Automatic Interpretable Retail Forecasting with Promotional Scenarios

Özden Gür Ali[*] [†] **and Ragıp Gürlek**[‡]

Business Administration, Koç University, College of Administrative Sciences and Economics, Rumeli Feneri Yolu, Sariyer, 34450, Istanbul, TURKEY

Abstract

Budgeting and planning processes require medium-term sales forecasts with marketing scenarios. The complexity in modern retailing necessitates consistent, automatic forecasting and insight generation. Remedies to the high dimensionality problem have drawbacks: Black box machine learning methods require voluminous data and lack insights, while regularization may bias causal estimates in interpretable models.

The proposed FAIR (Fully Automatic Interpretable Retail forecasting) method supports the retail planning process with multi-step-ahead category-store level forecasts, scenario evaluation and insights. It considers category-store specific seasonality, focal- and cross-category marketing, and adaptive base sales while dealing with regularization induced confounding.

We show with three chains from the IRI dataset involving 30 categories that regularization induced confounding decreases forecast accuracy substantially. Including focal- and cross-category marketing, and random disturbances, all increase forecast accuracy. FAIR has better accuracy than the black box machine learning method Boosted Trees and other benchmarks, while providing insights that are in line with the marketing literature.

Keywords: causality, decomposition, marketing, multivariate time series, panel data, machine learning

[*]Corresponding author.
[†]E-mail: oali@ku.eu.tr
[‡]E-mail: rgurlek16@ku.edu.tr

1 Introduction

Retailers use medium-term category-store level forecasts as an input to the master planning process that involves decisions about budgeting and allocation of resources, coordination of negotiations with suppliers, promotion and personnel planning (Hübner et al., 2013). Promotion effects can be difficult to forecast and judgmental adjustments to baseline forecast may be biased (Fildes et al., 2019). There is a potential for conflict and lack of coordination among functions in the absence of a formal planning process with quality information, process and alignment (Oliva & Watson, 2011). It can be argued that category and store level predictions are particularly relevant for retailers as they are mainly interested in the performance of the overall category rather than individual brands or SKUs (Ailawadi et al., 2009). SKU forecasts, on the other hand, are critically important for inventory decisions. Category-store level sales forecasting at a retail chain usually involves thousands of sales and marketing time series, making manual model specification almost impossible. Supporting the planning process requires an objective tool to forecast category-store level sales that can be trusted to perform scenario analyses with planning elements. There are three further issues that complicate this task.

First, the large number of categories, marketing variables of focal[1] and other categories, and category-store specific seasonality introduce high dimensionality[2] (Feng & Shanthikumar, 2018). The use of online reviews, social media, or demand data from different markets and regions to enhance forecasting accuracy magnifies the high dimensionality issue (Choi et al., 2018; Cui et al., 2018). High dimensionality can cause overfitting where the model fits irrelevant "noise" or overcompensates for relevant inputs resulting in poor out-of-sample accuracy, and lack of computational scalability as the number of categories and input variables increase.

Machine learning methods, particularly regression trees and their bagged and boosted ensembles, constitute one of the remedies for dealing with the high dimensionality of inputs and enhancing sales forecasting accuracy (e.g., Cui et al., 2018; Ferreira et al.,

[1] Category sales are expected to be affected by promotions in the same category and other categories. Focal category effects refer to the former, while cross-category effect refers to the latter.

[2] IRI/Nielsen lists more than 300 categories, while a focused retailer could have around thirty categories. The granularity of category definition, and therefore the number of categories varies by retailer. http://www.supermarketnews.com/center-store/2015-data-table-supermarket-categories-dollar-unit-sales?full=1

2015; Gür Ali et al., 2009). A recent Kaggle contest featured a challenge to forecast daily sales for 1115 stores and 6 weeks based on promotions, competition, holidays and locality for Rossman drugstores as an input to staff scheduling. The evaluation was based only on the accuracy of the forecasts and the winning strategy employed an ensemble of Boosted Trees[3].

Out of sample accuracy is the common criterion for forecasting, but it can be insufficient for evaluating the impact of decisions. As an extreme case, when machine learning models are used as a black box they may predict well without using the decision variables by replacing them with other variables that they are correlated with. Variables that affect both sales and causal variables (confounders) can result in models that are misleading when the promotion decisions follow a different pattern than in the training data (Ebbes et al., 2011). An external validation of whether the estimated impact of the causal variables "makes sense" is not easy when the model is not transparent. Retail forecasting is characterized by sparse and non-repetitive events[4], which makes it more difficult for flexile machine learning algorithms to learn the underlying patterns (Feng & Shanthikumar, 2018). Even if there is sufficient data, interpretability is particularly important to gain acceptance for system-generated forecasts from category managers, promotion planners and store managers who make operational decisions based on the forecasts, as well as from IT and business functions that provide the budget to implement the forecasting system. Yelland et al. (2019) report that users of forecasts at the retail chain Target were prepared to forgo some forecast accuracy to gain a more interpretable forecast. An example of the value of interpretability was that it allowed the supply-chain planners to make adjustments to the automatically generated forecast if it did not account adequately for an idiosyncratic influence on demand in a particular area/item.

Several methods use regularization with interpretable models, to deal with the high dimensionality problem in retail forecasting with promotions (e.g., Gür Ali & Yaman, 2013; Huang et al., 2014; Ma et al., 2016). Regularization is effective for selecting important variables, but it introduces bias to the parameter estimates. In general, due to the bias-

[3] https://www.kaggle.com/c/rossmann-store-sales

[4] Frequent introduction of new products and new types of promotions, exacerbated by the sensitivity of the demand on seasonality and environmental conditions like competition, fashion or economic conditions are some of the sources of non-repetitive events.

variance tradeoff, a model with biased but lower variance parameter estimates may yield more accurate predictions than a model with unbiased but higher variance estimates (e.g., Hastie et al., 2009; Kolassa et al., 2016). However, confounding variables pose an additional threat. While including the confounding variables along with the treatment variable solves the confounding problem in a linear ordinary least squares (OLS) model (Verbeek, 2008), the biased estimates in a regularized regression create the "regularization induced confounding" problem (Hahn et al., 2018). The implications for our problem are a) misleading estimates of price and promotion effects and b) lower forecast accuracy.

Another relevant issue is that a model that reflects the impact of causal variables properly does not necessarily forecast sales more accurately than a naïve time series model (Chevillon & Hendry, 2005). Random disturbances due to factors that are not captured by the model, such as changes in the (local) economy, competition, or consumer tastes, affect category-store base sales. Ignoring them reduces forecasting accuracy. In the literature, efforts have focused on either estimating the impact of promotions (e.g., Cooper et al., 1999; Van Heerde et al., 2002) while eliminating "nuisance" random disturbances due to other factors (Duncan et al., 2001); or extrapolating time series based on historical patterns and recent random disturbances using, for example, the exponential smoothing method (e.g., Dekker et al., 2004; Taylor, 2007). The latter captures the effect of "usual" marketing actions but misses those that deviate from their established patterns.

Our first contribution in this paper is a new method to support the retail planning process with accurate category-store level sales forecasts, as well as causal insights. The proposed FAIR (Fully Automatic Interpretable Retail forecasting) method considers category-store specific seasonality, focal- and cross-category marketing effects and their lags, and random disturbances to base sales. The medium-term multi-step ahead forecasts allow marketing scenario evaluation.

Compared to flexible machine learning methods in retail forecasting with promotions eg., Boosted Trees or Regression Trees (Gür Ali et al., 2009), FAIR has a domain knowledge driven data model that decomposes sales into what is expected based on category-store specific seasonal patterns, the impact of deviation from own- and cross-category marketing patterns, and the impact of random disturbances on base sales. We would ex-

pect similar or better accuracy than the machine learning models if the model captures the essential components of the data generation mechanism. Additionally, FAIR provides insights that are useful for the planning process, unlike black box models.

Compared to other methods with structured models that use regularization to deal with high dimensionality (e.g., Ma et al., 2016; Gür Ali & Yaman, 2013; Huang et al., 2014) FAIR first orthogonalizes sales and causal variables with respect to the common confounding source of category-store specific seasonality, to protect against regularization induced confounding. The resulting marketing coefficients can be interpreted as the elasticity of sales to marketing variables.

Since the sparseness and infrequent nature of promotional events is a challenge, FAIR pools observations across stores, which has been shown to increase forecasting accuracy with many methods (Gür Ali et al., 2009).

Unlike other methods that use Lasso regression, FAIR allows selecting variables or shrinking their coefficients as needed in an elastic net regression where the parameters are determined based on cross-validation. There is some evidence that using multiple "redundant" variables together may provide better predictive power than selecting one, especially in noisy environments (Guyon & Elisseeff, 2003). FAIR also uses domain knowledge to summarize a large number of SKU level marketing variables with meaningful features.

Finally, FAIR extrapolates the residual random disturbance components that affect the category-store base sales, building on the findings of Gür Ali & Pınar (2016) that extrapolating residuals resulted in better accuracy than an equivalent regression model with autoregressive error terms.

We apply the method and benchmarks to an extensive grocery dataset on three different grocery chains with 151 total stores. One to thirteen week ahead sales forecasts in 31 categories are evaluated. To ensure robustness, six non-overlapping test time periods are used.

The empirical evaluation shows that FAIR has better accuracy than benchmark methods. The focal and cross-category marketing effects, as well as random disturbance to base sales all contribute significantly and substantially to forecasting accuracy, thus justifying the high dimensional inputs and components of the FAIR method. The insights

make sense: The estimated coefficients of the FAIR model show post-promotion dip, carryover effects, and higher cross-category interaction among complementary categories without constrained estimation. These observations are in line with the theory developed in the marketing literature and provide face-validity.

Our second contribution is to empirically show that regularization induced confounding has significant and substantial impact on the accuracy of category-store level forecasts. This finding has implications for other retail forecasting methods beyond the proposed FAIR method.

Our third contribution is to provide an example that to gain interpretability we do not necessarily have to give up accuracy in retail forecasting, contrary to findings of numerous papers that show superior accuracy with machine learning models compared to interpretable models (see Fildes et al. (2020) for a review). In fact, FAIR has significantly better accuracy compared to Boosted Trees using the same features for every chain, origin and lead time horizon. While Boosted Trees are very flexible, they lack the guidance to navigate the numerous effects, functional forms, interactions and correlations of category-store-specific seasonality, marketing effects and random disturbances with the available data.

The rest of the paper is structured as follows. Section 2 provides an overview of the literature on retail forecasting methods in the presence of promotions, and cross category interactions. Section 3 introduces the FAIR method. The empirical evaluation is described in Section 4, including the dataset and derived features, accuracy comparisons, and interpretation of the estimated model coefficients for insights. Section 5 concludes the paper.

2 Retail forecasting with promotions

There are three main modes of incorporating causal (marketing) effects in retail forecasts. The first is judgmental adjustment of time series forecasts for promotional outcomes. The adjustments done by experts are costly, sometimes infeasible, and not very accurate due to biases (Fildes et al., 2009; Trapero et al., 2013). Apart from accuracy and scalability issues, expert adjustments are likely to be questionable when objectivity and consistency

are desired. This is particularly relevant during the planning process in the absence of complete alignment between the business functions.

The second mode involves the prediction of sales with an unstructured model that takes historical sales and causal variables with their lags as input. Machine learning techniques, such as Support Vector Machines (SVM), Neural Nets, Regression Trees, and their boosted, bagged versions do not assume a functional form for the effect of the causal factors on sales. These computing-intensive methods are heavily used for forecasting tasks in data mining competitions, like Kaggle (see Taieb & Hyndman (2014) for an example). They can learn arbitrary functions and interactions of many variables, identify those that are relevant and provide high predictive accuracy when large amounts of observations regarding repetitive events are present.

Unstructured forecasting models such as Boosted Trees or Regression Trees enjoy the advantage that they are very flexible and can estimate heterogeneous nonlinear relationships and arbitrary interaction effects and select and combine variables automatically. Gür Ali et al. (2009) show that when promotions and high dimensionality are present, regression trees improve SKU forecast accuracy compared to Support Vector Regression with Radial Basis Function (SRV-RBF) and linear models. Ferreira et al. (2015) estimate their demand prediction model for an apparel retailer by using regression trees with bagging, which they show to outperform parametric regression models. Cui et al. (2018) employ Gradient Boosting, Random Forest, and SVM for a daily aggregate sales forecast of an online apparel retailer by including social media information and observe that they perform better than linear models. Outside the retail scope, there are works that aim to improve supply chain efficiency by accurately forecasting demand with neural networks and/or SVM (Carbonneau et al., 2008; Efendigil et al., 2009).

Nevertheless, the large number of variables, many functional forms and sparse and non-repetitive events in retail forecasting pose a challenge (Feng & Shanthikumar, 2018), resulting in models that overfit or use spurious relationships. Such confounding would result in biased predictions when the retailer wants to evaluate a promotional scheme that is different from established patterns. Lack of interpretability, which is particularly acute in the case of ensembles such as Boosted Trees, Random Forests or deep neural network

models makes detecting the confounding problem very difficult.

Recognizing these issues, the machine learning community is developing methods to explain the trained black box models. Other scholars are arguing for working within the set of interpretable models and refuting the perception that there is a tradeoff between interpretability and accuracy. They point out that when meaningful features are constructed using domain knowledge, there tends to be little difference between the accuracy of the algorithms (Rudin, 2018). Interpretability serves first to establish trust by conforming to the established norms and validating the expected patterns. Further, providing new and useful insights as a by-product of the forecast ensures engagement in the process.

Several researchers introduce domain knowledge into unstructured models through features. Aburto & Weber (2007) forecast SKU sales at a Chilean supermarket with a neural network that takes residuals of a SARIMA model, as well as the promotional inputs and special day dummies as input. Gür Ali et al. (2009) construct dynamic features of SKU sales and promotions, such as promotion stock, based on findings in the marketing literature. Similarly, Ferreira et al. (2015) use a relative price feature that becomes instrumental for the subsequent price optimization. Sun et al. (2008) forecast fashion retail demand considering complex factors, such as design.

The third mode of incorporating causal (marketing) effects in retail forecasts is to constrain the model with a functional form to specify how promotions impact sales. This type of approach has been used in PromoCast (Cooper et al., 1999), SCAN*PRO (Van Heerde et al., 2002, 2001), CHAN4CAST (Divakar et al., 2005), Driver Moderator (Gür Ali, 2013), and in (Gür Ali & Pınar, 2016; Huang et al., 2014; Ma et al., 2016). By specifying the assumed data generation mechanism, these methods allow measurement of effects and enable insight generation. Their accuracy depends on how closely the assumed data generation mechanism resembles the actual data that is to be forecasted. The promotion event forecasting system PromoCast used regression with promotional variables to forecast SKU sales, while its extension identified rules with additional variables in the pooled residuals (Cooper & Giuffrida, 2000). The SCAN*PRO model focuses on the impact of promotions by decomposing the store-brand unit sales into multiplicative own- and cross-brand effects of price, feature and display, store and week effects. This model

has been used commercially and as basis for a number of marketing papers with variations on heterogeneity and estimation procedure (Andrews et al., 2008). On the other hand, it does not incorporate seasonality or extrapolate the effect of recent random disturbances that can make a significant difference in the resulting category-store specific sales volume, which is important for operational decisions such as compensation goal setting or staffing. CHAN4CAST is a decision support system providing sales forecasts of consumer-packaged goods by pack size, category, channel, region and customer account. The tool forecasts sales and simulates forecast scenarios based on possible marketing decisions and other variables. It has been developed with a substantial amount of analyst resources, testing many models for accuracy, validity of parameters and interpretability. They report using a log-linear random coefficient model where temperature, price, seasonality coefficients are allowed to vary across regions (Divakar et al., 2005).

Several models with interpretable functional form use regularization or dimensionality reduction to deal with the dimensionality problem. Gür Ali & Yaman (2013) use L1 norm regularization of dual and primal variables in an Epsilon Support Vector Regression (ϵ-SVR) to select the important observations and features. The Driver-Moderator method uses an L1 norm regularized epsilon insensitive regression (Gür Ali, 2013). Huang et al. (2014) use factor analysis before fitting an Autoregressive Distributed Lag (ADL) model using own and competing SKU marketing variables to forecast SKU sales. Ma et al. (2016) develop a three-stage lasso regression model where the first stage uses seasonality and promotions of the focal SKU and lags along with past sales, the second stage fits a model to the residuals with sales and promotions of other SKUs in the same category, and the third stage fits the residuals with data of selected SKUs from other categories. They include only the important SKUs or principal components, and restrict interactions among categories to those identified using the Lasso-Granger algorithm by Arnold et al. (2007). Other studies use domain knowledge driven features to replace numerous SKU level promotion variables in the category, eg., with a weighted average of discount, display and feature (Natter et al., 2007).

Using regularization is an effective method for selecting important variables but it introduces bias to the parameter estimates (Hastie et al., 2009). In a linear ordinary least

squares (OLS) model, including the confounding variables along with the treatment variable to predict y solves the confounding problem (Verbeek, 2008). However, the use of regularization while controlling for confounding variables is subject to regularization induced confounding (Hahn et al., 2018).

Causal models often predict less accurately than naïve formulations and they disregard the effect of random disturbances on base sales (Chevillon & Hendry, 2005). In the Two Stage Information Sharing method for category-store level forecasting method, Gür Ali & Pınar (2016) extrapolate the residual series that remain after taking into account the seasonal and causal effects. The empirical evaluation indicates that there are substantial accuracy gains by extrapolating the residual time series beyond what is attainable with a comparable regression model with autoregressive error components, or an ADL model that includes lags of sales and marketing variables in a single model. Vector Auto Regressive (VAR) approaches consider temporal correlation within and across the series. Curry et al. (1995) treat sales and marketing variables of competing entities as endogenous variables. Stock & Watson (2003) differentiates between three variations of VAR models: (i) reduced form, (ii) recursive, and (iii) structural. Only the reduced form VAR is applicable for prediction purposes because the other two require the current values of dependent variables to predict some other dependent variable.

High dimensionality complicates the estimation of VAR models (Harvey, 1990), where the number of free parameters increases quadratically with the number of variables (Ma et al., 2016). Hsu et al. (2008) suggest using the lasso technique to reduce dimensionality and generate a more robust estimation. Gelper et al. (2016) uses a similar approach, namely Sparse VAR; and they represent the VAR model in a regression form while penalizing the L1 norm of the coefficients and non-diagonal elements of the covariance matrix of contemporaneous error terms. By doing so, they can prevent the model from being overparameterized. Alternatively, one can reduce the dimensionality by assuming some variables are exogenous (VARx), if the exogeneity of the variables can be theoretically supported. In addition to that, Bandyopadhyay (2009) uses Bayesian VAR (BVAR) that has fewer degrees of freedom. BVAR is an estimation of a VAR model by assuming a prior distribution of the parameters which makes the coefficients for longer

lags more likely to have zero means. Carriero et al. (2012) also use BVAR to overcome the over-parameterization they encounter while they forecast correlated government bond yields.

There is substantial work in the marketing literature that shows change in one category's sales affect other categories through substitution, income, or complement effects (Leeflang & Parreno-Selva, 2012). Considering cross-category effects improves forecasting accuracy (Ma et al., 2016). Gelper et al. (2016) show that the cross-category effect exists beyond the categories that are traditionally assumed to be related, like substitutes and compliments with a sparse vector autoregression (VAR) model. We create two sets of features summarizing SKU level marketing variables in the category, providing the level and distribution of marketing resources in each category.

For an extensive review of retail forecasting and many of these models, we refer the reader to Fildes et al. (2020). They report retail promotion optimization solutions (e.g., Ma & Fildes, 2017), SAP Customer Activity repository) do not have widespread use among retailers, and that these decisions are made by the demand planning team who like to interpret and interact with the statistical models.

3 FAIR (Fully Automatic Interpretable Retail forecasting)

FAIR aims to provide an objective platform where different functions in a retailer can collaborate to assess promotion impact or shape demand by evaluating scenarios of planning variables. This requires considering causal effects, seasonality and random sustained disturbances that come together to form the resulting sales; as well as interpretability of sales drivers.

While the general method can be extended to other planning (causal) variables, in this work we consider the marketing variables, including those for the focal category and other categories. We create two sets of features summarizing SKU level marketing variables, providing the level and distribution of marketing resources in each category, respectively.

We consider the following major sources of confounding that affect both marketing and sales: a) time-invariant unobserved category-store characteristics such as clientele of

store, b) seasonal unobserved category-store characteristics such as weather or pay-day patterns, c) unobserved factors varying in time such as price inflation or growth in client interest in the category-store. Corresponding to each of these sources, we have the following observed confounding variables a) store indicator, b) category-store specific seasonal variables, c) category-store specific linear trend t. We do not claim that the formulation eliminates all confounding, but as we show empirically, addressing these sources of confounding has a substantial impact on forecast accuracy as a result of removing regularization induced confounding bias from the estimates of the marketing variable effects.

3.1 The data model

Category-store time series from multiple stores of a given chain constitute multivariate (including sales and marketing variables), unbalanced panel data with occasional missing values and shorter series. There is cross-sectional variation due to multiple stores within the retail chain and variation across time.

The FAIR data model specifies the causal planning variables (marketing in this case) and sales as a function of their common determinants, i.e., the confounding variables. The status quo in sales is described as

$$\begin{aligned} &\ln(y_{ijt}) = b_{ij} + \beta_{1,ij} H_t + \beta_{2,ij} t + z_{ijt} \\ &i = 1...C,\ j = 1...S \end{aligned} \tag{1}$$

Here, y_{ijt}[5] is the sales amount in category *i,* store *j,* and time *t*; H_t is the vector of seasonality variables; b_{ij} is the category-store fixed effect; z_{ijt} is the "unexpected" sales component. *C* is the number of categories, *S* is the number of stores. Although in this example, the confounding variables are restricted to the seasonality variables, depending on the specific application, other confounding variables can be added to this expression. Similarly, the status quo in the marketing variables is described as

$$\begin{aligned} &\ln(m_{ijat}) = c_{ija} + \gamma_{1,ija} H_t + \gamma_{2,ija} t + w_{ijat} \\ &i = 1...C,\ j = 1...S,\ a = 1...A_1 + A_2 \end{aligned} \tag{2}$$

where m_{ijat} is the ath marketing feature for category *i*, store *j* and time *t*; c_{ija} is the category-store fixed effect for the marketing feature; w_{ijat} is the "unexpected" marketing

[5] 1 is added to sales and marketing variables to avoid undefined value in case of 0.

component. The marketing features which are assumed to have cross-category effects are indexed from 1 to A_1 , whereas the ones that have only focal-category effects are indexed from $A_1 + 1$ to $A_1 + A_2$. Next, the unexpected sales is explained with unexpected focal- and cross-marketing activity.

$$\begin{aligned} & z_{ijt} = \sum_{a=1}^{A_1+A_2} \sum_{l=0}^{L} \varphi_{ial} w_{ija,t-l} + \sum_{k \neq i} \sum_{a=1}^{A_1} \sum_{l=0}^{L} \theta_{ikal}\ w_{kja,t-l} + r_{ijt} \\ & i = 1, ..., C \end{aligned} \tag{3}$$

L is the maximum lag of marketing features to be considered. The final residuals, r_{ijt}, are assumed to be correlated over time.

$$cov(r_{ijt}, r_{ijt+l}) \neq 0 \qquad i = 1, ..., C;\ j = 1, ..., S \tag{4}$$

Inserting (2) and (3) into (1) provides the log-linear data model

$$\begin{aligned} & \ln(y_{ijt}) = b_{ij} + \beta_{1,ij} H_t + \beta_{2,ij} t \\ & + \sum_{a=1}^{A_1+A_2} \sum_{l=0}^{L} \varphi_{ial} (\ln(m_{ija,t-l}) - c_{ija} - \gamma_{1,ija} H_{t-l} - \gamma_{2,ija}(t-l)) \\ & + \sum_{k \neq i} \sum_{a=1}^{A_1} \sum_{l=0}^{L} \theta_{ikal} (\ln(m_{kja,t-l}) - c_{kja} - \gamma_{1,kja} H_{t-l} - \gamma_{2,kja}(t-l)) \\ & + r_{ijt} \end{aligned} \tag{5}$$

where sales is a function of seasonality, trend, focal and cross-category marketing variables and their lags. We explain the motivation for this representation in the next section. The φ_{ial} coefficient captures the marketing impact that the lag *l* of the marketing instrument *a* in category *i* has on the sales in the same category-store. Similarly, the θ_{ikal} coefficient captures the cross-category marketing impact that the lag *l* of the marketing instrument *a* in category *k* has on the sales in category *i* in the same store. Due to the *log – log* structure of the models, they represent the elasticity of sales to marketing instruments. The residual terms, r_{ijt}, represent the random component that also contains changes to the category-store base sales over time, with no distributional assumptions.

Notice that while (1) and (2) imply category-store specific seasonality and trend patterns for sales and marketing variables, the multiplicative impact of marketing that deviates from these patterns is assumed to be common across stores. This is in line with findings that store-level heterogeneity does not improve marketing mix elasticity estimates (Andrews et al., 2008), and provides greater sample size to estimate the effects.

3.2 Estimation with a multistage approach

FAIR is designed to address the complicating issues introduced earlier. High dimensionality is present even after using domain knowledge for specifying the model and summarizing SKU level variables. There are a large number of promotion variables due to cross-category interactions, distributional features and their lags relative to the number of observations with significant variation from the status quo. To address this issue, FAIR pools observations across stores to estimate category-level marketing elasticities and uses elastic net regularization to select variables and shrink parameters as needed (Hastie et al., 2009).

In an OLS regression the confounding issue can be solved by including the confounding variables along with the treatment variable to predict y (Verbeek, 2008). However, regularized regression may introduce bias to treatment (e.g., marketing) effect estimation (Hahn et al., 2018).On the other hand, regressing (3) with OLS provides the same parameter estimates as regressing (5), according to the Frisch–Waugh–Lovell theorem (Greene, 2012). So, we orthogonalize sales and marketing variables with respect to the confounding variables using OLS in (1) and (2), and use the residuals to estimate the marketing effects in a regularized regression in (3). A similar orthogonalization approach has been shown to eliminate the regularization induced confounding (Chernozhukov et al., 2018). In section 5.2, we empirically show that orthogonalization before causal effect estimation substantially improves the accuracy of the forecasts.

The third complicating issue of changes to the base category-store sales due to other random factors is addressed by extrapolating the category-store residual series. FAIR model estimation consists of three stages.

The first stage estimates the coefficients in Equations (1) and (2) with category-store specific OLS regressions: category-store specific seasonalities $\widehat{\beta}_{1,ij}$ and $\widehat{\gamma}_{1,ija}$, trends $\widehat{\beta}_{2,ij}$ and $\widehat{\gamma}_{2,ija}$, and the fixed effects $\widehat{b}_{ij}$ and $\widehat{c}_{ija}$. They are used to estimate the residuals, $\widehat{z}_{ijt}$ and $\widehat{w}_{ijat}$. The second stage estimates the focal category and cross-category marketing effects by fitting Equation (3) using $\widehat{z}_{ijt}$ and $\widehat{w}_{ijat}$ for each category with the elastic net penalty. There are $P = C \times L \times A_1 + L \times A_2$ marketing parameters to be estimated per category. To simplify the exposition, we call the vector containing all φ_{ial} and θ_{ikal}

coefficients for category *i*, ξ_i and we let W_i represent the matrix containing $w_{ija,t-l}$ and $w_{kja,t-l}$, and Z_i the vector of z_{ijt}. Elastic net penalized regression determines $\hat{\xi}$ by minimizing the following loss function that trades off model fit in terms of squared error with model complexity.

$$(Z_i - W_i\xi_i)^T(Z_i - W_i\xi_i) + \lambda(\alpha\xi_i^T\xi_i + (1-\alpha)\|\xi_i\|_1) \tag{6}$$

The λ hyperparameter determines the amount of regularization, larger values indicating simpler models. The hyperparameter $\alpha \in [0, 1]$ determines the weights of lasso and ridge penalty in the penalty term, $\alpha = 1$ being the lasso penalty (L1) and $\alpha = 0$ ridge penalty (L2). Elastic net penalty induces dimension reduction and continuous shrinkage at the same time (Zou & Hastie, 2005). By using a convex combination of lasso and ridge penalty elastic net inherits advantages of both methods. While the lasso penalty is very useful in eliminating irrelevant variables, the ridge penalty shrinks the coefficient values toward zero. In the presence of correlated relevant variables, lasso tends to select one and leave others out[6], while ridge would retain the variables but shrink their coefficients, which makes for a more robust estimator in a noisy environment with many inputs. To sum up, we use the elastic net penalty to eliminate irrelevant variables while keeping relevant correlated variables with shrunk coefficients.

The third stage extrapolates the $\widehat{r}_{ijt}$ residual series remaining after the Stage 2 estimation for each category *i*, and store *j* as of the last observed time period t_0. It provides $\widehat{r}_{ij,t_0+h}$. As mentioned in the previous section, we expect the residuals to be autocorrelated within a store, but to be contemporaneously independent across stores. We use the Seasonal and Trend decomposition using Loess (STL) method (Cleveland et al., 1990) that decomposes a time series into seasonal, trend-cycle, and remainder components which are respectively represented by S_t, T_t, and E_t in the following equation:

$$r_t = S_t + T_t + E_t$$

STL is a locally estimated method which allows seasonal and trend components to change over time. We use the *stlf* function in *forecast* package (Hyndman et al., 2019; Hyndman & Khandakar, 2008) in the R Language (R Core Team, 2017) to estimate the STL models

[6]For example (Ma et al., 2016) resorts to a multistage lasso regression to overcome this problem.

and predict the future values. The package determines the smoothing parameters of trend and seasonality components automatically.

The *h* step ahead forecasts of ln sales are denoted as f_{ijt}; they are produced with the estimated parameters and future m_{ijat} using Equation (5). Finally, as we take the anti-log we multiply by the ratio estimator $R_{ij} = \sum_{t=1}^{t_0} y_{ijt} / \sum_{t=1}^{t_0} exp(f_{ijt})$ to account for the log transformation bias (Snowdon, 1991).

3.3 Setting hyper-parameters with time series cross-validation

We use the *glmnet* package (Friedman et al., 2010) in R to implement the elastic net. Following the convention in regularization literature, we standardize the variables before elastic net estimation. The elastic net regularization method requires hyperparameters governing a) to what extent the regression model should be regularized, and b) to what extent it should rely on the ridge versus lasso penalty, namely the α and λ hyperparameters. α_i and λ_i refer to the Stage 2 model hyperparameters for category *i*. Cross-validation is the standard approach to set machine learning hyperparameters: the training data is split into v folds, the model is trained on v-1 folds and its accuracy is evaluated on the remaining fold. Repeating this procedure for each candidate hyper-parameter set and fold yields v realizations of the accuracy measure, which provides estimates for mean and standard deviation of accuracy for each candidate parameter set.

However, cross-validation has two important assumptions that the data are 1) independent and 2) identically distributed (Arlot & Celisse, 2010). Time series are temporally interdependent. While v-fold cross-validation can be used if the time series is stationary and the residuals of the model are uncorrelated, these assumptions could be violated in real-life applications. One remedy for this problem is blocked cross-validation, where the data is split sequentially instead of randomly, but the validation set may precede the training set in time (Bergmeir & Benítez, 2012).

Our cross-validation approach maintains the time order of the data chunks by only using those sets where the validation set follows immediately after the training set. Thus, by preserving the precedence relationship, it also addresses the potential issue that the data generation mechanism may evolve over time. Figure 1 illustrates the approach, where the training data is split into *v* sequential and overlapping blocks, where the first block con-

tains a minimum number of observations to train and validate the model, and each block start time is offset by n_v observations. In each block the last n_v observations constitute the validation set, and the rest the training set.

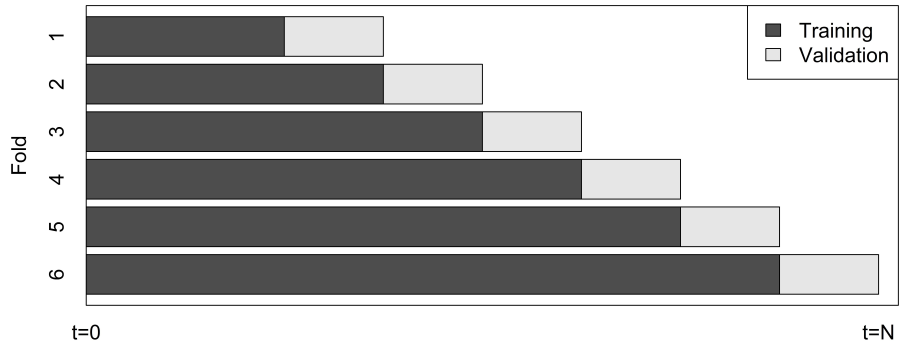


FIGURE 1: Cross validation procedure overview

This cross-validation scheme is used to evaluate the accuracy of the grid of candidate α and λ hyperparameter sets. The combination with the lowest mean(RMSE) + std(RMSE) is set as ${\alpha^*}_i$ and ${\lambda^*}_i$. The complete training data is used to train the final model with ${\alpha^*}_i$ and ${\lambda^*}_i$.

Category-store models can be run in parallel in Stage 1, while category models can be run in parallel in Stages 2 and 3. However, outputs of sister categories after Stage 1 become inputs for the next stage model.

4 Empirical evaluation

In this section, we first introduce the extensive grocery dataset for empirical evaluation, then introduce the benchmark methods. Next, we report the accuracy results quantifying the contribution of FAIR method components and compare FAIR with benchmark methods. Further, we evaluate the face-validity of the model estimates by discussing the estimated marketing effects including cross-category interactions.

4.1 Data and the forecasting task

The data (Bronnenberg et al., 2008), is provided by market research company IRI[7], one of the two largest retail data providers. It consists of store-level weekly scanner data for grocery stores operating in North America. In this study, we focus on the three largest chains[8] in terms of sales volume, during the period from the beginning of 2003 to the end of 2007. There are 42, 69 and 43 stores in each chain respectively, and all the 31 grocery categories included in the IRI dataset are available in all stores.

The forecasting task is to predict category-store level dollar sales for 1 to 13 weeks ahead, given the historical sales and marketing information, and the marketing plans for the forecasting horizon.

In order to ensure that the results are not particular to the characteristics of one chain or time period, we use six non-overlapping test time periods and three chains to produce forecasts. In other words, in our experiment we have 18 replications with 31 categories each. The first evaluation set spans from December 30, 2002 to October 1, 2006. The data until July 2, 2006 are used for training and validation. The following 13 weeks are set aside for testing. The other origins are obtained by rolling the dataset by 13 weeks each time. The cross-validation blocks within each set are determined as described in section 3.3. The resulting length of the training data varies from 183 to 248 weeks, providing 3.5 to 4.5 years of training data that is relevant for seasonality estimation.

The original data is provided at the SKU-store-week level and includes dollar and unit sales, marketing variables in category *i*, store *j*, SKU *s* and week *t* as follows. Dollars_{ijst}: dollar sales, Units_{ijst}: unit sales in units PR_{ijst}: price reduction flag, D_{ijst}: display feature, F_{ijst}: advertisement feature. In addition to these variables, we are provided with market, chain, estimated activity, and status (open/close during a week) information of the stores. We assume the sticker price (price without discount) to be the maximum observed price within the last three months.

We aggregate the data from the SKU-store-week level to the category-store-week level as follows. We sum the dollar sales based on the SKU-category hierarchy. For aggre-

[7]We would like to thank IRI for making the data available. All estimates and analysis in this paper, based on data provided by IRI, are by the authors and not by IRI.

[8]Large chains may have been partitioned into multiple de-identified chains by IRI.

gating marketing variables, we follow an approach similar to the one by Divakar et al. (2005), who aggregated the data geographically to the SKU-region level. We calculate the category-store level marketing variables as the weighted average of the SKU-store level marketing variable using the sales volume of the category SKUs in the last 13 weeks as weights. Disaggregate information in the aggregate model has been shown to improve the aggregate forecast accuracy for certain applications (Harvey & Cushing, 2014; Zeng, 2017). Thus, in addition to these point summary measures for each marketing instrument, we also create and use distributional variables, which are assumed to have only focal-category effects.

The vector of seasonality variables H_t includes indicator variables indicating whether the week of the observation includes (1) Thanksgiving, (2) Christmas, (3) other holidays (Labor Day, Memorial Day, New Year), (4) the first day of the month, (5) the fifteenth day of the month, as well as indicator variables for the month[9]. The week variable captures the overall trend.

The marketing features, m_{ijat} include four basic and eight distributional features for each category *i*, store *j* and week *t*, and their first and second lags; i.e. $A_1 = 4,\ A_2 = 8,\ L = 2$.

Basic features measure the level of each marketing instrument, as follows.

Regular_Price$_{ijt}$ is the volume weighted average unit price in the category

Display$_{ijt}$ is the percent of dollar weighted category volume on display

Discount$_{ijt}$ is the category discount percentage, and measures the overall resource allocation

Ad$_{ijt}$ is the percent of dollar weighted category volume advertised.

Distributional features capture the distribution of the discount resource. A 5% discount on the category can result from a 5% price discount on all category items, but it may also result from discounting 10% of half of the category volume. We define dist_d$_{ijt}$ as the percent of category volume discounted at rate range d to capture the distribution

[9] An alternative to indicator variables for the months is a trigonometric representation using fewer degrees of freedom for the parameters and providing a smooth transition between months. A sensitivity analysis in our case study with $\sin(j \times g \times t)$ and $\cos(j \times g \times t)$, where $g = \left(\frac{2\pi}{\#weeks\ in\ a\ year}\right)$ and $j = 1, 2, 3$ (Shumway and Stoffer, 2017) resulted in a slightly lower accuracy than using the monthly indicator variables that also have the advantage of being readily interpretable.

of the discount resource. We use eight rate ranges and define corresponding distributional variables, *dist_5, dist_10, dist_15, dist_20, dist_30, dist_40, dist_50* and *dist_100*. Here, *dist_5* is the percent of category volume discounted less than 5%, excluding non-discounted items, *dist_10* is the percent of category volume discounted at 5% to 10%, and so on.

Figure 2 provides a time series plot of sales and basic marketing features for a sample category-store in Chain 48 and the Beer category. We clearly observe seasonality in sales and marketing features. As seen in Figure 3, category sales volumes are very diverse, with Carbonated Beverages, Milk and Beer in the top three among the study categories. The relative size of categories is similar across the chains considered.

Chain 117 relies more on discounting than the other two chains, while Chain 14 uses more advertising, as seen in Table 1 which provides the average and standard deviation of the variables by chains within the overall data including train and test sets. Carbonated Beverages, Beer and Cigarettes categories stand out in terms of promotion levels.

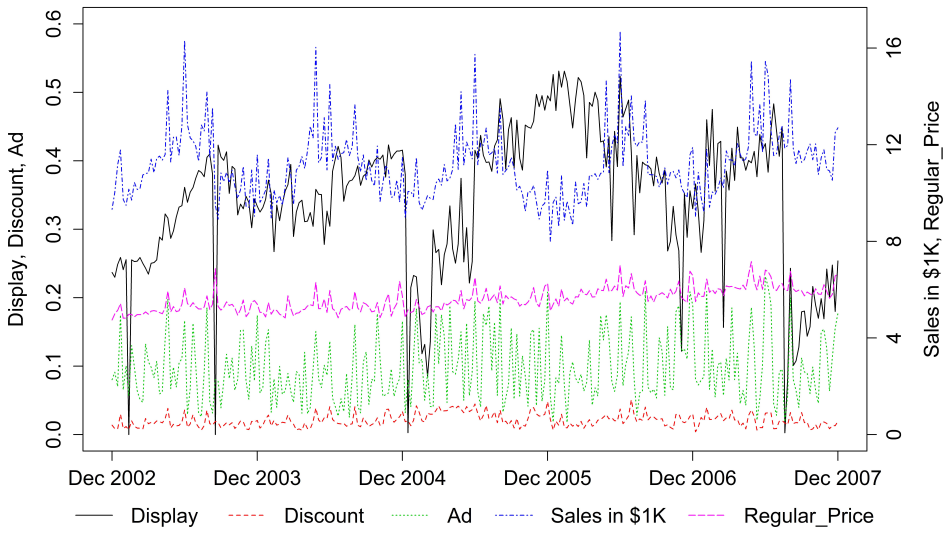


FIGURE 2: Sample category-store-level sales and basic marketing feature time series

In the dataset, some category-store-week observations are missing. The potential reasons are either that the store was closed or no transaction occurred. If the data is missing for all categories in the store for that week we assume that the store was closed. We interpolate the series to fill the missing values. If the missing values amount to more than 30%

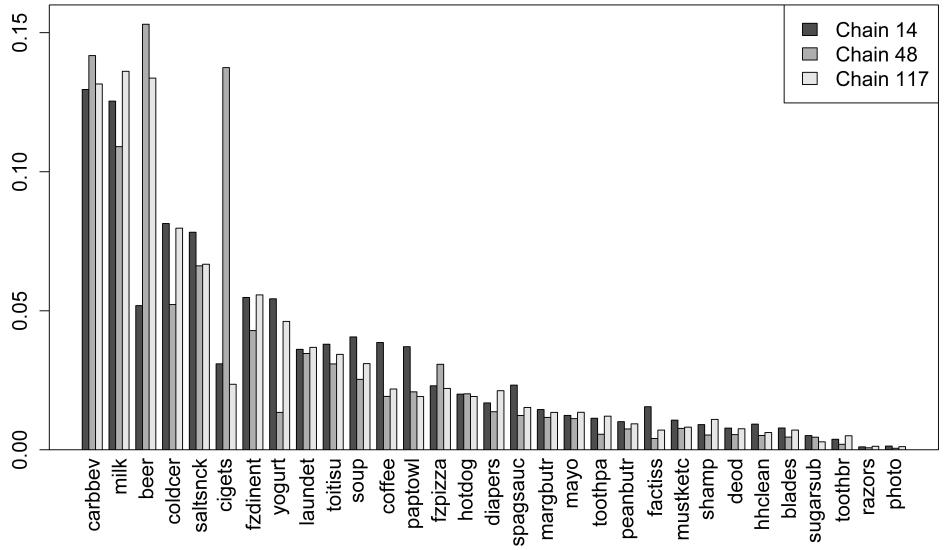


FIGURE 3: Relative sales volume of categories by chain

TABLE 1
Descriptive statistics by retail chain – all data

		Average			Standard Deviation		
		Chain14	Chain48	Chain117	Chain14	Chain48	Chain117
Basic Features	Regular Price	4.39	3.64	4.38	2.82	2.48	2.98
	Discount	0.05	0.05	0.07	0.05	0.04	0.05
	Display	0.06	0.08	0.07	0.11	0.13	0.13
	Ad	0.11	0.08	0.08	0.13	0.12	0.12
Distributional Features	dist_5	0.15	0.16	0.16	0.13	0.14	0.14
	dist_10	0.03	0.04	0.07	0.06	0.07	0.09
	dist_15	0.03	0.04	0.06	0.06	0.06	0.08
	dist_20	0.02	0.03	0.05	0.04	0.06	0.07
	dist_30	0.04	0.05	0.07	0.06	0.08	0.09
	dist_40	0.04	0.03	0.05	0.07	0.06	0.07
	dist_50	0.03	0.01	0.03	0.06	0.04	0.06
	dist_100	0.01	0.01	0.01	0.04	0.03	0.04
	Sales	2,225	1,692	2,823	2,748	2,619	3,882
	N	322,573	534,301	332,523			
	# of stores	42	69	43			

of the category-store training data, we exclude the entire category-store from the analysis for methods that cannot handle missing data. If the data is missing for some categories in the same week, we assume that no transaction occurred and the dollar sales amount is assumed to be zero.

4.2 Benchmark methods

We compare the performance of the FAIR method with a number of methods. Table 2 provides an overview of the benchmark methods. ETS, Auto-Arima and STL all forecast using only the time series characteristics of the focal series. VAR considers the interaction among the series, and VARx additionally considers the impact of external causal variables. Boosted Trees considers all relevant information. Being a very flexible, unstructured black box model, it does not offer interpretation of the fitted model. Therefore, the evaluation of the estimated relationships for face-validity is not possible. Ensembles of Boosted Trees have been used by the winning solution of a Kaggle competition for forecasting daily drugstore sales with promotions and other data[10]. We benchmark ETS, Auto-Arima, STL and VAR for comparing their forecasting accuracy, although they do not meet the requirement of taking causal variables into account.

TABLE 2
Information taken into account by each forecasting method

	ETS	Auto-Arima	VAR	VARx	STL	FAIR	Boosted Trees
Causal Variables	N	N	N	Y	N	Y	Y
Cross-category Interactions	N	N	Y	Y	N	Y	Y
Time series effects	Y	Y	Y	Y	Y	Y	Y
Interpretability	N	N	N	Y	N	Y	N
Scales to 31 categories	Y	Y	Y	N	Y	Y	Y

Boosted Trees

Classification and regression trees are one of the fundamental tools in machine learning. Their decision tree-like structure makes them easy to understand for non-experts. Regression trees split the predictor space into simple regions with bisection and determine a simple prediction, like the average, for all the observations in a region (Hastie et al., 2009). Boosting additively combines many trees, that are grown successively giving higher weight to observations with high error in the previous stages. The assumed equation for boosting is

$$f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m)$$

[10] https://www.kaggle.com/c/rossmann-store-sales/discussion/18024

where each $T(x; \Theta_m)$ is a tree fitted to the data. Gradient boosting (GB) for regression trees is a "greedy function approximation" to Boosted Trees (Friedman, 2001) which does not reweight the sample but adjusts the output at each step. If the loss function is squared error, the "gradient" is the residual from the previous step, i.e., each tree predicts the residual of the previous one. In the context of numerical optimization, GB is an analogue to gradient descent methods. As one can control the step size in gradient descent methods, GB allows controlling the learning rate. In GB algorithms, learning rate and the number of trees are two essential parameters that regularize the model. The maximum depth of the trees is another parameter that determines the complexity of the model by adjusting the maximum level of the interaction among the features. Friedman (2002) introduced a stochastic element into the method, by taking a random sample without replacement from the observations to grow each tree. Hence, the algorithm is called stochastic gradient boosting (SGB). This randomization results in lower correlation among the trees, which usually leads to improved accuracy.

Boosted Trees is attractive because of its several properties: (i) it can handle less than clean data, (ii) non-linear relations are easily captured, and (iii) it automatically selects features. Cui et al. (2018) utilize both Random Forest and Boosted Trees for a daily aggregate sales forecast for an online apparel retailer. Guelman (2012) uses the technique to model and predict the loss cost of an auto insurance firm and observes that GB outperforms Generalized Linear Models, which is the established approach to the problem. Lemmens & Croux (2006) show that the Boosted Trees increase the accuracy in predicting churn for a telecommunication company.

With a complex relationship between the outcome and the predictors, high dimensionality and missing values, our forecasting problem is expected to benefit from Boosted Trees. We provide the same inputs to Boosted Trees model as the FAIR method with the addition of category-store sales average: 1) seasonality variables, 2) basic promotional variables, 3) distributional variables, 4) lags of sales, 5) lags of basic promotional variables, and cross-category counterparts of 2, 4, 5; up to 26 lags. We train chain-category-leadtime specific models as in the FAIR method. The total number of features exceeds 4000.

To train the models, we use the XGBoost R package (Chen and Guestrin, 2016) which implements the SGB algorithm efficiently. To determine the hyperparameters of the algorithm, we utilize the same CV scheme as in the FAIR method. Following the range of parameters recommended by (Hastie et al., 2009), we use the following grid search space: Learning rate $\in \{0.1, 0.3\}$, Number of trees $\in \{100, 250\}$, Subsample rate $\in \{0.5, 1\}$. We also require the algorithm to have nodes with at least 50 observations, have a maximum depth of 6 and to stop if the performance does not improve on a validation set for 10 iterations.

Other benchmarks

STL method is discussed in Section 3.2.

ETS is a popular automatic forecasting tool that uses exponential smoothing method in a state-space framework, taking seasonality, trend and random disturbances into account (Hyndman et al., 2002). We implement it with the *ets* function in R package *forecast* (Hyndman et al., 2019; Hyndman & Khandakar, 2008).

Autoregressive integrated moving average ARIMA(p, d, q) model use the univariate time series for forecasting

$$y'_t = c + \theta_1 y'_{t-1} + \cdots + \theta_p y'_{t-p} + \varphi_1 e_{t-1} + \cdots + \varphi_q e_{t-q} + e_t$$

where e_{t-i} is the white noise in period *t-i*, y'_t is obtained by first differencing y_t d times, p and q are orders of autoregressive and moving average parts respectively (Hyndman & Athanasopoulos, 2014). We use *auto.arima* function in the R package *forecast* to select the parameters p, d, and q; and to estimate the model. Auto-Arima can model seasonality through AR terms that correspond to the seasonal frequencies.

The vector autoregression (VAR) model is an extension of the AR model that assumes not just temporal correlation within a time series but also correlation among the series. Verbeek (2008) formulates the reduced form VAR(p) that is useful for forecasting sales given marketing scenarios as

$$Y_t = \delta + \Theta_1 Y_{t-1} + \cdots + \Theta_p Y_{t-p} + \varepsilon_t$$

where Y_t is k-dimensional vector, each Θ_j is a $k \times k$ matrix, and ε_t is a k-dimensional vector of error terms with covariance matrix Σ.

We implement a VAR model for each store using its category sales time series using the R package *vars* (Pfaff, 2008). Before providing the data to the packages, we deseasonalized and detrended the data, using Equation (1), simplifying the task for VAR. We also attempted to estimate a VARx model that explains the endogenous sales with both the sales and exogenous marketing variables, again deseasonalizing and detrending the data. However, we were not able to estimate this model due to high dimensionality, although we used only a limited set of marketing features (basic variables introduced in Section 4.1). The method we used to estimate the model, *BigVAR* (Nicholson et al., 2017) estimated the model for one store and time origin in 5 days and 19 hours[11], and did not provide forecasts for leadtimes greater than 1. Thus, we excluded VARx from the analysis.

All benchmark model forecasts are adjusted for the log transformation bias in the same manner as the FAIR forecasts.

4.3 Accuracy Evaluation

In this section, we quantify the contribution of FAIR model components to forecasting accuracy and compare its forecasting accuracy to external benchmark methods.

FAIR model components

To quantify the accuracy contribution of each model component in FAIR, we train and test five additional versions of FAIR, as shown in Table 3. Model S involves forecasting using only Equation (1) which forecasts Base Sales with category-store specific constant Seasonality and Trend. Model Sm involves Stages 1 and 2, using only focal category marketing features. Model SM also includes Stages 1 and 2, but uses cross-category marketing features in addition to the focal category marketing features. FAIR, in addition includes Stage 3 that extrapolates random disturbances to base sales, as described in Section 3[12].

The FAIR No Orthogonalization model replaces Stages 1 and 2 with a model that

[11] The algorithm was run on a system with CPU: Intel(R) Xeon(R) CPU E5-2695 v4 @ 2.1GHz, Memory: 16 GB

[12] FAIR handles observations with missing values by defaulting to the forecast provided by the next simpler model in the complexity order of models, which is as follows: "S, Sm, SM, FAIR". Notice that the simplest model S will always provide a forecast.

TABLE 3
Internal accuracy evaluation design

Models	Components of FAIR					
	Base Sales Constant Seasonality and Trend	Focal Marketing	Cross-category Marketing	Random disturbances to base sales	Distributional marketing features	Orthogonalization wrt confounders
S	+					
Sm	+	+			+	+
SM	+	+	+		+	+
FAIR	+	+	+	+	+	+
FAIR No Orthogonalization	+	+	+	+	+	

estimates

$$\ln(y_{ijt}) = b'_{ij} + \beta'_{1,ij} H_t + \beta'_{2,ij} t + \sum_{a=1}^{A_1+A_2} \sum_{l=0}^{L} \varphi'_{ial} (\ln(m_{ija,t-l}) + \sum_{k \neq i} \sum_{a=1}^{A_1} \sum_{l=0}^{L} \theta'_{ikal} (\ln(m_{kja,t-l}) + r'_{ijt} \quad (7)$$

with the elastic net. Subsequent Stage 3 proceeds in the same manner in both versions with the respective r'_{ijt}. In section 3.2 we explained that Stage 1 was introduced to avoid regularization induced confounding. The comparison of accuracy between FAIR and the No Orthogonalization version quantifies the accuracy contribution of orthogonalization with respect to confounding variables.

Accuracy results

Table 4 provides the forecasting accuracy and bias of the five model versions and external benchmarks by lead time buckets. The forecast accuracy measures are as follows: Root Mean Squared Error (RMSE); Weighted Mean Absolute Percentage Error (wMAPE) which is the ratio of the sum of the absolute errors to the sum of the actual values; and the Average Relative Mean Absolute Error (AvgRelMAE) (Davydenko and Fildes, 2013), which is the geometric mean of the MAE ratios relative to the naïve forecast – defined as the sales last year the same week. The bias of the forecast is measured in weighted Mean Percentage Error (wMPE). The results are obtained on the test dataset which has 340,950 category-store-week observations, with average sales of $2,289.

The results indicate that FAIR provides substantially better forecasts than all external benchmarks for all forecasting horizons, and in all chains and origins[13]. Overall, compared to Boosted Trees, FAIR has 18%, 12%, and 4% lower RMSE, wMAPE, and AvgRelMAE, respectively. AutoARIMA, ETS and VAR perform much worse than FAIR and Boosted Trees. As expected, FAIR forecasting accuracy is better for shorter lead times, but the deterioration is mild: from 10.7% in the 1-4 weeks horizon to 11.4% in the 10-13 weeks horizon.

Among the external benchmarks, STL performs the best, although its accuracy is still 8% to 9% worse than FAIR in terms of all the accuracy measures. While this performance in capturing evolving sales patterns is remarkable, STL will not be useful for evaluating new marketing scenarios. Unlike the other time series benchmarks, the VAR model con-

[13]The tables are provided in the Appendix

TABLE 4
Accuracy and Bias of the benchmark methods and FAIR versions

	RMSE				wMAPE			
Models	1-4 Weeks	5-9 Weeks	10-13 Weeks	Overall	1-4 Weeks	5-9 Weeks	10-13 Weeks	Overall
S	748	716	703	722	14.8%	14.6%	14.8%	14.7%
Sm	659	622	626	635	12.9%	12.6%	13.0%	12.8%
SM	623	589	592	601	12.3%	12.1%	12.5%	12.3%
FAIR	**553**	**539**	**550**	**547**	**10.7%**	**10.9%**	**11.4%**	**11.0%**
FAIR No Orthogonalization	620	620	635	625	12.3%	12.5%	13.0%	12.6%
Boosted Trees	630	658	641	644	11.5%	12.5%	12.9%	12.3%
STL	594	588	596	592	11.6%	12.0%	12.4%	12.0%
AutoArima	627	678	671	660	12.0%	13.3%	14.0%	13.1%
ETS	675	786	869	781	12.8%	15.1%	17.2%	15.1%
VAR	693	675	696	687	13.4%	13.6%	14.5%	13.8%

	AvgRelMAE				wMPE			
Models	1-4 Weeks	5-9 Weeks	10-13 Weeks	Overall	1-4 Weeks	5-9 Weeks	10-13 Weeks	Overall
S	84.2%	86.0%	88.8%	86.2%	5.7%	4.3%	3.7%	4.6%
Sm	74.8%	76.4%	79.7%	76.8%	4.8%	3.4%	3.1%	3.8%
SM	73.4%	75.0%	78.1%	75.4%	3.7%	2.4%	1.9%	2.6%
FAIR	**67.7%**	**70.9%**	**74.0%**	**70.8%**	2.6%	1.6%	1.2%	1.8%
FAIR No Orthogonalization	74.9%	78.6%	81.6%	78.3%	2.9%	1.8%	1.3%	2.0%
Boosted Trees	68.0%	74.5%	77.1%	73.3%	3.5%	3.2%	2.8%	3.2%
STL	74.1%	77.0%	79.7%	77.0%	3.1%	2.2%	1.6%	2.3%
AutoArima	73.5%	81.7%	86.1%	80.4%	2.4%	2.6%	1.3%	2.1%
ETS	76.3%	88.7%	97.9%	87.7%	**1.5%**	**1.5%**	**-0.5%**	**0.9%**
VAR	79.6%	84.3%	90.4%	84.8%	4.1%	3.3%	2.7%	3.4%

Note: The best performing method for each metric-leadtime combination is indicated with **bold**.

siders interaction among categories in a store, but it performs worse than univariate AutoARIMA, potentially due to the high number of parameters representing cross-category interactions.

All components of the FAIR algorithm contribute to improving the bias (wMPE). FAIR has a lower bias than Boosted Trees and all the other benchmarks except for ETS which does not allow scenario evaluation.

Boosted Trees is the only benchmark method that can take into account the causal variables, cross-category interactions and time series effects. The potential reason for the Boosted Trees not having better accuracy is that the amount of available data was not sufficient for determining the right structure empirically. The only insight we can get into Boosted Trees models is the variable importance measure: Across leadtimes, 62% of the Gain measure is due to category-store specific historical mean, 20% is due to the lags of the focal category-store sales, 6% is due to marketing features of the focal category, 5% each are due to cross-category marketing and cross-category sales, finally about 1% is due to seasonal features. As the leadtime increases, the impact of focal category sales lags decreases (from 51% to 6%), while the importance of other features increases: historical mean (from 36% to 75%), focal marketing (5% to 7%), cross-category marketing (4% to 6%).

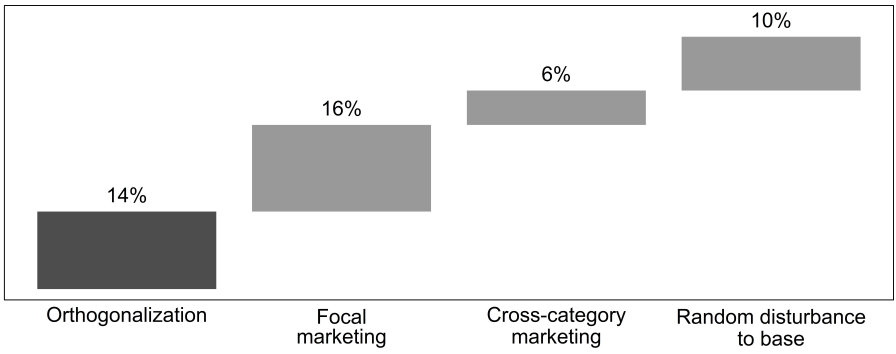


FIGURE 4: Accuracy contribution of components of the proposed FAIR method

Turning our attention to FAIR model components, we quantify the accuracy improvement due to consideration of a) focal category marketing, b) cross-category marketing and c) random disturbances to base sales as the difference in accuracy of successive models

(S, Sm, SM, FAIR) as a percentage of the accuracy of the FAIR model. The blue bars in Figure 4 show the RMSE contribution of each component. Accordingly, focal category marketing accounts for 16% improvement in RMSE. Cross-category marketing, as expected, contributes less (6% improvement), but still a substantial amount. Extrapolating random disturbances to the base sales contributes an additional 10% improvement in RMSE. The orange bar shows that the "No Orthogonalization" model has 14% worse accuracy than FAIR. In other words, orthogonalization with respect to the confounding variables contributes to the accuracy of FAIR substantially.

One of the important design elements of the proposed method is the use of the elastic net penalty versus the lasso penalty used commonly in forecasting methods. We see that the elastic net uses the full spectrum between a pure lasso ($\alpha = 1$) and pure ridge regression ($\alpha = 0$) depending on the data characteristics in the particular chain, category and origin, where the most frequent value is 0.1.

4.4 Interpreting the marketing effects

Figure 5 provides the average absolute value of estimated focal marketing feature coefficients, $\widehat{\varphi}_{ial}$, over all chain, category and time origin models. We observe that Regular_Price is the most important feature followed by Discount, Display, and Ad. How the discount resource is allocated is next in importance, particularly dist_5 and dist_100, which measure the category volume that is discounted within intervals 5-10% and 50-100% respectively. Basic marketing variables have repercussions for subsequent weeks too.

Now we look at a particular chain and time origin. Table A3 (Appendix) lists the estimated focal marketing coefficients, $\widehat{\varphi}_{ial}$. It includes the top two categories in terms of top marketing promotions, as identified in Section 5.1, Beer and Carbonated Beverages, and the category that has the most nonzero coefficients, which is Diapers. First, we see that the elastic net penalty does select features. Out of the 36 available features, Beer and Carbonated Beverages categories have 3 and 10 nonzero coefficients, respectively. For the Beer category, Display, Discount and dist_5, i.e. the percent of category volume that is discounted at less than 5% drive the sales. For the Carbonated Beverages category, Regular_Price, Discount, Display, Ad are all important; allocating discounts ei-

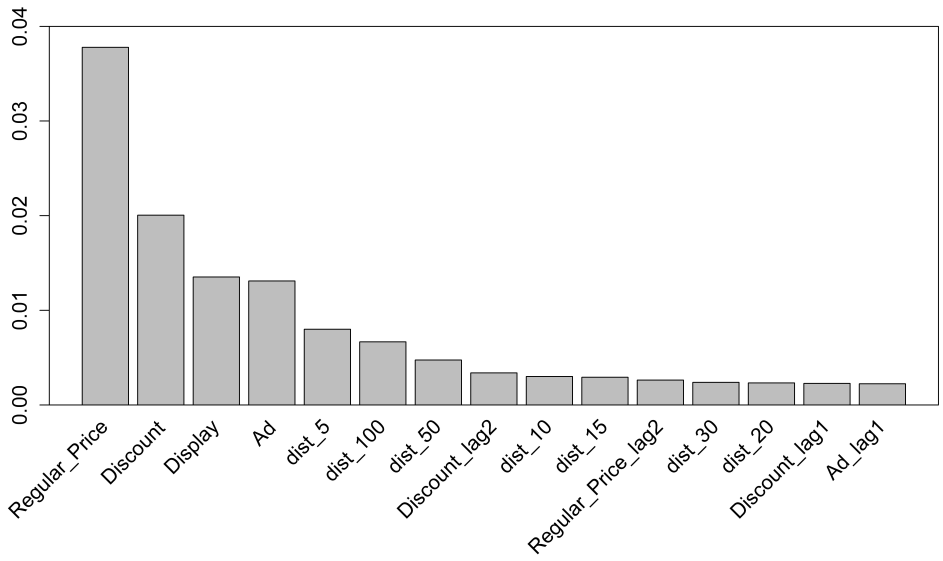


FIGURE 5: Average absolute value of estimated focal marketing feature coefficients, $\widehat{\varphi}_{ial}$

ther at deep (dist_100) or nominal (dist_5) percentages increases dollar sales in the same week and in the following weeks (dist_5_lag1, dist_100_lag2). The Diapers category exhibits many of the effects that the marketing literature has been documenting, such as the post-promotion-dip: While Discount increases sales in the same week, subsequent weeks see lower sales, pulling demand forward as customers stock diapers and reduce their purchases in subsequent weeks (Macé & Neslin, 2004). Ads have effects that carry over to multiple weeks (Givon & Horsky, 1990). Increase in the Regular_Price increases the sales revenue in the same week, while reducing it in the subsequent weeks as customers shop at alternative channels and the decrease in the sales quantity offsets the increase in the price. The estimated coefficients of the marketing variables in the FAIR method provide meaningful insights.

Turning our attention to cross-category effects of marketing, Figure 6 shows the average absolute value of the cross-category marketing feature coefficients, $\left|\widehat{\theta}_{ikal}\right|$, averaged over *a* and *l* for the same chain. Notice that the coefficients are for standardized features. The color scale in the figure indicates the strength of the relationship. We observe that Diaper is the most influential category, followed by Shampoo, while Household Cleaning is the most influenced category by cross-category marketing, followed by Photo. Based on the observed figure, the size of the cross-category interaction appears to be determined to



FIGURE 6: Average of absolute cross-category marketing coefficients in a sample chain

a large extent by the Influencer and Influenced category scores. There are a few category combinations where the actual interaction is stronger than what would be expected based on the general degree of the two categories to influence and be influenced. A couple of them are Spaghetti Sauce and Mustard_Ketchup category promotions affecting the Hot-dog category sales, which is a classic example of complementary goods effect (Leeflang & Parreno-Selva, 2012).

These are examples of insights that can be gleaned from the model for specific chain and situation and have face-validity in light of marketing literature.

4.5 Scenario Evaluation

The *h* step ahead forecasts of category-store sales are produced with the trained model and decisions about the level of marketing variables in the forecasting horizon, as explained in section 3.2. A potential base-case scenario may be to keep the marketing variables at their "expected levels" based on historical seasonality patterns. In that case, the category-store sales forecasts would simply be the sales expected based on seasonality patterns, multiplied by the factor $exp(\widehat{r}_{ijt})$ representing the expected impact of the recent random disturbances *h* steps ahead. The sales forecast for a marketing scenario that deviates from the historical patterns can be decomposed into the base-case scenario forecast introduced above and multipliers for each marketing feature. Suppose the coefficient for the marketing feature is 0.05, and the scenario feature value is 1% higher than in the base-case. This deviation will contribute a 0.05% increase in the sales forecast compared to the base-case.

5 Conclusions

In this paper, we proposed the FAIR method for multi-step ahead category-store level sales forecasting. The method is well positioned to support the medium-term retail planning process by providing an objective and data-driven tool. It evaluates the impact of different marketing scenarios on sales at the particular category-store at the particular time period in a consistent manner. FAIR is designed to deal with the large dimensionality of explanatory variables and regularization induced confounding, while incorporating heterogeneous and dynamic time series effects. The FAIR forecasts can be decomposed into what is expected based on historical patterns, the impact of own- and cross-category marketing, and the impact of random disturbances on base sales. The estimated parameters can be interpreted to gain insights into the impact of promotions on sales in the same or subsequent weeks and in the same category or other categories. The interpretability of the model is expected to promote trust and support better decisions. We illustrated insights into marketing impact based on interpretation of model coefficients, such as post-promotion dip, carryover, higher cross-category interaction among complementary categories - even though these effects were not forced by the estimation method.

Beyond providing insights and allowing scenario analysis with planning variables, the

extensive empirical evaluation shows that FAIR has better accuracy than all benchmarks over all forecast horizons (1 to 13 weeks), retail chains, and time origins. While Ferreira et al. (2015), Cui et al. (2018) and Gür Ali et al. (2009) showed that ensembles of trees forecast better than linear models, Boosted Trees is outperformed by FAIR. This shows that the domain knowledge driven structure of the FAIR method helps in high dimensional setups where events are not highly repetitive. This is in line with the findings for demand forecasting at Target: Yelland et al. (2019) also report that black box models such as random forests and neural networks did not achieve sufficient out-of-sample accuracy.

Empirical evaluations show that each component of FAIR adds to the accuracy of the method. The orthogonalization of sales and causal variables with respect to confounding variables improves accuracy substantially. Considering focal category marketing, cross-category marketing, and random disturbances to base sales, all improve forecasting accuracy at all lead times. While previous studies concluded that cross-category marketing has a marginal effect on SKU forecasts (Ma et al., 2016), we find that for category level forecasts, cross-category marketing has a substantial impact on accuracy.

While the particular case study involved only marketing variables, other causal variables can be included within the same setup, such as assortment decisions or competitor actions.

References

Aburto, L. & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, 7(1):136–144.

Ailawadi, K. L., Beauchamp, J. P., Donthu, N., Gauri, D. K., & Shankar, V. (2009). Communication and promotion decisions in retailing: A review and directions for future research. *Journal of Retailing*, 85(1):42–55.

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.

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Arnold, A., Liu, Y., & Abe, N. (2007). Temporal causal modeling with graphical granger methods. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, page 66.

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

Bergmeir, C. & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

Bronnenberg, B. J., Kruger, M. W., & Mela, C. F. (2008). Database paper—The IRI marketing data set. *Marketing Science*, 27(4):745–748.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, 184(3):1140–1154.

Carriero, A., Kapetanios, G., & Marcellino, M. (2012). Forecasting government bond yields with large Bayesian vector autoregressions. *Journal of Banking and Finance*, 36(7):2026–2047.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. R package version 0.6.4.1.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chevillon, G. & Hendry, D. F. (2005). Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting*, 21(2):201–218.

Choi, T. M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. *Production and Operations Management*, 27(10):1868–1883.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73.

Cooper, L. G., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). PromoCast^{TM}: A new forecasting method for promotion planning. *Marketing Science*, 18(3):301–316.

Cooper, L. G. & Giuffrida, G. (2000). Turning datamining into a management science tool: New algorithms and Empirical Results. *Management Science*, 46(2):249–264.

Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and Operations Management*, 27(10):1749–1769.

Curry, D. J., Divakar, S., Mathur, S. K., & Whiteman, C. H. (1995). BVAR as a category management tool: An illustration and comparison with alternative techniques. *Journal of Forecasting*, 14(3):181–199.

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

Dekker, M., Donselaar, K. V., & Ouwehand, P. (2004). How to use aggregation and

combined forecasting to improve seasonal demand forecasts. *International Journal of Production Economics*, 90(2):151–167.

Divakar, S., Ratchford, B. T., & Shankar, V. (2005). Practice prize article—CHAN4CAST: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. *Marketing Science*, 24(3):334–350.

Duncan, G. T., Gorr, W. L., & Szczypula, J. (2001). Forecasting analogous time series. In *Principles of forecasting*, (pp. 195-213). Springer, Boston, MA.

Ebbes, P., Papies, D., & Van Heerde, H. J. (2011). The sense and non-sense of holdout sample validation in the presence of endogeneity. *Marketing Science*, 30(6):1115–1122.

Efendigil, T., Önüt, S., Kahraman, C., Systems, E., & Applications, W. (2009). A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. *Expert Systems with Applications*, 36(August 2015):6697–6707.

Feng, Q. & Shanthikumar, J. G. (2018). How research in production and operations management may evolve in the era of big data. *Production and Operations Management*, 27(9):1670–1684.

Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2015). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 18(1):69–88.

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., Goodwin, P., & Önkal, D. (2019). Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting*, 35(1):144–156.

Fildes, R., Ma, S., & Kolassa, S. (2020). Retail forecasting: research and practice. *International Journal of Forecasting*, forthcoming.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1). R package version 2.0-16.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378.

Gelper, S., Wilms, I., & Croux, C. (2016). Identifying demand effects in a large network of product categories. *Journal of Retailing*, 92(1):25–39.

Givon, M. & Horsky, D. (1990). Untangling the effects of purchase reinforcement and advertising carryover. *Marketing Science*, 9(2):171–187.

Greene, W. W. H. . (2012). *Econometric analysis (7th ed.)*. Pearson

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 39(3):3659–3667.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Gür Ali, Ö. (2013). Driver moderator method for retail sales prediction. *International Journal of Information Technology & Decision Making*, 12(06):1261–1286.

Gür Ali, Ö. & Pınar, E. (2016). Multi-period-ahead forecasting with residual extrapolation and information sharing - Utilizing a multitude of retail series. *International Journal of Forecasting*, 32(2):502–517.

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

Gür Ali, Ö. & Yaman, K. (2013). Selecting rows and columns for training support vector regression models with large retail datasets. *European Journal of Operational Research*, 226(3):471–480.

Hahn, P. R., Carvalho, C. M., Puelz, D., & He, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge university press.

Harvey, S. K. & Cushing, M. J. (2014). Does using disaggregate components help in producing better forecasts for aggregate inflation? *Journal of Economics and Development Studies*, 2(2):527–546.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data Mining, Inference, and Prediction (2nd ed.)*. Springer New York Inc., New York.

Hsu, N.-J., Hung, H.-L., & Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis*, 52(7):3645–3657.

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.

Hübner, A. H., Kuhn, H., & Sternbeck, M. G. (2013). Demand and supply chain planning in grocery retail: an operations planning framework. *International Journal of Retail & Distribution Management*, 41(7):512–530.

Hyndman, R. J. & Athanasopoulos, G. (2014). *Forecasting: Principles and Practice*. OTexts.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2019). forecast: Forecasting functions for time series and linear models. R package version 8.2.

Hyndman, RJ & Khandakar, Y (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22.

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

Kolassa, S. et al. (2016). Sometimes It's Better to Be Simple than Correct. *Foresight: The International Journal of Applied Forecasting*, (40):20–26.

Leeflang, P. S. H. & Parreno-Selva, J. (2012). Cross-category demand effects of price promotions. *Journal of the Academy of Marketing Science*, 40(4):572–586.

Lemmens, A. & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research*, 43(2):276–286.

Ma, S. & Fildes, R. (2017). A retail store SKU promotions optimization model for category multi-period profit maximization. *European Journal of Operational Research*, 260(2):680–692.

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.

Macé, S. & Neslin, S. A. (2004). The determinants of pre-and postpromotion dips in sales of frequently purchased goods. *Journal of Marketing Research*, 41(3):339–350.

Natter, M., Reutterer, T., Mild, A., & Taudes, A. (2007). Practice prize report —An assortmentwide decision-support system for dynamic pricing and promotion planning in diy retailing. *Marketing Science*, 26(4):576–583.

Nicholson, W. B., Matteson, D. S., & Bien, J. (2017). VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, 33(3):627–651.

Oliva, R. & Watson, N. (2011). Cross-functional alignment in supply chain planning: A case study of sales and operations planning. *Journal of Operations Management*, 29(5):434–448.

Pfaff, B. (2008). VAR, SVAR and SVEC models: Implementation within R rackage vars. *Journal of Statistical Software*, 27(4). R package version 1.5-2.

R Core Team (2017). R: A language and environment for statistical computing.. *R Foundation for Statistical Computing, Vienna, Austria.* https://www.R-project.org/. Version 3.4.2.

Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*.

Shumway, R. H. and Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples*. Springer.

Snowdon, P. (1991). A ratio estimator for bias correction in logarithmic regressions. *Canadian Journal of Forest Research*, 21(5):720–724.

Stock, J. H. & Watson, M. W. (2003). Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115.

Sun, Z. L., Choi, T. M., Au, K. F., & Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems*, 46(1):411–419.

Taieb, S. B. & Hyndman, R. J. (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting*, 30(2):382–394.

Taylor, J. W. (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research*, 178(1):154–167.

Trapero, J. R., Pedregal, D. J., Fildes, R., and Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting*, 29(2):234–243.

Van Heerde, H. J., Leeflang, P. S., & Wittink, D. R. (2001). Semiparametric analysis to estimate the deal effect curve. *Journal of Marketing Research*, 38(2):197–215.

Van Heerde, H. J., Leeflang, P. S., & Wittink, D. R. (2002). How promotions work: SCAN*PRO-based evolutionary model building. *Schmalenbach Business Review*, 54(July):198–220.

Verbeek, M. (2008). *A guide to modern econometrics*. John Wiley & Sons.

Yelland, P., Baz, Z. E., & Serafini, D. (2019). Forecasting at scale: The architecture of a modern metail forecasting system. *Foresight: The International Journal of Applied Forecasting*, 55:10-18.

Zeng, J. (2017). Forecasting aggregates with disaggregate variables: Does boosting help to select the most relevant predictors? *Journal of Forecasting*, 36(1):74–90.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Statist. Soc. BJ. R. Statist. Soc. B*, 67(67):768–301.

A Appendix

TABLE A1
RMSE of the benchmark methods and FAIR versions by origin

Origin week starting on	Jul 3 2006	Oct 2 2006	Jan 1 2007	Apr 2 2007	Jul 2 2007	Oct 1 2007
S	552	568	837	766	803	773
Sm	493	512	804	674	662	619
SM	493	518	**802**	590	584	563
FAIR	**443**	**488**	806	**492**	**467**	**483**
FAIR No Orthogonalization	553	560	847	587	574	561
Boosted Trees	546	518	882	599	630	621
STL	517	533	845	513	524	534
AutoArima	574	572	914	611	581	631
ETS	777	661	1016	738	753	671
VAR	575	648	872	712	654	614

Note: The best performing method for each origin is indicated with **bold**.

TABLE A2
RMSE of the benchmark methods and FAIR versions by chain

	chain 14	chain 48	chain 117
S	715	494	991
Sm	622	417	890
SM	617	410	809
FAIR	**569**	**367**	**737**
FAIR No Orthogonalization	672	431	816
Boosted Trees	655	467	850
STL	655	399	768
AutoArima	747	454	834
ETS	839	620	942
VAR	757	461	895

Note: The best performing method for each chain is indicated with **bold**.

TABLE A3
Estimated focal category marketing coefficients, $\widehat{\varphi}_{ial}$, for a sample chain and origin, showing top promotion categories and the category with most nonzero coefficients

Feature	Beer	Feature	carbbev	Feature	diapers
Display	0.154	Regular_Price	0.896	Ad	0.270
dist_5	0.080	Discount	0.134	Ad_lag1	0.011
Discount	0.026	dist_100	0.116	Ad_lag2	0.053
		Ad_lag1	0.050	Discount	0.796
		Display	0.046	Discount_lag1	-1.152
		dist_5	0.031	Discount_lag2	-0.469
		dist_5_lag1	0.031	Display	0.991
		Ad	0.026	Display_lag1	0.105
		dist_100_lag2	0.018	Display_lag2	-0.103
		dist_10_lag2	0.007	Regular_Price	0.378
				Regular_Price_lag1	-0.224
				Regular_Price_lag2	-0.087
				dist_10	0.083
				dist_10_lag1	0.180
				dist_10_lag2	0.130
				dist_100	2.695
				dist_100_lag1	0.480
				dist_100_lag2	-0.214
				dist_15	0.250
				dist_15_lag1	0.193
				dist_15_lag2	0.183
				dist_20	0.170
				dist_20_lag1	0.074
				dist_20_lag2	0.127
				dist_30	0.265
				dist_30_lag1	0.101
				dist_30_lag2	-0.005
				dist_40	0.868
				dist_40_lag1	0.459
				dist_40_lag2	0.101
				dist_5	0.254
				dist_5_lag1	0.030
				dist_5_lag2	0.064
				dist_50	4.106
				dist_50_lag1	1.545
				dist_50_lag2	2.273