 | Non-Promoted period | | | | Promoted period | | | |
| Model/measure | MAE | SMAPE | MASE | AvgRelMAE | MAE | SMAPE | MASE | AvgRelMAE |
| Base-lift | 9.64 | 41.7% | 0.588 | 1.005 | 84.46 | 82.1% | 2.186 | 1.504 |
| ADL-own | 9.36 | 40.3% | 0.582 | 1.000 | 49.49 | 49.4% | 1.653 | 1.000 |
| ADL-intra | 9.13 | 40.1% | 0.582 | 0.996 | 48.78 | 48.0% | 1.630 | 0.972 |
| ADL-own-EWC | 9.33 | 40.2% | 0.581 | 0.996 | 49.23 | 49.3% | 1.652 | 0.995 |
| ADL-intra-EWC | 9.10 | 40.0% | 0.581 | 0.994 | 48.34 | 47.9% | 1.626 | 0.965 |
| ADL-own-IC | 9.23 | 40.2% | 0.575 | 0.995 | 51.54 | 49.7% | 1.673 | 1.017 |
| ADL-intra-IC | 9.03 | 40.0% | 0.577 | 0.992 | 49.71 | 48.3% | 1.647 | 0.988 |
| ADL-EWC-IC | 9.03 | 40.0% | 0.577 | 0.992 | 48.34 | 47.9% | 1.626 | 0.965 |
| Forecast horizon= 1 | Non-promoted period | | | | Promoted period | | | |
| Model/measure | MAE | SMAPE | MASE | AvgRelMAE | MAE | SMAPE | MASE | AvgRelMAE |
| Base-lift | 9.41 | 40.9% | 0.573 | 0.990 | 85.32 | 81.9% | 2.180 | 1.531 |
| ADL-own | 9.28 | 39.9% | 0.575 | 1.000 | 49.89 | 49.1% | 1.649 | 1.000 |
| ADL-intra | 9.07 | 39.7% | 0.575 | 0.995 | 48.54 | 47.5% | 1.614 | 0.959 |
| ADL-own-EWC | 9.23 | 39.8% | 0.574 | 0.996 | 49.57 | 48.9% | 1.642 | 0.996 |
| ADL-intra-EWC | 9.04 | 39.6% | 0.574 | 0.993 | 48.21 | 47.4% | 1.608 | 0.952 |
| ADL-own-IC | 9.08 | 39.7% | 0.566 | 0.991 | 51.59 | 49.3% | 1.659 | 1.018 |
| ADL-intra-IC | 8.92 | 39.5% | 0.568 | 0.988 | 49.29 | 47.6% | 1.624 | 0.966 |
| ADL-EWC-IC | 8.92 | 39.5% | 0.568 | 0.988 | 48.21 | 47.4% | 1.608 | 0.952 |
| Forecast horizon= 1 | Non-promoted period | | | | Promoted period | | | |
| Model/measure | MAE | SMAPE | MASE | AvgRelMAE | MAE | SMAPE | MASE | AvgRelMAE |
| Base-lift | 9.43 | 39.5% | 0.562 | 0.975 | 93.64 | 85.2% | 2.220 | 1.469 |
| ADL-own | 9.19 | 39.3% | 0.567 | 1.000 | 52.93 | 48.1% | 1.632 | 1.000 |
| ADL-intra | 9.00 | 39.0% | 0.569 | 0.986 | 50.49 | 46.7% | 1.613 | 0.937 |
| ADL-own-EWC | 9.11 | 39.1% | 0.565 | 0.994 | 53.22 | 48.0% | 1.629 | 0.995 |
| ADL-intra-EWC | 8.96 | 38.9% | 0.567 | 0.984 | 50.65 | 46.5% | 1.613 | 0.929 |
| ADL-own-IC | 9.00 | 38.9% | 0.558 | 0.979 | 54.23 | 48.0% | 1.635 | 1.004 |
| ADL-intra-IC | 8.87 | 38.7% | 0.561 | 0.970 | 51.18 | 46.6% | 1.613 | 0.940 |
| ADL-EWC-IC | 8.87 | 38.7% | 0.561 | 0.970 | 50.65 | 46.5% | 1.613 | 0.929 |

In Table 5, we compare the forecasting performance of the ADL-intra model, the ADL-intra-EWC model, and the ADL-inter-IC model, for each individual product category. We select the three models because the ADL-intra-EWC model and the ADL-inter-IC model have the best forecasting performance overall and the ADL-intra model is their counterpart model which overlooks the issue of structural break. We show the forecasts based on one to eight weeks horizon for simplicity and the results for other horizons are generally consistent. Figure 3 and 4 show further details using boxplot for the MASE. In the boxplot, positive values indicate the percentage improvements by the ADL-intra-EWC model or the ADL-intra-IC model compared to the ADL-intra model. Both the ADL-intra-EWC model and the ADL-inter-IC models outperform the ADL-intra model for most of the categories. For example, the ADL-intra-EWC model outperforms the ADL-intra model for 20 out of 28 product categories. The ADL-intra-IC model outperforms the ADL-intra model for 19 product categories. The ADL-intra-EWC model and the ADL-inter-IC model do not outperform the ADL-intra model for every product category due to the heterogeneity of the data characteristics across different product categories ([e.g., Ma et al., 2016](#_ENREF_46" \o "Ma, 2016 #733)).

Table 5. Comparing forecasting performance for each product category for one to eight week forecast horizon

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ADL-intra | | |  | ADL-intra-EWC | | |  | ADL-intra-IC | | |  |
|  | MAE | MASE | SMAPE | AvgRelMAE | MAE | MASE | SMAPE | AvgRelMAE | MAE | MASE | SMAPE | AvgRelMAE |
| Beer | 5.92 | 0.729 | 52.8% | 0.993 | 5.92 | 0.728 | 52.66% | 0.992 | 5.97 | 0.731 | 52.07% | 0.994 |
| Blades | 3.86 | 0.822 | 51.5% | 0.992 | 3.84 | 0.820 | 51.56% | 0.989 | 3.81 | 0.801 | 50.59% | 0.968 |
| Carbonated Beverages | 41.48 | 0.501 | 54.0% | 0.923 | 41.05 | 0.499 | 54.43% | 0.926 | 43.61 | 0.500 | 55.99% | 0.930 |
| Cigarette | 6.56 | 0.890 | 51.1% | 0.997 | 6.53 | 0.889 | 51.10% | 0.996 | 6.27 | 0.879 | 51.25% | 0.988 |
| Coffee | 5.66 | 0.818 | 56.4% | 0.994 | 5.64 | 0.819 | 56.60% | 0.996 | 5.68 | 0.812 | 55.46% | 0.983 |
| Cold Cereal | 36.47 | 0.425 | 74.0% | 0.975 | 36.50 | 0.423 | 74.26% | 0.974 | 37.63 | 0.432 | 78.60% | 0.998 |
| Deodorant | 2.92 | 0.772 | 80.6% | 1.006 | 2.93 | 0.772 | 80.53% | 1.005 | 2.87 | 0.760 | 78.16% | 0.988 |
| Face Tissue | 13.63 | 0.576 | 31.9% | 0.914 | 13.03 | 0.566 | 31.32% | 0.897 | 13.83 | 0.579 | 33.08% | 0.917 |
| Frozen Dinner | 21.87 | 0.516 | 99.5% | 1.017 | 22.07 | 0.520 | 99.46% | 1.022 | 22.10 | 0.517 | 92.05% | 1.024 |
| Frozen pizza | 11.10 | 0.702 | 56.0% | 0.994 | 11.16 | 0.714 | 57.52% | 0.995 | 11.43 | 0.713 | 57.91% | 1.012 |
| Household Cleaner | 7.42 | 0.812 | 31.8% | 1.004 | 7.30 | 0.802 | 31.24% | 0.990 | 7.33 | 0.811 | 32.00% | 0.997 |
| Hotdog | 36.31 | 0.616 | 83.7% | 0.989 | 36.74 | 0.618 | 84.12% | 0.988 | 38.22 | 0.636 | 91.02% | 1.017 |
| Laundry Detergent | 10.91 | 0.523 | 76.0% | 0.989 | 10.68 | 0.521 | 74.84% | 0.981 | 10.77 | 0.521 | 76.86% | 0.986 |
| Margarine/Butter | 19.01 | 0.630 | 58.5% | 1.023 | 19.42 | 0.633 | 58.52% | 1.031 | 18.89 | 0.635 | 56.06% | 1.018 |
| Mayonnaise | 13.87 | 0.896 | 29.9% | 0.984 | 13.88 | 0.896 | 29.84% | 0.985 | 13.83 | 0.893 | 29.53% | 0.979 |
| Milk | 21.80 | 0.941 | 27.4% | 1.020 | 21.70 | 0.933 | 26.96% | 1.012 | 21.61 | 0.886 | 24.94% | 0.966 |
| Mustard & Ketchup | 11.08 | 0.748 | 50.9% | 0.968 | 11.04 | 0.743 | 50.24% | 0.963 | 11.20 | 0.755 | 50.28% | 0.962 |
| Peanut butter | 10.23 | 1.184 | 31.1% | 1.048 | 10.25 | 1.185 | 30.72% | 1.049 | 9.54 | 1.124 | 31.23% | 0.997 |
| Photo | 2.52 | 0.644 | 66.2% | 0.996 | 2.49 | 0.638 | 65.17% | 0.985 | 2.49 | 0.640 | 64.43% | 0.989 |
| Salty snacks | 17.12 | 0.612 | 64.9% | 1.001 | 17.25 | 0.613 | 65.05% | 0.999 | 17.07 | 0.611 | 64.97% | 0.994 |
| Shampoo | 3.83 | 0.679 | 80.7% | 1.000 | 3.81 | 0.677 | 79.89% | 0.998 | 3.57 | 0.669 | 76.39% | 0.973 |
| Soup | 19.66 | 0.562 | 80.4% | 0.979 | 19.42 | 0.556 | 79.82% | 0.970 | 20.31 | 0.580 | 80.44% | 0.997 |
| Spaghetti sauce | 11.66 | 0.781 | 46.7% | 1.000 | 11.17 | 0.768 | 45.10% | 0.963 | 11.43 | 0.770 | 45.15% | 0.978 |
| Sugar substitutes | 4.94 | 0.761 | 52.8% | 0.984 | 4.91 | 0.759 | 52.77% | 0.982 | 4.67 | 0.735 | 51.69% | 0.959 |
| Toilet Tissue | 37.95 | 0.696 | 119.5% | 0.982 | 37.14 | 0.693 | 117.99% | 0.974 | 36.69 | 0.686 | 113.88% | 0.951 |
| Toothbrush | 3.53 | 0.733 | 90.1% | 0.980 | 3.53 | 0.734 | 90.68% | 0.983 | 3.58 | 0.742 | 90.89% | 0.988 |
| Toothpaste | 28.54 | 0.807 | 218.4% | 0.996 | 26.70 | 0.793 | 216.56% | 0.980 | 26.89 | 0.809 | 211.49% | 0.985 |
| Yogurt | 30.80 | 0.807 | 26.3% | 1.008 | 30.19 | 0.793 | 25.94% | 0.993 | 29.29 | 0.771 | 25.30% | 0.971 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

Figure 3. The ADL-intra-EWC model versus the ADL-intra model, for the MASE, for h=8



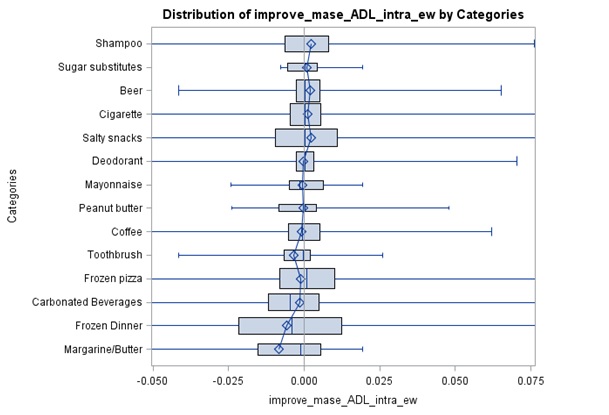
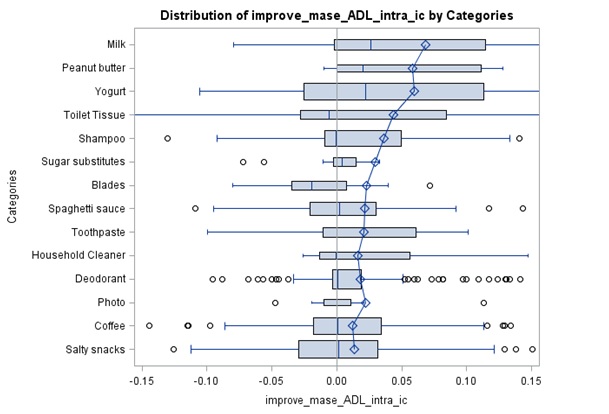
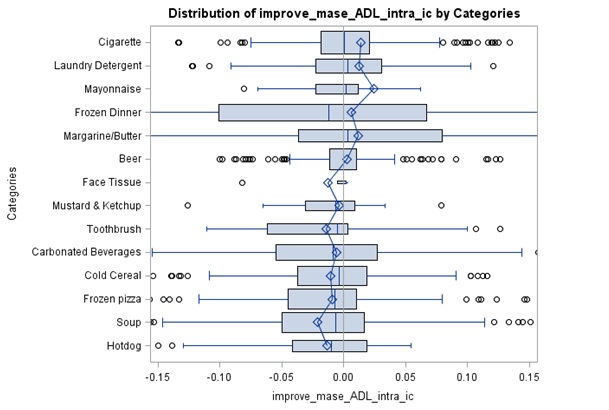


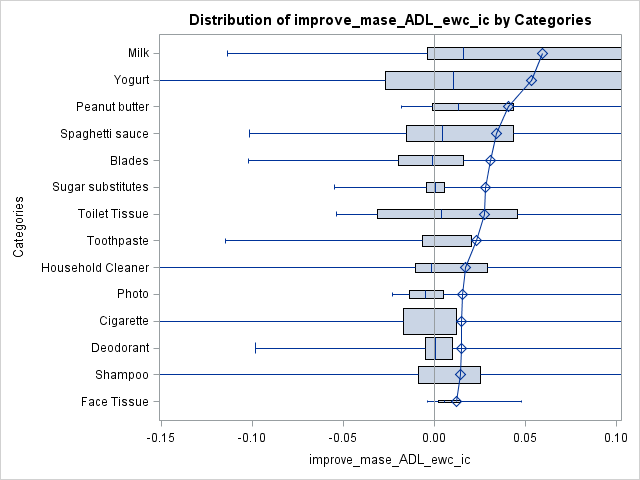
Figure 4. The ADL-intra-IC model versus the ADL-intra model, for the MASE, for h=8





The ADL-intra-IC model has the best forecasting performance for the non-promoted period but only moderate performance for the promoted period. A possible explanation is that the estimated bias used for the correction gets submerged by high variations in the high product sales. This allows us to complement the ADL-intra-IC model for the promoted period. As the ADL-intra-EWC model has the best performance for the promoted period, we forge a combined model between these two models, named as the ADL-EWC-IC model. The forecasts by the ADL-EWC-IC model will be identical to the ADL-intra-EWC model for the promoted period and the ADL-intra-IC model for the non-promoted period. Table 3 also shows the forecasting performance by the ADL-EWC-IC model. The results indicate that the ADL-EWC-IC model generates the most accurate forecasts. Table 5 also includes the performance of the ADL-EWC-IC model for the promoted and non-promoted forecast periods. In Figure 5 we depict the performance of the ADL-EWC-IC model against the ADL-intra model for each of the product categories using boxplot. The ADL-EWC-IC model outperforms the ADL-intra model for more product categories (21 out of 28) compared to either the ADL-intra-EWC model or the ADL-intra-IC model.

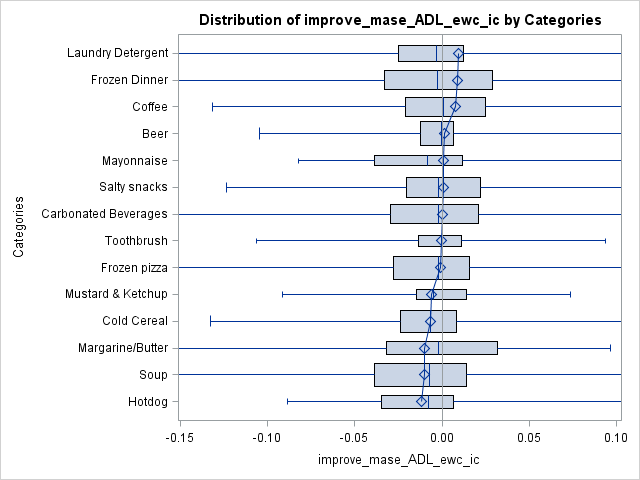
Figure 5. The ADL-EWC-IC model versus the ADL-intra model, for the MASE, for h=8











1. **Exploring the determinants of the forecasting improvement**

The results show that our proposed models generate more accurate forecasts especially for some product categories (e.g., Yogurt, Milk, Toilet Tissue etc.). We further explore the determinants of the improvement of the forecasting performance of our proposed models at SKU level. This provides insights into what types of SKUs may benefit most from the proposed models. We consider the following types of potential determinants: 1) basic statistical measures for both the prices and sales including the average, standard deviation, skewness, range, kurtosis, and coefficient of variation; 2) the frequency of the feature and display promotions for each SKU. 3) more advanced statistical measures which capture the characteristics of the data series designed by [Robert Fildes (1992)](#_ENREF_31" \o "Fildes, 1992 #198). For example, we measure the proportion of outliers for the sales of each SKU that we used in the empirical analysis. The value of the sales for product *i* will be identified as an outlier if or , where is the differenced value of the sales for product *i*. and are the first and third quantiles of . This measure may indicate the dispersion of the product sales. We measure randomness by regressing on , where is the sales value for product *i* at week *t* and *T* is the time trend. The fitness of this autoregressive model (e.g., the R square) tries to approximate the systematic variation in the sales data series which may be captured by simple models. Lastly, we measure the linear trend for the sales of the SKU as the absolute correlation between and the time trend.

We develop five orthogonal factors out of the fourteen explanatory variables to mitigate the issue of multicollinearity[[10]](#footnote-18). Table 6 shows the correlation between the original fourteen explanatory variables and the construct factors[[11]](#footnote-19). We define factor 1 as “Outliers and general variations”, factor 2 as “Sales level and variation”, factor 3 as “Central tendency of sales”, factor 4 as “Price level and variation”, and factor 5 as “Randomness and growth”. We then regress the percentage improvement by the models based on these 5 factors. For robustness, we construct the model with and without dummy variables for product categories.

Table 6. The pattern of the factors

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 |
| Proportion of outliers | 0.855 |  |  |  |  |
| Coefficient of variation (price) | 0.759 |  |  |  |  |
| Coefficient of variation (sales) | 0.714 |  |  |  |  |
| Frequency of Feature | 0.703 |  |  |  |  |
| Standard deviation of sales |  | 0.964 |  |  |  |
| Range of sales |  | 0.929 |  |  |  |
| Average sales |  | 0.807 |  |  |  |
| Frequency of Display |  | 0.281 |  |  |  |
| Kurtosis of sales |  |  | 0.973 |  |  |
| Skewness of sales |  |  | 0.881 |  |  |
| Standard deviation of price |  |  |  | 0.987 |  |
| Average price |  |  |  | 0.831 |  |
| Randomness |  |  |  |  | 0.992 |
| Trend |  |  |  |  | 0.703 |

Table 7 The determinants of improvement (MASE)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model with 5 factors | | | | | | | | |
| Horizon = 8 | ADL-intra-EWC | | ADL-own-EWC | | ADL-intra-IC | | ADL-own-IC | |
| Parameter/estimate and p-values | Estimate | P-value | Estimate | P-value | Estimate | P-value | Estimate | P-value |
| *Outliers and general variations* | 0.001 | 0.319 | 0.001 | 0.321 | -0.010 | 0.000 | -0.014 | 0.000 |
| *Sales level and variation* | 0.001 | 0.134 | 0.002 | 0.085 | -0.002 | 0.277 | -0.010 | 0.000 |
| *Central tendency of sales* | -0.001 | 0.508 | -0.001 | 0.530 | -0.007 | 0.001 | -0.008 | 0.001 |
| *Price level and variation* | -0.001 | 0.170 | -0.002 | 0.121 | 0.000 | 0.824 | -0.001 | 0.761 |
| *Randomness and growth* | 0.004 | 0.000 | 0.004 | 0.000 | 0.006 | 0.008 | 0.007 | 0.003 |
| *Intercept* | 0.003 | 0.001 | 0.003 | 0.001 | -0.002 | 0.234 | -0.004 | 0.094 |
| Model with 5 factors and category dummy variables | | | | | | | | |
| Horizon = 8 | ADL-intra-EWC | | ADL-own-EWC | | ADL-intra-IC | | ADL-own-IC | |
| Parameter/estimate and p-values | Estimate | P-value | Estimate | P-value | Estimate | P-value | Estimate | P-value |
| *Outliers and general variations* | 0.002 | 0.085 | 0.004 | 0.013 | -0.005 | 0.155 | -0.007 | 0.075 |
| *Sales level and variation* | 0.001 | 0.150 | 0.002 | 0.054 | -0.001 | 0.539 | -0.009 | 0.000 |
| *Central tendency of sales* | 0.000 | 0.679 | 0.000 | 0.851 | -0.005 | 0.044 | -0.005 | 0.047 |
| *Price level and variation* | -0.001 | 0.370 | -0.003 | 0.066 | -0.001 | 0.795 | -0.003 | 0.367 |
| *Randomness and growth* | 0.003 | 0.001 | 0.004 | 0.000 | 0.004 | 0.053 | 0.005 | 0.055 |
| *Intercept* | 0.015 | 0.001 | 0.016 | 0.001 | 0.026 | 0.015 | 0.042 | 0.001 |

Table 7 reports the estimated parameters of the models with different dependent variables. We construct the models with and without category dummy variables. For example, we regress the percentage improvement of the MASE by the ADL-intra-EWC model over the ADL-intra model on the five factors. The estimate of the parameter “Randomness and growth” is positive (e.g., 0.004) and statistical significant (e.g., p-value<0.000). For robustness, we regress the same dependent variable on the five factors with category dummy variables and we find that the estimate for “Randomness and growth” stays positive (e.g., 0.003) and statistical significant (e.g., p-value= 0.001). We also explore the determinants of the percentage improvement of the MASE by the ADL-own-EWC model (compared to the ADL-own model), the ADL-intra-IC model (compared to the ADL-intra model), and the ADL-own-IC model (compared to the ADL-own model).

We find the following from the analysis. First, the coefficients for the factor “Randomness and growth” are all positive and statistically significant, which suggests that our proposed models tend to be more advantageous for the SKU’s which are difficult to forecast and exhibiting a trend in sales. Second, the ADL-intra-IC model and the ADL-own-IC model tend to have disadvantages for the SKU’s with a high proportion of outliers and general variations and for the SKU’s with the high central tendency of sales. This may indicate that the ‘intercept correction’ for the bias can be submerged by high sales pikes which are usually ‘outliers’ and caused by promotions. For simplicity, we only show the results for the MASE and when the horizon is one to eight weeks ahead. The results are consistent across all the error measures and all the forecast horizons. This indicates that we may pre-test these features of the SKU and then determine the optimal sales forecasting method specifically for that SKU.

1. **Conclusions, limitations and future research**

Grocery retailers need to effectively manage their inventory and, to achieve that, they rely use forecasting models and welcome new approaches that will enable them to improve on what they are currently doing. Recent studies focus on incorporating more information (e.g., Gur Ali et al., 2009; Huang et al., 2014; Ma et al. (2016). however, they all assume a constant effect of marketing activities such as price reductions and feature and display promotions which may actually change over time because of the impact of external factors including a change in economic conditions, the change of the consumer taste, and new competition entry etc. However, the data on these factors is not always available to incorporate into the model. Or, we do not actually know which of these external factors are actually causing the structural break. As a result, the conventional models with the data that is used in building these models will be subject to a structural break and potentially generate biased and consequently produce less accurate forecasts.

Our research focuses on how to mitigate the problem with data of the marketing activities which retailers typically have control over. We propose a set of models which take into account the potential forecast bias caused by a structural break. The ADL-intra-EWC model generates forecasts which are the combination of various sets of forecasts by the ADL-intra model with different estimation windows under the condition where structural breaks are detected. It tries to obtain an effective trade-off between the forecast bias and the forecast error variance. The ADL-intra-IC model tries to offset the potential forecast bias by adding the estimate of the forecast bias back to the error term at a cost of inflated forecast error variance when structural breaks are detected. Based on our experiments, we find that these models outperform the ADL-intra model across all the 28 product categories. Table 7 shows the percentage reductions of various error measures by these models compared to compared to different benchmark models. For example, for the forecast horizon of one to eight weeks ahead, the ADL-intra-EWC model reduces the SMAPE of the ADL-intra model by 0.22% and reduces the SMAPE of the Base-lift model by 13.97%. The ADL-intra-IC model reduces the SMAPE of the ADL-intra model by 0.18% and reduces the SMAPE of the Base-lift model by 13.94%. At the category level, the ADL-intra-EWC model and the ADL-intra-IC model have superior forecasting performance compared to the ADL-intra model for most of the product categories.

We also observe that the ADL-intra-EWC model has the best performance for the promoted forecast period while the ADL-intra-IC model dominates the non-promoted forecast period. We, therefore, combine the ADL-intra-EWC model with the ADL-intra-IC model based on whether or not the focal product is being promoted. The so-called ADL-EWC-IC model thus generates the most accurate forecasts across all the candidate models. The ADL-EWC-IC model has superior forecasting performances compared to the ADL-intra model for 21 out of 28 product categories.

We also evaluate the forecasting performance of the ADL-own-EWC model and the ADL-own-IC model. These methods are especially valuable to manufacturers since, under certain circumstances, competitive promotional information may not be available to them ([Ali & Boylan, 2011](#_ENREF_1" \o "Ali, 2011 #742)). In our experiment, the ADL-own -EWC model and the ADL-own -IC model both outperform the ADL-own model across all the product categories. Table 8 also shows the percentage reductions of various error measures by these models compared to compared to different benchmarks. For example, for the forecast horizon of one to eight weeks head, the ADL-own-EWC model reduces the SMAPE of the ADL-own model by 0.31% and reduces the SMAPE of the Base-lift model by 13.40%. The ADL-own-IC model reduces the SMAPE of the ADL-own model by 0.15% and reduces the SMAPE of the Base-lift model by 13.26%.

Table 8. Forecasting performance regarding percentage reductions in various error measures

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Forecast horizon | Proposed model | Benchmark | percentage of increase/decrease | | | | |
| MAE | SMAPE | MASE | AvgRelMAE |
| h=8 | ADL-intra-EWC | ADL-intra | -0.57% | -0.22% | -0.18% | -0.33% |
| ADL-intra-IC | ADL-intra | 0.86% | -0.18% | -0.45% | -0.36% |
| ADL-own-EWC | ADL-own | -0.52% | -0.31% | -0.17% | -0.42% |
| ADL-own-IC | ADL-own | 3.00% | -0.15% | -0.52% | -0.32% |
| ADL-intra | ADL-own | -2.02% | -0.75% | -0.30% | -0.86% |
| ADL-own-EWC | Base-lift | -31.61% | -13.40% | -10.22% | -12.36% |
| ADL-own-IC | Base-lift | -29.20% | -13.26% | -10.53% | -12.27% |
| ADL-intra-EWC | Base-lift | -33.03% | -13.97% | -10.49% | -13.04% |
| ADL-intra-IC | Base-lift | -32.07% | -13.94% | -10.74% | -13.06% |
|  | ADL-EWC-IC | Base-lift | -33.27% | -14.02% | -11.09% | -13.32% |
| h=4 | ADL-intra-EWC | ADL-intra | -0.54% | -0.34% | -0.26% | -0.46% |
| ADL-intra-IC | ADL-intra | 1.96% | -0.49% | -0.97% | -0.72% |
| ADL-own-EWC | ADL-own | -0.54% | -0.34% | -0.26% | -0.46% |
| ADL-own-IC | ADL-own | 1.96% | -0.49% | -0.97% | -0.72% |
| ADL-intra | ADL-own | -3.03% | -0.84% | -0.57% | -1.13% |
| ADL-own-EWC | Base-lift | -31.42% | -12.82% | -9.62% | -9.97% |
| ADL-own-IC | Base-lift | -29.70% | -12.95% | -10.25% | -10.21% |
| ADL-intra-EWC | Base-lift | -33.47% | -13.45% | -10.07% | -10.88% |
| ADL-intra-IC | Base-lift | -32.99% | -13.66% | -10.62% | -11.21% |
|  | ADL-EWC-IC | Base-lift | -33.85% | -13.70% | -10.88% | -11.36% |
| h=1 | ADL-intra-EWC | ADL-intra | -37.33% | -13.17% | -10.04% | -4.55% |
| ADL-intra-IC | ADL-intra | -37.63% | -13.38% | -10.27% | -4.88% |
| ADL-own-EWC | ADL-own | -0.47% | -0.36% | -0.44% | -0.58% |
| ADL-own-IC | ADL-own | 2.11% | -0.80% | -1.22% | -1.80% |
| ADL-intra | ADL-own | -6.01% | -1.10% | -0.51% | -2.03% |
| ADL-own-EWC | Base-lift | -33.64% | -12.52% | -9.97% | -3.14% |
| ADL-own-IC | Base-lift | -31.92% | -12.91% | -10.68% | -4.33% |
| ADL-intra-EWC | Base-lift | -37.63% | -13.38% | -10.27% | -4.88% |
| ADL-intra-IC | Base-lift | -37.59% | -13.75% | -10.99% | -5.85% |
|  | ADL-EWC-IC | Base-lift | -37.97% | -13.74% | -10.97% | -6.02% |

We also explore the relationship between the relative advantage of the proposed models and the data characteristics of the product SKU. We find that the models with the estimation window combining (e.g., the ADL-intra-EWC model and the ADL-own-EWC model)have better forecasting performances compared to their counterparts for the SKU’s with high randomness and trend, while the models with intercept corrections (e.g., the ADL-intra-IC model and the ADL-own-IC model) tend to have more advantages compared to their counterparts for the SKU’s with high randomness and trend, with low proportion of outliers and low level of general variations, and with low level of sales kurtosis and skewness.

The approach that we propose here is new to the area of SKU forecasting but we have also identified some areas where we feel there can be further improvements to the forecasting models that we have described in the paper. For example, for the EWC method, we combine five sets of forecasts based on ten different estimation windows using equal weights. The forecasting performance may potentially be improved by changing the number of the estimation windows, by changing the minimum length of the estimation windows, and by exploring alternative forecasting combination schemes (e.g., based on k-fold evaluation). For the IC method, Clements and Hendry (1999) summarize various correction schemes each of which may have different effects on the trade-off between the bias and the error variance[[12]](#footnote-20). Furthermore, Ma et al. (2016) propose models which integrate both the intra- and the inter-category promotional information. We may investigate if we can further improve the forecasting performance of the ADL-intra-EWC model and the ADL-intra-IC model with inter-category information. 3) A method alternative to the EWC method and the IC method is directly modeling the changing process of the effect of the marketing activities into the model so that the structural break may potentially be eliminated or be mitigated even when the influencing factors are not observed. [Foekens, Leeflang, and Wittink (1999)](#_ENREF_36" \o "Foekens, 1999 #145) modeled the effect of the price variables using previous prices and the recency and the frequency of previous prices. The models are for descriptive purposes and not evaluated in terms of forecasting performance. However, one of the challenges for this type of model is the complexity and potential convergence issue in estimation.

**Acknowledgement**

We thank the IRI company for making the data available. All the analysis and findings in this paper based on the IRI dataset are by the author and not by IRI company.

**Appendix A:**

In this Appendix, we illustrate the impact of structural break on forecasting accuracy with an example using simulation. For example, we construct a price variable with its values being 2.99 for most of the observations (say, weeks) but occasionally reduced to 2.29 or 1.99[[13]](#footnote-21). We assume the following true product sales[[14]](#footnote-22):

, , when

, , when

, , when

where and represent the product sales and the price at week *t*, and is the error term. There are two structural breaks for the model: the parameter of the price changes at week 25 and then at week 50. In practice, this may be due to new product introduction (which reduces the price elasticity of the focal product) and a credit crunch (e.g., customers become more price sensitive). The sales and price are represented in Figure A1 by the solid black line and the solid red line respectively.

Figure A1. Simulated sales with a structural break: model with post-break data



Suppose that we have the data from week 1 to week 75 and we forecast the product sales for the period from week 76 to week 100. If we know that there are changes for the effect of the product price, we may develop a congruent model (i.e., ) exclusively based on the post-break data (i.e., data from week 51 to week 75) and generate unbiased forecasts. Figure A1 represents the predictions/forecasts using the black dashed line (e.g., *ybar\_post breaks*). Table A1 shows the forecasting performance of the model with post break data (e.g., with MAE= 0.3, MSE= 0.18, MAPE= 5.0%, and SMAPE= 4.3%).

However, the changes of the effect of the price are usually unknown. If we overlook the two structural breaks and estimate the model using all the available data (i.e., from week 1 to week 75), we would obtain estimates of the parameters as the weighted average of the true parameters before and after the breaks and generate biased forecasts. In this example, we tend to over-predict the sales from week 1 to week 25, under-predict the sales from week 26 to week 50, then again over-predict the sales from week 51 to week 70 and finally generate downwards-biased out-of-sample forecasts from week 76 to week 100. Figure 2 shows the biased predictions/forecasts with the black dashed line (as *ybar\_all data*). Table 1 shows the forecasting performance of the model with the full data (e.g., with MAE= 0.7, MSE= 0.52, MAPE= 12.2%, and SMAPE= 11.5%). The forecasts are substantially inferior compared to the model with post-break data.

Figure A2. Simulated sales with a structural break: model with full data[[15]](#footnote-23)



Table A1. The forecasting performance of different models in the simulation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MAE | MSE | MAPE | SMAPE |
| Model with post-break data | 0.3 | 0.18 | 5.0% | 4.3% |
| Model with full data | 0.7 | 0.52 | 12.2% | 11.5% |
| Model with full data, with IC | 0.1 | 0.01 | 1.7% | 1.8% |
| Model with full data, with EWC | 0.6 | 0.43 | 11.0% | 10.5% |

**Appendix B:**

Based on the same example in Appendix A, we may improve the accuracy of the forecasts using the intercept correction (IC) method. First, we construct a congruent model as but presuming no priori knowledge of the structural breaks. We conduct a sequential [Chow (1960)](#_ENREF_16" \o "Chow, 1960 #260) test based on every observation during the estimation period[[16]](#footnote-24). The rejection of the null hypothesis of no structural break for any of the observations would suggest that the model is subject to structural break though without indicating how many structural breaks and their locations. Figure B1 shows the *p*-values of the sequential Chow test. The results reject the null hypothesis of no structural break (especially for weeks closed to week 25 and week 50) [[17]](#footnote-25). More advanced tests are available (e.g., considering multiple breaks, heteroskedasticity, and unit roots etc.) are available but require additional priori knowledge such as the number of potential structural breaks ([Donald W K Andrews, 1993](#_ENREF_3" \o "Andrews, 1993 #237); [Donald W. K. Andrews & Ploberger, 1994](#_ENREF_4" \o "Andrews, 1994 #238); [Bai & Perron, 1998](#_ENREF_9" \o "Bai, 1998 #227), [2003](#_ENREF_10" \o "Bai, 2003 #657)).

Figure B1 *P*-values of the sequential Chow test



We confirm that the model is subject to structural break and assume the forecasts are biased. We may estimate the forecast bias as the average value of an ad hoc number (e.g., we may choose 4) of the errors close to the forecast origin. e.g., where is the estimated forecast bias. We can obtain the final corrected forecasts by add the estimated bias back to the forecasts by the original model, e.g., , where are the final forecasts by the IC model. Figure B2 shows the predictions/forecasts with the black dashed line (as *ybar\_IC*). Table A1 shows the forecasting performance of the model with the full data (e.g., with MAE= 0.1, MSE= 0.01, MAPE= 1.7%, and SMAPE= 1.8%). The intercept corrected model substantially outperforms the model with the full data.

Figure B2 Simulated Sales with a structural break: model with intercept correction



We may improve the accuracy of the forecasts using the estimation window combining (EWC) method. We first conduct the sequential Chow test. As the results suggest that the model is subject to structural break, we combine the forecasts by the same model but with different estimation windows. For example, we may u maya second set of pre-break may , which gives us 60 sets of forecasts. Thus wof these 60 where are the final forecasts by the EWC model. Figure B3 represents the predictions/forecasts with the black dashed line (as *ybar\_EWC*). Table A1 shows the forecasting performance of the model with the full data (e.g., 0.6 for MAE, 0.43 for MSE, 11.0% for MAPE, and 10.5% for SMAPE). The EWC method outperforms the conventional model with the full data.

B3

Reference**:**

Ali, M., & Boylan, J. (2011). Feasibility principles for Downstream Demand Inference in supply chains. *Journal of the Operational Research Society, 62*.

Allen, P. G., & Fildes, R. (2001). Econometric forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Boston: Kluwer Academic Publishers.

Andrews, D. W. K. (1993). Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica, 61*, 825-851.

Andrews, D. W. K., & Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica, 62*, 1383-1414.

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*(1), 22-33.

Ang, A., & Bekaert, G. (2002). Regime Switches in Interest Rates. *Journal of Business & Economic Statistics, 20*(2), 163-182.

Arenas, T., Pedregal, J., J., D., Fildes, R., & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting, 29*(2).

Armstrong, J. S. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*: Kluwer Academic Publishers.

Bai, J., & Perron, P. (1998). Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica, 66*, 47- 78.

Bai, J., & Perron, P. (2003). Computation and Analysis of Multiple Structural-Change Models. *Journal of Applied Econometrics, 18*, 1-22.

Blattberg, R. C., Briesch, R., & Fox, E. J. (1995). How promotions work? *Marketing Science, 14*(3).

Bronnenberg, B. J., Kruger, M. W., & Mela, C. F. (2008). The IRI Marketing Data Set. *Marketing Science, 27*(4), pp. 745–748.

Bucklin, R. E., Gupta, S., & Siddarth, S. (1998). Determining Segmentation in Sales Response across Consumer Purchase Behaviors. *Journal of Marketing Research, 35*(2), 189-197. doi: 10.2307/3151847

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2008). Model Selection when there are Multiple Breaks. *Working paper No. 407, Economics Department, University of Oxford*.

Chevillon, G. (2016). Multistep forecasting in the presence of location shifts. *International Journal of Forecasting, 32*(1), 121-137. doi: <https://doi.org/10.1016/j.ijforecast.2015.04.004>

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica, 28*(3).

Clark, T. E., & McCracken, M. W. (2007). Forecasting with Small Macroeconomic VARs in the Presence of Instabilities *Finance and Economics Discussion Series*

Divisions of Research & Statistics and Monetary Affairs

Federal Reserve Board, Washington, D.C.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583. doi: <http://dx.doi.org/10.1016/0169-2070(89)90012-5>

Clements, M., & Hendry, D. (1998). *Forecasting Economic Time Series*: Cambridge University Press.

Clements, M. B., & Hendry, D. F. (1994). Towards a theory of economic forecasting. In C. P. Hargreaves (Ed.), *Nonstationary Time Series Analysis and Cointegration*: Oxford University Press.

Clements, M. P., & Hendry, D. F. (1999). *Forecasting non-stationary economic time series*. London: The MIT Press.

Cooper, J. P., & Nelson, C. R. (1975). The Ex Ante Prediction Performance of the St. Louis and FRB-MIT-PENN Econometric Models and Some Results on Composite Predictors. *Journal of Money, Credit and Banking, 7*(1).

Cooper, L. G., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). Promocast": a New Forecasting Method for Promotion Planning. *Marketing Science, 18*(3), 301-316. doi: <http://dx.doi.org/10.1287/mksc.18.3.301>

Cooper, L. G., & Giuffrida, G. (2000). Turning Datamining into a Management Science Tool: New Algorithms and Empirical Results. *Management Science, 46*(2), 249.

Corsten, D., & Gruen, T. (2003). Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out-of-stocks. *International Journal of Retail & Distribution Management, 31*(12).

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

Dekimpe, M., Hanssens, D. M., & Silva-Risso, J. M. (1999). Long-run effects of price promotions in scanner markets. *Journal of Econometrics, 89*(1/2), 261-291.

Dekker, M., van Donselaar, K., & Ouwehand, P. (2004). How to use aggregation and combined forecasting to improve seasonal demand forecasts. *International Journal of Production Economics, 90*(2), 151-167.

Epprecht, C., Guegan, D., & Veiga, Á. (2013). Comparing variable selection techniques for linear regression: LASSO and Autometrics: Université Panthéon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne.

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

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting, 8*, 81-98.

Fildes, R., & Goodwin, P. (2007). Fine judgements: do organizations follow best practice when applying management judgement to forecasting? *Interfaces, 37*, 570-576.

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*(1), 3-23.

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

Fildes, R., & Stekler, H. (2002). The state of macroeconomic forecasting. *Journal of Macroeconomics, 24*(4), 435-468. doi: Pii S0164-0704(02)00055-1

Foekens, E. W., Leeflang, P., & Wittink, D. R. (1999). Varying parameter models to accommodate dynamic promotion effects. *Journal of Econometrics, 89*(1-2), 249-268.

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega, 30*(2), 127-135. doi: Doi: 10.1016/s0305-0483(01)00062-7

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).

Hendry, D. F. (1995). *Dynamic Econometrics: Advanced Texts in Econometrics*. Oxford, UK: Oxford University Press.

Houston, F. S., & Weiss, D. L. (1975). CUMULATIVE ADVERTISING EFFECTS: THE ROLE OF SERIAL CORRELATION\*. *Decision Sciences, 6*(3), 471-481. doi: 10.1111/j.1540-5915.1975.tb01036.x

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. doi: <http://dx.doi.org/10.1016/j.ejor.2014.02.022>

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

Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting, 24*(1), 163-169. doi: <http://dx.doi.org/10.1016/j.ijforecast.2007.06.001>

Little, J. D. C. (1966). A Model of Adaptive Control of Promotional Spending. *Operations research, 14*(6).

Loeb, W. (2015). Unrelenting Competition: The Biggest Retail Story of 2015, 2016

Ma, S., Fildes, R., & Huang, T. (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. doi: <http://dx.doi.org/10.1016/j.ejor.2015.08.029>

Mace, S., & Neslin, S. A. (2004). The determinants of pre- and postpromotion dips in sales of frequently purchased goods. *Journal of Marketing Research, XLI*, 339-350.

Mahajan, V., Bretschneider, S. I., & Bradford, J. W. (1980). Feedback Approaches to Modeling Structural Shifts in Market Response. *Journal of Marketing, 44*, 71-80.

Martin, R., & Kolassa, S. (2009). *Challenges of Automated Forecasting in Retail.* Paper presented at the International Symposium on Forecasting, Hong Kong.

Meeran, S., Jahanbin, S., Goodwin, P., & Quariguasi Frota Neto, J. (2017). When do changes in consumer preferences make forecasts from choice-based conjoint models unreliable? *European Journal of Operational Research, 258*(2), 512-524. doi: <https://doi.org/10.1016/j.ejor.2016.08.047>

Moinpour, R., McCullough, J. M., & MacLachlan, D. L. (1976). Time Changes in Perception: A Longitudinal Application of Multidimensional Scaling. *Journal of marketing research, 13*(3), 245-253.

Monroe, K. B., & Guiltinan, J. P. (1975). A Path-Analytic Exploration of Retail Patronage Influences. *The Journal of Consumer Research, 2*(1), 19-28.

Morrison, D. G. (1966). Interpurchase Time and Brand Loyalty. *Journal of Marketing Research, 3*.

Muellbauer, J. (1994). The Assessment: Consumer Expenditure. *Oxford Review of Economic Policy, 10*(2), 1-41.

Myers, J. G. (1971). The Sensitivity of Dynamic Time-Path Typologies. *Journal of marketing research, 8*(4), 472-479.

Myers, J. G., & Nicosia, F. M. (1970). Time-Path Types: From Static to Dynamic Typologies. *Management Science, 16*(10), B584-B596.

Nijs, V. R., Dekimpe, M. G., Steenkamps, J.-B. E. M., & Hanssens, D. M. (2001). The Category-Demand Effects of Price Promotions. *Marketing Science, 20*(1), 1-22.

Nikolopoulos, K. (2010). Forecasting with quantitative methods: the impact of special events in time series. *Applied Economics, 42*, 947-955.

OrderDynamics. (2015). Retailers and the Ghost Economy: The Haunting of Returns. <http://engage.dynamicaction.com/WS-2015-06-IHL-Ghost-Economy-Haunting-of-Returns-AR_LP.html>.

Ouyang, Y. (2007). The effect of information sharing on supply chain stability and the bullwhip effect. *European Journal of Operational Research, 182*, 1107-1121.

Perez-Quiros, G., & Timmermann, A. (2000). Firm Size and Cyclical Variations in Stock Returns. *The Journal of Finance, 55*(3), 1229-1262. doi: 10.1111/0022-1082.00246

Pesaran, H. M., & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics, 137*, 134-161.

Pesaran, M. H., & Pick, A. (2011). Forecast Combination Across Estimation Windows. *Journal of Business & Economic Statistics, 29*(2), 307-318. doi: 10.1198/jbes.2010.09018

Pesaran, M. H., Schuermann, T., & Smith, L. V. (2009). Forecasting Economic and Financial Variables with Global VARs. *International Journal of Forecasting, 25*, 642-675.

Pesaran, M. H., & Timmermann, A. (2002). Market timing and return prediction under model instability. *Journal of Empirical Finance, 9*(5), 495-510. doi: <http://dx.doi.org/10.1016/S0927-5398(02)00007-5>

Pesaran, M. H., & Timmermann, A. (2005). Small sample properties of forecasts from autoregressive models under structural breaks. *Journal of econometrics, 129*(1-2), 183-217. doi: DOI: 10.1016/j.jeconom.2004.09.007

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). ‘Horses for Courses’ in demand forecasting. *European Journal of Operational Research, 237*(1), 152-163. doi: <http://dx.doi.org/10.1016/j.ejor.2014.02.036>

Sodhi, M. S., & Tang, C. S. (2011). The incremental bullwhip effect of operational deviations in an arborescent supply chain with requirements planning. *European Journal of Operational Research, 215*, 374-382.

Song, H., & Witt, S. F. (2003). Tourism Forecasting: The General-to-Specific Approach. *Journal of Travel Research, 42*, 65-74.

Stock, J. H., & Watson, M. W. (1996). Evidence on Structural Instability in Macroeconomic Time Series Relations. *Journal of Business and Economic Statistics, 14*.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review *International Journal of Forecasting, 16*(4).

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

Trusov, M., Bodapati, A. V., & Cooper, L. G. (2006). Retailer Promotion Planning: Improving Forecasting Accuracy And Interpretability. *Journal of Interactive Marketing, 20*(3-4), 71-81.

Van 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.

Van Heerde, H. J., Srinivasan, S., & Dekimpe, M. G. (2008). Decomposing the Demand for a Pioneering Innovation. *Working paer, University of Waikato, Department of Marketing*.

Wedel, M., & Zhang, J. (2004). Analyzing brand competition across subcategories. *Journal of Marketing Research, 41*(4), 448-456.

Wichern, D. W., & Jones, R. H. (1977). Assessing the Impact of Market Disturbances Using Intervention Analysis. *Management Science, 24*(3), 329-337.

Wildt, A. R. (1976). The empirical investigation of time dependent parameter variation in marketing models. In E. proceedings (Ed.), *American Marketing Association* (pp. 466-472).

Wildt, A. R., & Winer, R. S. (1983). Modeling and Estimation in Changing Market Environments. *The Journal of Business, 56*(3).

Winer, R. S. (1979). An Analysis of the Time-varying Effects of Advertising: The Case of Lydia Pinkham. *The Journal of Business, 52*(4).

Wittink, D., Addona, M., Hawkes, W., & Porter, J. (1988). *SCAN\*PRO: the estimation, validation and use of promotional effects based on scanner data*. Internal paper. Internal Paper. Cornell University.

1. Corresponding author at Surrey Business School, University of Surrey, GU2 7XH, UK. Tel: 01483 68 6359, email: [t.huang@surrey.ac.uk](mailto:t.huang@surrey.ac.uk); [r.fildes@lancaster.ac.uk](mailto:r.fildes@lancaster.ac.uk) (r.Fildes); [d.Soopramanien@lancaster.ac.uk](mailto:d.Soopramanien@lancaster.ac.uk) [↑](#footnote-ref-1)
2. Analytical evidence for the models with endogenous explanatory variables can be found in Clements and Hendry (1999) and Pesaran and Timmerman (2005, 2007). [↑](#footnote-ref-2)
3. We select the SKU’s with positive movements for at least 90% of the time. [↑](#footnote-ref-9)
4. In Figure 6, the calendar events include Halloween, Thanksgiving, Christmas, New Year’s Day, President’s Day, Easter, Memorial Day, the 4th of July, and Labour Day. The promotional events include Feature and Display. [↑](#footnote-ref-10)
5. Compared to cross validation, the 10-fold cross validation repeat the validation for multiple times and ensure all the observations are used for training and validation. [↑](#footnote-ref-11)
6. We find little difference between these two schemes. [↑](#footnote-ref-12)
7. We include the following US calendar events including *Halloween*, *Thanksgiving*, *Christmas*, *New Year’s Day*, *President’s Day*, *Easter*, *Memorial Day*, the *4th of July*, and *Labour Day*. [↑](#footnote-ref-13)
8. We conduct the sequential Chow test and find the models for 99.89% of SKU’s are subject to structural break. [↑](#footnote-ref-14)
9. The ADL-EWC-IC model in Table 3 will be discussed in later sections. [↑](#footnote-ref-15)
10. With 5 factors, we retain 90% of the variations. [↑](#footnote-ref-18)
11. In Table 6, we omit all small values for simplicity. [↑](#footnote-ref-19)
12. For example, one of the alternative is to first make adjustments to the one-step-ahead forecast, and then calculate the two-step-ahead forecast based on the value of the one-step-ahead forecast which has adjusted, and so forth. [↑](#footnote-ref-20)
13. This setting is typical in a retailer context. In this example, we artificially make up the data series (but we keep the data series to be stationary). [↑](#footnote-ref-21)
14. In this example, for simplicity, we choose to illustrate the impact of structural breaks on forecasting accuracy using two structural breaks and also by holding the error variance to be constant before and after the breaks. Alternative settings (e.g., with different number of structural breaks and with changing error variance before and after the structural breaks) would provide the same indication. [↑](#footnote-ref-22)
15. In Figure 1, we use the blue area to represent the period before the first structural break (e.g., week [1,25]), use the yellow area to represent the period after the second structural break until the forecast origin (e.g., week [51, 75]), use the green area to represent the period between the two structural breaks (e.g., [26, 50]), and we use the red area to represent the forecast period (e.g., week [76, 100]). [↑](#footnote-ref-23)
16. The Chow test is a variant of F-test which compares the fitting of the model before and after the structural break. It assumes the locations of one structural break known a priori and also invariant error variations before and after the break. For a sequential Chow test, we conduct the Chow test assuming the break occurs at each individual week. For example, we may conduct the Chow test assuming there is a structural break at a specific week (e.g., week 10) and we obtain the p-value, and so forth. We plot the p-values for all the weeks in Figure 3 excluding the first and the last a few weeks. [↑](#footnote-ref-24)
17. We would consider the model not being subject to structural breaks only when all the p-values are above the threshold. To mitigate the multiple comparison problem, we adopt very small threshold (e.g., 0.001 rather than the usual 0.05) for the p-values. [↑](#footnote-ref-25)