ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

*Appl. Statist.* (2018)
**67**, *Part* 2, *pp.* 435–452

# Multiclass vector auto-regressive models for multistore sales data

Ines Wilms,

*KU Leuven, Belgium, and Cornell University, Ithaca, USA*

Luca Barbaglia

*KU Leuven, Belgium*

and Christophe Croux

*EDHEC Business School, Lille, France*

[Received May 2016. Final revision May 2017]

**Summary.** Retailers use the vector auto-regressive (VAR) model as a standard tool to estimate the effects of prices, promotions and sales in one product category on the sales of another product category. Besides, these price, promotion and sales data are available not just for one store, but for a whole chain of stores. We propose to study cross-category effects by using a multiclass VAR model: we jointly estimate cross-category effects for several distinct but related VAR models, one for each store. Our methodology encourages effects to be similar across stores, while still allowing for small differences between stores to account for store heterogeneity. Moreover, our estimator is sparse: unimportant effects are estimated as exactly 0, which facilitates the interpretation of the results. A simulation study shows that the multiclass estimator proposed improves estimation accuracy by borrowing strength across classes. Finally, we provide three visual tools showing clustering of stores with similar cross-category effects, networks of product categories and similarity matrices of shared cross-category effects across stores.

*Keywords*: Fused lasso; Multiclass estimation; Multistore sales application; Sparse estimation; Vector auto-regressive model

## 1. Introduction

Successful cross-category management requires retailers to understand 'cross-category demand effects', i.e. the effects of prices, promotions and sales of a certain product category on the sales (or demand) of another product category. The vector auto-regressive (VAR) model is ideal to measure such cross-category demand effects. In the $J$-dimensional VAR model of order $P$, the values of the $J$-price, promotion and sales time series are modelled as a function of their own past values, up to $P$ periods ago. As such, the VAR model accounts for time inertia in marketing spending and treats price and promotion variables as endogenous, thereby allowing feedback effects (e.g. Dekimpe and Hanssens (1995)). The relevance of cross-category analysis in the marketing literature has been widely acknowledged (e.g. Leeflang and Selva (2012) and references therein).

To analyse cross-category demand effects, retailers typically prefer to work with store level

*Address for correspondence*: Ines Wilms, Faculty of Economics and Business, KU Leuven, Naamsestraat 69, Leuven 3000, Belgium.
E-mail: ines.wilms@kuleuven.be

© 2017 Royal Statistical Society 0035–9254/18/67435

data. However, information on prices, promotions and sales is available not for only one store but typically also for an entire chain of stores. A 'multiclass' VAR approach where we jointly estimate several distinct but related VAR models—one for each store—is to be preferred to a standard VAR model. Our multiclass approach has several important advantages.

(a) Cross-category demand effects are expected to be *similar* for the different stores since they belong to the same retail chain. We therefore encourage estimates to be similar between classes. As such, retailers can set a chainwide marketing strategy for the shared dynamics across stores.
(b) At the same time, we allow for differences between stores stemming from the *heterogeneity* in shopping behaviour at the different stores. As such, retailers can fine-tune their chainwide strategy to accommodate store-specific effects.
(c) By jointly estimating the multiple VAR models, we borrow strength across classes, which results in *improved estimation accuracy*, as will be illustrated by means of a simulation study.
(d) Our estimation method is '*sparse*' in the sense that some parameters are estimated as 0. Sparse estimation techniques have proven their worth in delivering highly interpretable VAR models in high dimensional settings; see among others Hsu *et al*. (2008), Abegaz and Wit (2013), Basu *et al*. (2015), Davis *et al*. (2016) and Gelper *et al*. (2016).

Sparse multiclass estimators have been recently introduced for graphical models (Danaher *et al*., 2014), and regression models (Kim and Xing, 2009). Our sparse multiclass estimator of the VAR model differs from the method of Kim and Xing (2009) in that

(a) we consider a time series framework instead of a regression framework,
(b) we allow for a multivariate instead of a univariate response model for each class,
(c) we account for the correlation structure between the error terms of different equations of the VAR model and
(d) we use the smoothing proximal gradient algorithm.

The remainder of this paper is structured as follows. Section 2 introduces the multiclass VAR model, the corresponding estimator and algorithm. Simulation studies in Section 3 show the good performance of the proposed estimator in terms of estimation accuracy. Section 4 presents the data and model for the multistore sales application; Section 5 discusses the results. Section 6 discusses three possible extensions of the basic multiclass VAR approach. Finally, Section 7 concludes.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

## 2. Multiclass vector auto-regressive models

### 2.1. Model and estimator

Price, promotion and sales of several categories are available for each store (i.e. class) $1 \leqslant k \leqslant K$ over a certain period of time. Let $\mathbf{y}_t^{(k)} = (y_{t,1}^{(k)}, \ldots, y_{t,J}^{(k)})'$ be a $J$-dimensional multivariate time series containing these price, promotion and sales data for store $k$ at a given point in time $1 \leqslant t \leqslant T$ where $T$ is the length of the time series. The multiclass VAR model of order $P$ with $K$ classes and $J$ time series is given by

$$\mathbf{y}_t^{(k)} = B_1^{(k)} \mathbf{y}_{t-1}^{(k)} + \ldots + B_P^{(k)} \mathbf{y}_{t-P}^{(k)} + \mathbf{e}_t^{(k)}. \tag{1}$$

The parameters $B_p^{(k)}$, for $1 \leqslant p \leqslant P$ and $1 \leqslant k \leqslant K$, are $J \times J$ matrices including all the auto-regressive coefficients at lag $p$ for class $k$. The $ij$th entry of $B_p^{(k)}$ is denoted by $[B_p^{(k)}]_{ij} := \beta_{p,ij}^{(k)}$, for $1 \leqslant i, j \leqslant J$. This element measures the direct effect for class $k$ of time series $j$ on time series $i$ at lag $p$. As such, we measure for each store $k$ the direct lagged effects of prices, promotions and sales in one category on the prices, promotions and sales of another category (including its own). Assume that the error terms $\mathbf{e}_t^{(k)}$ are independent over time and follow a multivariate white noise distribution with mean 0 and covariance matrix $\Sigma^{(k)}$. We assume, without loss of generality, that all the time series are mean centred such that no intercept is included.

We estimate the model parameters by penalized generalized least squares (LS). For ease of notation, rewrite model (1) in stacked form as

$$\mathbf{y}^{(k)} = X^{(k)}\boldsymbol{\beta}^{(k)} + \mathbf{e}^{(k)}, \tag{2}$$

where $\mathbf{y}^{(k)}$ for $1 \leqslant k \leqslant K$ is an $NJ$-vector stacking all $J$ time series, with $N = T - P$. For each class $k$, $X^{(k)} = I_J \otimes X_0^{(k)}$, where the $N \times JP$ matrix $X_0^{(k)}$ is defined as $X_0^{(k)} = (Y_1^{(k)}, \ldots, Y_P^{(k)})$, with $Y_p^{(k)}$ being an $N \times J$ matrix collecting the observations at lag $p$ for the $J$ series in the $k$th class. The symbol '$\otimes$' is the Kronecker product. Furthermore, $\boldsymbol{\beta}^{(k)} = (\beta_{1,11}^{(k)}, \ldots, \beta_{P,JJ}^{(k)})'$, and $\mathbf{e}^{(k)}$ is the $NJ$-vector of stacked error components for each class $k$.

We define the estimator $\hat{\boldsymbol{\beta}}$ of the vector $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)\prime}, \ldots, \boldsymbol{\beta}^{(K)\prime})'$, collecting the auto-regressive parameters for all classes, as the minimizer of the following penalized LS criterion:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{K} (\mathbf{y}^{(k)} - X^{(k)}\boldsymbol{\beta}^{(k)})'(\mathbf{y}^{(k)} - X^{(k)}\boldsymbol{\beta}^{(k)}) + \lambda_1 P_1(\boldsymbol{\beta}) + \lambda_2 P_2(\boldsymbol{\beta}), \tag{3}$$

where $\lambda_1, \lambda_2 > 0$ are regularization parameters and $P_1(\boldsymbol{\beta})$ and $P_2(\boldsymbol{\beta})$ are two penalty functions.

For the first penalty function, we take the $l_1$-penalty on the absolute value of the differences of corresponding auto-regressive parameters across classes (e.g. Tibshirani *et al.* (2005) and She (2010)):

$$P_1(\boldsymbol{\beta}) = \sum_{k<k'}^{K} \sum_{i,j=1}^{J} \sum_{p=1}^{P} |\beta_{p,ij}^{(k)} - \beta_{p,ij}^{(k')}|. \tag{4}$$

The pairwise absolute difference penalty that is used in equation (4) is commonly referred to in the literature as the fused lasso. The aim of this penalty is to induce similarity across classes. The larger the value of $\lambda_1$, the more differences of corresponding auto-regressive parameters will be set to 0. As a consequence, the more elements of $\hat{B}_p^{(1)}, \ldots, \hat{B}_p^{(K)}$, for $1 \leqslant p \leqslant P$, will be identical across classes. If $\lambda_1 \to \infty$, all the corresponding auto-regressive parameters across classes will be identical and, hence, the same VAR model is obtained for each class $k$. If $\lambda_1 = 0$, then each class $k$ has its own VAR model and there are no similarities across classes. Since some of the estimated auto-regressive parameters will be identical for some classes, a 'clustering' of classes arises for each estimated auto-regressive parameter, where all classes with the same estimated parameter value form a cluster.

For the second penalty function, we consider an adaptive $l_1$-penalty on the absolute value of the auto-regressive parameters

$$P_2(\boldsymbol{\beta}) = \sum_{k=1}^{K} \sum_{i,j=1}^{J} \sum_{p=1}^{P} \hat{w}_{p,ij}^{(k)} |\beta_{p,ij}^{(k)}|, \tag{5}$$

where $\hat{w}_{p,ij}^{(k)}$ are weights. This adaptive $l_1$-penalty (Zou, 2006) generalizes the popular $l_1$-penalty (Tibshirani, 1996) and enjoys good theoretical properties. We take the weights as $\hat{w}_{p,ij}^{(k)} = 1/|\hat{\beta}_{p,ij}^{(k),\mathrm{R}}|$, where $\hat{\beta}^{\mathrm{R}}$ is the ridge estimator (Hastie *et al.* (2009), chapter 3). The aim of this adaptive $l_1$-penalty is twofold. First, by adding this penalty to the objective function, estimation remains feasible if the number of parameters exceeds the time series length. Second, it induces

sparsity in the estimated auto-regressive parameters by setting some coefficients equal to 0. The larger the value of $\lambda_2$, the sparser the estimate of $\boldsymbol{\beta}$. The combination of the first and second penalty in the objective function in equation (3) leads towards shared sparsity patterns across classes.

We further improve estimator (3) by simultaneously estimating the correlation structure of the error terms. For this, we include the inverse error covariance matrices $\Omega = (\Omega^{(1)}, \ldots, \Omega^{(K)})'$ in the objective function, where $\Omega^{(k)} = (\Sigma^{(k)})^{-1}$ for $1 \leqslant k \leqslant K$. The $ij$th entry of $\Omega^{(k)}$ is denoted by $[\Omega^{(k)}]_{ij} := \omega_{ij}^{(k)}$, for $1 \leqslant i, j \leqslant J$. We define the *multiclass* estimator as the minimizer of the following penalized generalized LS criterion:

$$(\hat{\boldsymbol{\beta}}, \hat{\Omega}) = \underset{\boldsymbol{\beta}, \Omega}{\arg\min} \sum_{k=1}^{K} \{(\mathbf{y}^{(k)} - X^{(k)}\boldsymbol{\beta}^{(k)})' \tilde{\Omega}^{(k)} (\mathbf{y}^{(k)} - X^{(k)}\boldsymbol{\beta}^{(k)}) - NJ \log |\Omega^{(k)}|\} + \lambda_1 P_1(\boldsymbol{\beta}) + \lambda_2 P_2(\boldsymbol{\beta}) + \gamma_1 P_1(\Omega) + \gamma_2 P_2(\Omega), \tag{6}$$

where $\tilde{\Omega}^{(k)} = \Omega^{(k)} \otimes I_N$, and $\gamma_1, \gamma_2 > 0$ are regularization parameters for the elements of the inverse error covariance matrices. Similarly to equation (4), we use a penalty on the corresponding elements of the inverse error covariance matrices of the different classes, together with a standard $l_1$-penalty on the elements of the inverse error covariance matrices:

$$\begin{aligned} P_1(\Omega) &= \sum_{k<k'}^{K} \sum_{i,j=1}^{J} |\omega_{ij}^{(k)} - \omega_{ij}^{(k')}|, \\ P_2(\Omega) &= \sum_{k=1}^{K} \sum_{i,j=1}^{J} |\boldsymbol{\omega}_{ij}^{(k)}|. \end{aligned} \tag{7}$$

Hence, the larger the value of $\gamma_1$, the more elements of $\hat{\Omega}^{(1)}, \ldots, \hat{\Omega}^{(K)}$ will be identical across classes. The larger the value of $\gamma_2$, the sparser the estimate of $\Omega$ will be. Moreover, the penalty $P_2(\Omega)$ ensures that the estimate of the inverse error covariance matrix exists even when the number of parameters exceeds the time series length. The elements of the inverse error covariance matrices $\Omega^{(k)}$ have a natural interpretation as the partial correlations between the error terms of the $J$ equations for class $k$. If the $ij$th element of $\Omega^{(k)}$ is equal to 0, this means that the error terms of equation $i$ and $j$ for class $k$ are uncorrelated given all the other equations.

### 2.2. Algorithm

This section provides technical details on the implementation of the algorithm. We iteratively solve the optimization problem (6) first considering $\boldsymbol{\beta}$ conditional on $\Omega$ and then $\Omega$ conditional on $\boldsymbol{\beta}$. The code of the algorithm is available from the Web page of the first author (`http://feb.kuleuven.be/ines.wilms/software`) and from `http://wileyonlinelibrary.com/journal/rss-datasets`. The on-line supplementary file contains a step-by-step illustration on how to use the code to replicate the application results.

#### 2.2.1. Solving for $\boldsymbol{\beta}$ conditional on $\Omega$

We build on Chen *et al.* (2012) and extend their smoothing proximal gradient algorithm for sparse estimation of regression models. The algorithm optimizes a smooth approximation of the objective function (see also Nesterov (2005)):

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min}\ g(\boldsymbol{\beta}) + h_\mu(\boldsymbol{\beta}) + \lambda_2 P_2(\boldsymbol{\beta}), \tag{8}$$

where $g(\boldsymbol{\beta})$ is the first term in the objective function in problem (6) with $\Omega$ kept constant, and we replace the term $\lambda_1 P_1(\boldsymbol{\beta})$ with its smooth approximation

$$h_\mu(\boldsymbol{\beta}) = \max_{\|\boldsymbol{\alpha}\|_\infty \leqslant 1} \left( \boldsymbol{\alpha}' C\boldsymbol{\beta} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2 \right),$$

with $\mu > 0$ a smoothing parameter, $\boldsymbol{\alpha}$ a vector of auxiliary variables and $C = I_P \otimes \tilde{C}$ a $(K-1)(d/2) \times d$ matrix, with $d = \dim(\boldsymbol{\beta})$, representing the pairs of coefficients that are coupled across classes. We take $\tilde{C} = ((\tilde{C}_1 \otimes I_{J^2})', (\tilde{C}_2 \otimes I_{J^2})', \ldots, (\tilde{C}_{K-1} \otimes I_{J^2})')'$ with

$$[\tilde{C}_k]_{ij} = \begin{cases} \lambda_1 & \text{if } j = i, \\ -\lambda_1 & \text{if } j = i+k, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

for $1 \leqslant k \leqslant K-1$, $1 \leqslant i \leqslant K-k$ and $1 \leqslant j \leqslant K$. The solution of the objective function in equation (8) is approximated by using the fast iterative shrinkage thresholding algorithm (Beck and Teboulle, 2009).

Note that we choose the smoothing proximal gradient algorithm over other standard first-order methods since it has a theoretically faster rate of convergence and it is more scalable to high dimensional problems because of its lower per-iteration time complexity; see Chen *et al.* (2012).

*2.2.1.1. Selection of regularization parameters.* We use a two-dimensional grid of regularization parameters $\lambda_1$ and $\lambda_2$ and search for the optimal parameters minimizing the Bayesian information criterion

$$\mathrm{BIC}_{\lambda_1,\lambda_2} = -2g(\tilde{\boldsymbol{\beta}}_{\lambda_1,\lambda_2}) + \mathrm{df}_{\lambda_1,\lambda_2} \log(N),$$

where $\tilde{\boldsymbol{\beta}}_{\lambda_1,\lambda_2}$ is the estimator using the regularization parameters $\lambda_1$ and $\lambda_2$ and $\mathrm{df}_{\lambda_1,\lambda_2}$ is the number of non-zero estimated components of $\tilde{\boldsymbol{\beta}}_{\lambda_1,\lambda_2}$.

An alternative to BIC is AICc, a corrected version of the Akaike information criterion AIC with better finite sample performance for small sample sizes (see for example Hurvich and Tsai (1989)). Our simulations indicate no significant difference in terms of estimation accuracy of the multiclass estimator when using either BIC or AICc to select the regularization parameters.

*2.2.2. Solving for $\Omega$ conditional on $\boldsymbol{\beta}$*

When $\boldsymbol{\beta}$ is fixed, the estimation in problem (6) corresponds to the joint graphical lasso (Danaher *et al.*, 2014) on the residuals $\mathbf{e}^{(k)} = \mathbf{y}^{(k)} - X^{(k)}\boldsymbol{\beta}^{(k)}$, for $1 \leqslant k \leqslant K$. The joint graphical lasso is computed by using the fast alternating directions method-of-multipliers algorithm. The optimal values of the regularization parameters $\gamma_1$ and $\gamma_2$ are selected by using BIC (e.g. Yuan and Lin (2007)).

*2.2.3. Starting value and convergence*

We start by taking $\Omega^{(1)} = \ldots = \Omega^{(K)} = I_J$, and then we solve for $\boldsymbol{\beta}$ conditional on $\Omega$ and for $\Omega$ conditional on $\boldsymbol{\beta}$. We iterate until the relative change in the value of the objective function in problem (6) in two successive iterations is smaller than the tolerance value $\varepsilon = 10^{-5}$. Convergence was reached in all simulation runs and the real data example.

## 3. Simulation study

We compare the performance of the proposed 'multiclass' estimator, i.e. the solution of equation (6), with three alternative estimators:

(a) the 'single-class' estimator, i.e. the solution of equation (6) with $\lambda_1 = \gamma_1 = 0$ (the VAR model is then estimated sparsely but no similarities across classes are induced);
(b) the 'LS' estimator, where every VAR model is estimated separately for each class;
(c) the 'LS clustered' estimator, as a simple alternative suggested by a referee, where hierarchical clustering of estimated effects of the LS estimator is performed.

The last two estimators are not regularized and can only be computed if the number of parameters does not exceed the time series length.

More specifically, the LS clustered estimator relies on a distance matrix given by the Euclidean distances between the LS-estimated effects for each pair of stores. As a clustering agglomeration method, we use Ward's minimum variances method. To determine the number of clusters for each estimated effect, we use the Tracew index (see for example Milligan and Cooper (1985)). Other choices result in very similar performance. Within each cluster, the estimated effects are then put equal to the average estimated effect across the cluster members.

We simulate from a multiclass VAR model of order $P = 1$ with $K = 15$ classes and $J = 10$ time series. These dimensions are similar to those of our multistore sales application. The data-generating process for each class $k$ is

$$\mathbf{y}_t^{(k)} = B_1^{(k)} \mathbf{y}_{t-1}^{(k)} + \mathbf{e}_t^{(k)},$$

for $t = P+1, \ldots, T = 100$ and $\mathbf{e}_t^{(k)}$ follows a multivariate normal distribution with zero mean and covariance matrix $\Sigma^{(k)}$.

### 3.1. Simulation designs

We consider five simulation designs.

(a) *Design 1, varying $\beta$*: we take

$$B_1^{(k)} = \begin{pmatrix} A_1^{(k)} & A_2^{(k)} \\ \mathbf{0} & A_1^{(k)} \end{pmatrix}$$

with

$$A_1^{(k)} = \begin{pmatrix} 0.5 & \eta^{(k)} & \eta^{(k)} & \eta^{(k)} & \eta^{(k)} \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \end{pmatrix}$$

and

$$A_2^{(k)} = \begin{pmatrix} \eta_{1\times 5}^{(k)} \\ \mathbf{0}_{4\times 5} \end{pmatrix},$$

where $\eta_{1\times 5}^{(k)}$ is a $1 \times 5$ matrix whose entries are all equal to $\eta^{(k)}$ and $\mathbf{0}_{4\times 5}$ is a $4 \times 5$ matrix of 0s. We include dynamics among the different time series: time series 2–10 lead time series 1, whereas time series 7–10 lead time series 6. Across classes, the auto-regressive coefficients have the same sparsity structure, whereas the magnitude of the non-zero effects varies with

$$\eta^{(k)} = \begin{cases} 0.20 & \text{if } 1 \leqslant k \leqslant 5, \\ 0.25 & \text{if } 6 \leqslant k \leqslant 10, \\ 0.30 & \text{if } 11 \leqslant k \leqslant 15. \end{cases}$$

Averaged across classes, the cross-effects are half the magnitude of the own lagged effects, and stationarity of the VAR model is ensured. The error covariance matrices are the same for all classes $\Sigma^{(1)} = \ldots = \Sigma^{(K)} = 0.5 I_J$.

(b) *Design 2, varying* $\Sigma$*, low dependence*: the error covariance matrices are dense matrices with the same sparsity pattern

$$[\Sigma^{(k)}]_{ij} = 0.5 \rho_{(k)}^{|i-j|}$$

across classes. The magnitude of the correlations varies across classes with

$$\rho_{(k)} = \begin{cases} 0.05 & \text{if } 1 \leqslant k \leqslant 5, \\ 0.10 & \text{if } 6 \leqslant k \leqslant 10, \\ 0.15 & \text{if } 11 \leqslant k \leqslant 15. \end{cases} \tag{10}$$

The auto-regressive parameters are the same across all classes. We take

$$B_1^{(1)} = \ldots = B_1^{(K)} = \begin{pmatrix} A_3 & A_4 \\ \mathbf{0} & 0.5 I_5 \end{pmatrix}$$

with

$$A_3 = \begin{pmatrix} 0.5 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \end{pmatrix}$$

and

$$A_4 = \begin{pmatrix} 0.25_{4\times 5} \\ \mathbf{0}_{4\times 5} \end{pmatrix}.$$

(c) *Design 3, varying* $\Sigma$*, high dependence*: compared with design 2, we increase the dependence structure in the error covariance matrices by adding 0.6 to each correlation in expression (10). All other parameter choices remain unchanged.

(d) *Design 4, varying* $\boldsymbol{\beta}$ *and* $\Sigma$*, low dependence*: the value of $\boldsymbol{\beta}$ is taken from design 1; the value of $\Sigma$ is taken from design 2.

(e) *Design 5, varying* $\boldsymbol{\beta}$ *and* $\Sigma$*, high dependence*: the value of $\boldsymbol{\beta}$ is taken from design 1; the value of $\Sigma$ is taken from design 3.

### 3.2. Performance measures

We compare the performance of the estimators in terms of their estimation accuracy. Estimation accuracy is evaluated by the mean absolute estimator error

$$\text{MAEE} = \frac{1}{R} \frac{1}{PKJ^2} \sum_{r=1}^{R} \sum_{k=1}^{K} \sum_{i,j=1}^{J} \sum_{p=1}^{P} |\hat{\beta}_{p,ij,r}^{(k)} - \beta_{p,ij,r}^{(k)}|,$$

where $\hat{\beta}_{p,ij,r}^{(k)}$ is the estimate of $\beta_{p,ij}^{(k)}$ in simulation run $r$. We take $R = 1000$ runs.

### 3.3. Results

Table 1 reports the MAEE of the four estimators for the five simulation designs. Standard errors around the reported numbers are all smaller than 0.001. For the 'varying $\boldsymbol{\beta}$' design, where the errors of the different equations of each VAR model are not correlated, the multiclass estimator attains a lower value of the MAEE than the single-class estimator: 0.0825 *versus* 0.0947

**Table 1.** Simulated mean absolute estimation error of the four estimators for the five simulation designs

| *Design* | *Results for the following estimators:* | | | |
|---|---|---|---|---|
| | *LS* | *LS clustered* | *Single class* | *Multi-class* |
| Varying $\beta$ | 0.1239 | 0.1189 | 0.0947 | 0.0825 |
| Varying $\Sigma$, low dependence | 0.1108 | 0.1053 | 0.0824 | 0.0727 |
| Varying $\Sigma$, high dependence | 0.1512 | 0.1419 | 0.0848 | 0.0834 |
| Varying $\beta$ and $\Sigma$, low dependence | 0.1241 | 0.1192 | 0.0962 | 0.0836 |
| Varying $\beta$ and $\Sigma$, high dependence | 0.1576 | 0.1492 | 0.1000 | 0.0971 |

respectively. Accounting for the shared sparsity patterns across classes thus improves estimation accuracy. The difference in estimation accuracy is significant, as confirmed by a paired $t$-test with $p$-value less than 0.01.

The two regularized estimators perform significantly better than the LS and LS clustered estimators. Since the number of parameters to be estimated is large compared with the time series length, the LS estimator suffers from imprecise estimation accuracy. The LS clustered estimator inherits this estimation imprecision.

The conclusions for the other four designs are similar:

(a) the multiclass estimator attains the best estimation accuracy,
(b) in each design, the multiclass estimator performs significantly better than the single-class estimator in terms of MAEE. In terms of mean-squared estimation error (unreported), the multiclass estimator is significantly better in designs 1, 2 and 4, and there is no significant difference in designs 3 and 5.
(c) the regularized estimators significantly outperform the LS and LS clustered estimators.

Since the *multiclass* estimator attains the best overall estimation accuracy, we use this estimator to study the cross-category demand effects across multiple stores.

## 4. Data and model

We use data from Dominick's Finer Foods, which is a large midwestern supermarket chain that operates in the Chicago metropolitan area. This database is well established in the literature on cross-category analysis (e.g. Wedel and Zhang (2004), Kamkura and Kang (2007) and Lang *et al.* (2015)). Weekly store level scanner data are available on prices, promotions and sales. For more information on the calculation of the prices, promotions and sales variables, see for example Srinivasan *et al.* (2004). We analyse cross-category demand effects between five categories involving drink items: soft drinks, SDR, refrigerated juices, RFJ, beer, BER, bottled juices, BJC, and frozen juices, FRJ. These data were collected for $K = 15$ stores over a period from January 1993 to July 1994: $T = 76$ weeks in total.

Store-specific information is provided in Table 2. Dominick adopts a price-tier-specific pricing strategy where each store belongs to one out of four price tier groups, i.e. cub fighter, low, medium or high price tier. Cub fighters pursue a more aggressive pricing policy in comparison with the other price tiers. We consider two cub fighter, two low price tier, seven medium price tier and four high price tier stores. Table 2 also presents demographic characteristics of the

**Table 2.** Store-specific price and demographic information†

| *store* | *price tier* | *income* | *educ* | *ethnic* | *hsizeavg* | *hvalmean* |
|---|---|---|---|---|---|---|
| 1 | Cub fighter | 10.716 | 0.178 | 0.105 | 3.110 | 120.134 |
| 2 | Cub fighter | 10.715 | 0.233 | 0.024 | 2.955 | 142.408 |
| 3 | Low | 10.597 | 0.095 | 0.035 | 2.770 | 97.501 |
| 4 | Low | 10.797 | 0.284 | 0.051 | 2.556 | 160.003 |
| 5 | Medium | 10.787 | 0.222 | 0.033 | 2.617 | 168.277 |
| 6 | Medium | 10.620 | 0.172 | 0.025 | 2.785 | 143.828 |
| 7 | Medium | 10.831 | 0.238 | 0.041 | 2.615 | 194.229 |
| 8 | Medium | 10.480 | 0.071 | 0.042 | 2.491 | 119.381 |
| 9 | Medium | 10.505 | 0.050 | 0.268 | 2.661 | 68.224 |
| 10 | Medium | 10.574 | 0.052 | 0.165 | 2.706 | 84.720 |
| 11 | Medium | 10.660 | 0.175 | 0.087 | 2.517 | 148.950 |
| 12 | High | 11.043 | 0.348 | 0.034 | 2.735 | 218.997 |
| 13 | High | 10.674 | 0.198 | 0.032 | 2.401 | 174.439 |
| 14 | High | 10.600 | 0.270 | 0.066 | 2.555 | 158.496 |
| 15 | High | 10.188 | 0.160 | 0.221 | 2.516 | 125.168 |

†price tier, price tier group to which the store belongs; income, logarithm of median income; educ, percentage of college graduates; ethnic, percentage of blacks and Hispanics; hsizeavg, average household size; hvalmean, mean household value.

consumers in each store's market area, namely income, the logarithm of median income, educ, the percentage of college graduates, ethnic, the percentage of blacks and Hispanics, hsizeavg, the average household size, and hvalmean the mean household value.

We analyse cross-category demand effects in a multiclass VAR model consisting of $J = 3 \times 5$ time series, for each of the $K = 15$ classes and $T = 76$ time points. The order of the VAR model is selected by using BIC and gives $P = 1$. The estimated auto-regressive parameters $\hat{B}_1^{(k)}$ from the multiclass VAR model in equation (1) capture the within- and cross-category effects for store $1 \leqslant k \leqslant K$. *Within-category* effects are the effects of prices, promotion or sales on its own prices, promotions or sales. *Cross-category* effects are the effects of prices, promotion or sales of a certain category on the prices, promotion or sales of another category.

Using the multivariate portmanteau statistic (Tsay (2014), page 72), we find that for none of the 15 stores is the null hypothesis of multivariate white noise for the residuals rejected. Furthermore, graphical inspection of the residuals also supports validity of the model.

## 5. Multistore sales application

As is common in the literature on cross-category analysis, we focus on the cross-category *demand* effects, i.e. the effects of prices, promotion and sales of a certain category on the sales (or demand) of another category. A good understanding of these demand effects is valuable to retailers to allocate their scarce marketing resources across categories better.

Previous studies on cross-category demand effects either

(a) focus on a single store (e.g. Leeflang and Selva (2012)),
(b) estimate separate models, one for each store (e.g. Wedel and Zhang (2004) and Gelper *et al*. (2016)), or
(c) aggregate information from several stores (e.g. Song and Chintagunta (2006)).

The first two approaches do not exploit the similarity between stores belonging to the same retail chain. Moreover, separate store models are likely to produce more noisy, less stable estimates

(Lang *et al.*, 2015). The third approach is likely to produce biased estimates since it ignores the fact that the data belong to different stores (Kamkura and Kang, 2007). Moreover, the differences between stores are of interest to retailers wanting to set a store-specific strategy. In contrast with these previous studies, we use the multiclass VAR approach from Section 2.

We compute standard errors of the parameter estimates by adapting the residual bootstrap procedure of Chatterjee and Lahiri (2011) to the time series setting, as in Kreiss and Lahiri (2012). Full details are provided in Appendix A. Note that an important area of current research on lasso-type methods focuses on inference after model selection (e.g. Lee *et al.* (2016)). Since we estimate $PKJ^2 = 3375$ auto-regressive parameters, we cannot report all of them. A small subset of the estimated parameters with corresponding standard errors is reported in Table 1 of the on-line supplementary file. To present the results parsimoniously, we provide three visual tools showing

(a) clustering of stores with similar cross-category demand effects,
(b) product category networks and
(c) similarity matrices of cross-category demand effects across stores.

### 5.1. Store clustering

In Fig. 1, we consider three typical examples of estimated cross-category demand effects. For each of the 15 stores, the estimated cross-category demand effect and the corresponding 95% confidence interval, computed by using the bootstrap standard errors, are presented in Fig. 1.

First, consider the estimated effects of beer prices on refrigerated juices sales; see Fig. 1(a). For most stores, refrigerated juices sales are unresponsive (i.e. non-significant estimates indicated in grey) to a change in Dominick's beer pricing. The low income, low educated shoppers (low values of income and educ in Table 2) at store 15 are more subject to substitution effects: a price increase of beer makes them substitute refrigerated juices for beer. In contrast, the small households with large homes (low value of hsizeavg, high value of hvalmean in Table 2) at store 6 and 13, or the high income, high educated shoppers at store 4 (high values of income and educ in Table 2) are less vulnerable to substitution effects. Pairwise comparisons indicate that the estimated effects of stores 4, 6 and 13 are not significantly different between each other, but are significantly different from store 15.

Next, consider the estimated effects of beer promotion on frozen juices sales; see Fig. 1(b). Frozen juices sales are either unresponsive (i.e. non-significant estimates indicated in grey) or respond negatively to an increase in Dominick's beer promotion intensity. This negative effect might be explained by substitutability: an increase in the promotion intensity of beer makes shoppers replace frozen juices by beer. The large households at store 2 (high value of hsizeavg in Table 2) and the low income shoppers at store 11 (low value of income in Table 2) might be most vulnerable to this substitution effect. There is, however, no significant difference between the estimated effect of store 2 and 11.

Finally, consider the estimated effects of beer sales on bottled juices sales; see Fig. 1(c). For all stores, bottled juices sales are unresponsive (i.e. non-significant estimates indicated in grey) to changes in beer sales. Cross-category effects of sales on sales mainly occur due to the budget constraints: if consumers spend more on one category, they might, all else equal, spend less on another because they hit their budget constraint. Such effects are more likely to occur for categories where consumers spend much of their budget, and less likely for categories where they spend less. Since consumers spend, on average, only 14% and 10% of their retail spending (in our data) on respectively beer and bottled juices, this might explain why bottled juices sales are unresponsive to changes in beer sales.

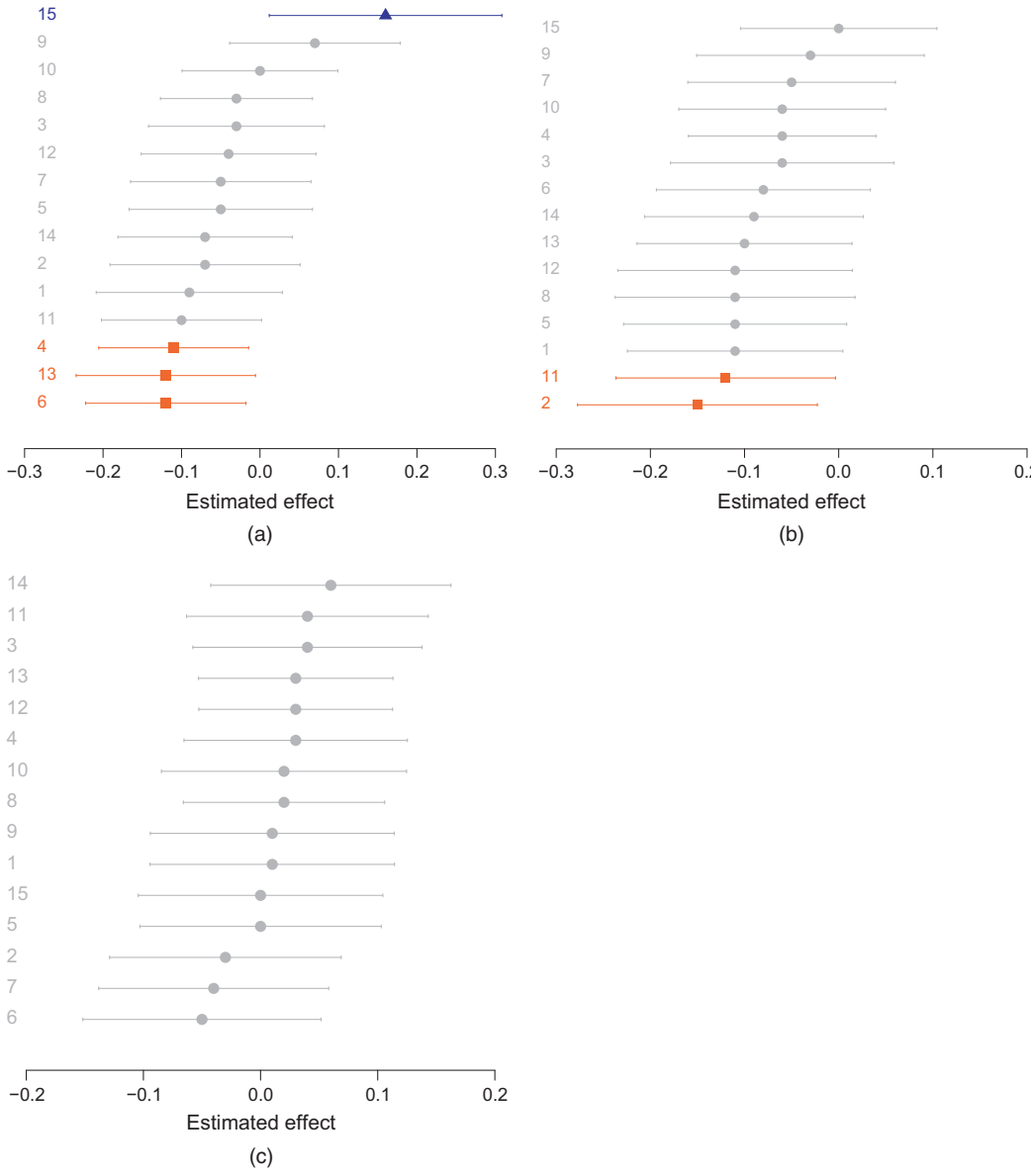


**Fig. 1.** For each store (labelled from 1 to 15), estimated effect (horizontal axis) with the corresponding 95% confidence interval of (a) beer prices on refrigerated juices sales, (b) beer promotion on frozen juices sales and (c) beer sales on bottled juices sales: ■, significant negative estimates; ▲, significant positive estimates; ●, non-significant estimates

In sum, for each estimated cross-category demand effect, the multiclass estimator indicates how the different stores cluster together. Three possible scenarios can occur:

(a) the stores vary in *sign and size* of the estimated effect (first example);
(b) the stores vary only in *size* of the estimated effect (second example);
(c) the stores all have a non-significant estimated effect (third example).

In scenario (a), retailers should set out a store-specific strategy. In scenario (b), a store-specific

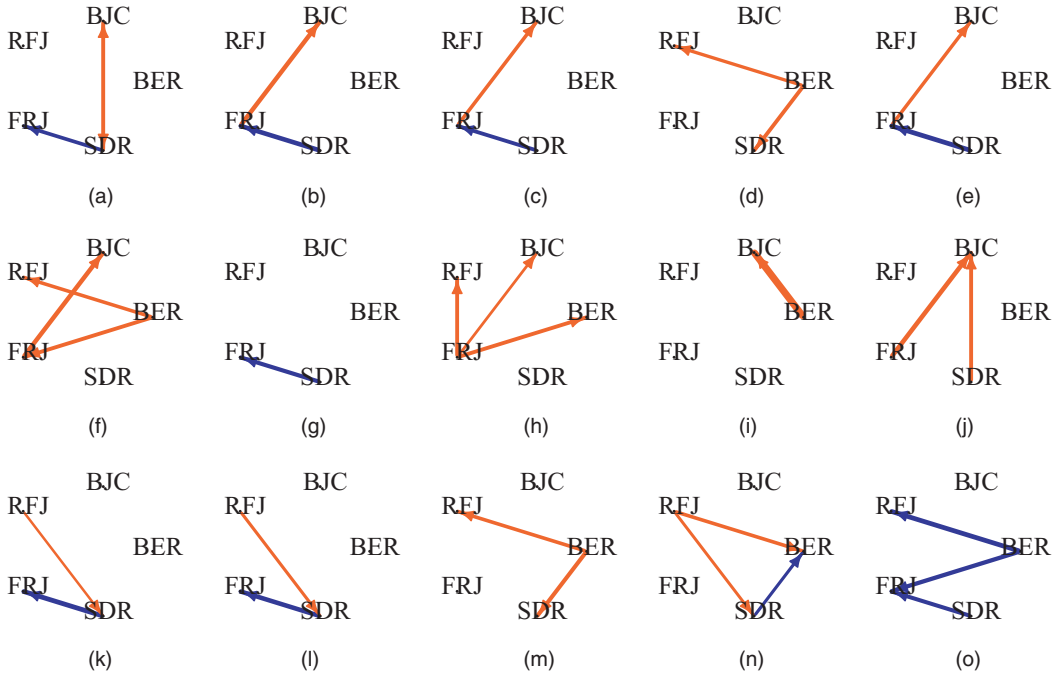


**Fig. 2.** Product category network of prices on sales for each of the 15 stores (a directed edge is drawn from one category to another if its prices influence sales in the other category; the edge width represents the magnitude of the effect; ——, positive effects; ——, negative effects): (a) store 1; (b) store 2; (c) store 3; (d) store 4; (e) store 5; (f) store 6; (g) store 7; (h) store 8; (i) store 9; (j) store 10; (k) store 11; (l) store 12; (m) store 13; (n) store 14; (o) store 15

strategy needs to be set only with respect to the expected degree of responsiveness of each store's market area. Scenario (c) allows retailers to set a chainwide strategy.

### 5.2. *Product category networks*

We use a network analysis to obtain insights into the estimated cross-category demand effects. The product category networks of *prices on sales* are given in Fig. 2. 15 networks are drawn: one for each store. The five product categories are the nodes of the networks. In each network, a directed edge is drawn from one category towards another if the multiclass estimator indicates, by giving an estimate that is statistically significantly different from 0 at the 5% level, that prices in the former category have a direct influence on sales in the latter category. The edge width represents the effect size. Positive effects are shown in blue; negative effects in red. Similar networks can be made for the effects of *promotion on sales* and *sales on sales*. For brevity, we discuss only the network of *prices on sales*.

#### 5.2.1. *Asymmetry of cross-category demand effects*

The cross-category effects of prices on sales are asymmetric. For example, a price increase in soft drinks makes consumers spend more on frozen juices as a compensation (edge from SDR to FRJ for eight stores in Fig. 2), yet a price increase on frozen juices does not affect the soft drinks sales. We typically find categories where consumers spend a large amount of their budget, like soft drinks, to be more influential than responsive: the soft drinks category has more outgoing than incoming edges in Fig. 2. Categories where consumers spend only a small

fraction of their budget, like bottled juices, are more responsive than influential: the bottled juices category has much more incoming than outgoing edges in Fig. 2. Similar conclusions regarding the asymmetry of cross-category effects of promotions on sales and sales on sales can be made. This observed asymmetry is in line with previous research (e.g. Briesch *et al*. (2013)).

An interesting finding concerns soft drinks at the high price tier stores 12–15. For these stores, soft drinks is more responsive to price changes in other categories (0.75 incoming edges per store, on average) than for the other stores (0.27 incoming edges per store, on average). Soft drinks are less frequently consumed by high price tier shoppers and less regularly purchased categories are typically expected to be more responsive to price changes in other categories, as is confirmed by our results.

#### *5.2.2. Drivers of cross-category demand effects*

We find considerably more negative cross-category effects of prices on sales than positive effects (66% *versus* 34%, on average, in Fig. 2). Positive effects might be driven by the substitutability of the products belonging to different categories. Substitution effects occur between products that are perceived by consumers as substitute goods. For instance, a price increase in soft drinks makes consumers purchase frozen juices instead of soft drinks (for eight stores in Fig. 2). The somewhat more surprising negative effects might be explained by either reduced store traffic and/or budget constraints. Price increases might reduce store traffic and, hence, lead towards lower overall sales. This especially holds for shoppers at the cub fighter and low price tier stores given their everyday low price positioning. At these stores, reduced store traffic is thus likely to be the main driver of the negative effects of prices on sales. Furthermore, price increases at one category might also constrain consumers' budget available for other categories, thereby leading towards lower sales of other categories and thus explaining the occurrence of negative cross-category effects of prices on sales.

Store 15 shows very specific cross-category effects of prices on sales: all observed effects are positive (Fig. 2). The market area of store 15 is characterized by a high percentage of blacks and Hispanics and low income, low educated people (high value of ethnic and low values of income and educ in Table 2). These demographic variables are likely to increase price sensitivity, making them more vulnerable to price substitution effects (i.e. positive price effects).

### *5.3. Similarity matrices*

Fig. 3 presents similarity matrices by computing for each pair of stores the proportion of shared within- and cross-category effects of prices on sales (Fig. 3(a)), promotions on sales (Fig. 3(b)) and sales on sales (Fig. 3(c)). We say that there is an effect if the corresponding estimate is statistically significantly different from 0 at the 5% level. Stores sharing a large proportion of effects are put close to each other. For instance, store 13 and store 4 share many prices-on-sales effects, as indicated by the large size and dark colour of the circle in the corresponding cell of Fig. 3(a): 100% of the prices-on-sales effects in store 13 are also present for store 4. In contrast, store 6 and store 4 share only a limited number of prices-on-sales effects, as indicated by the small size and light colour of the circle in the corresponding cell: only 20% of prices-on-sales effects in store 6 are also present for store 4.

The effects of sales on sales and promotions on sales show a considerably higher similarity across stores than the effects of prices on sales. This low similarity of prices-on-sales effects can be explained by Dominick's price tier and market area-specific pricing strategy (Wedel and Zhang, 2004). Since prices at Dominick's stores are set differently according to the price tier type to which they belong (Table 2), prices-on-sales effects are likely to vary considerable between

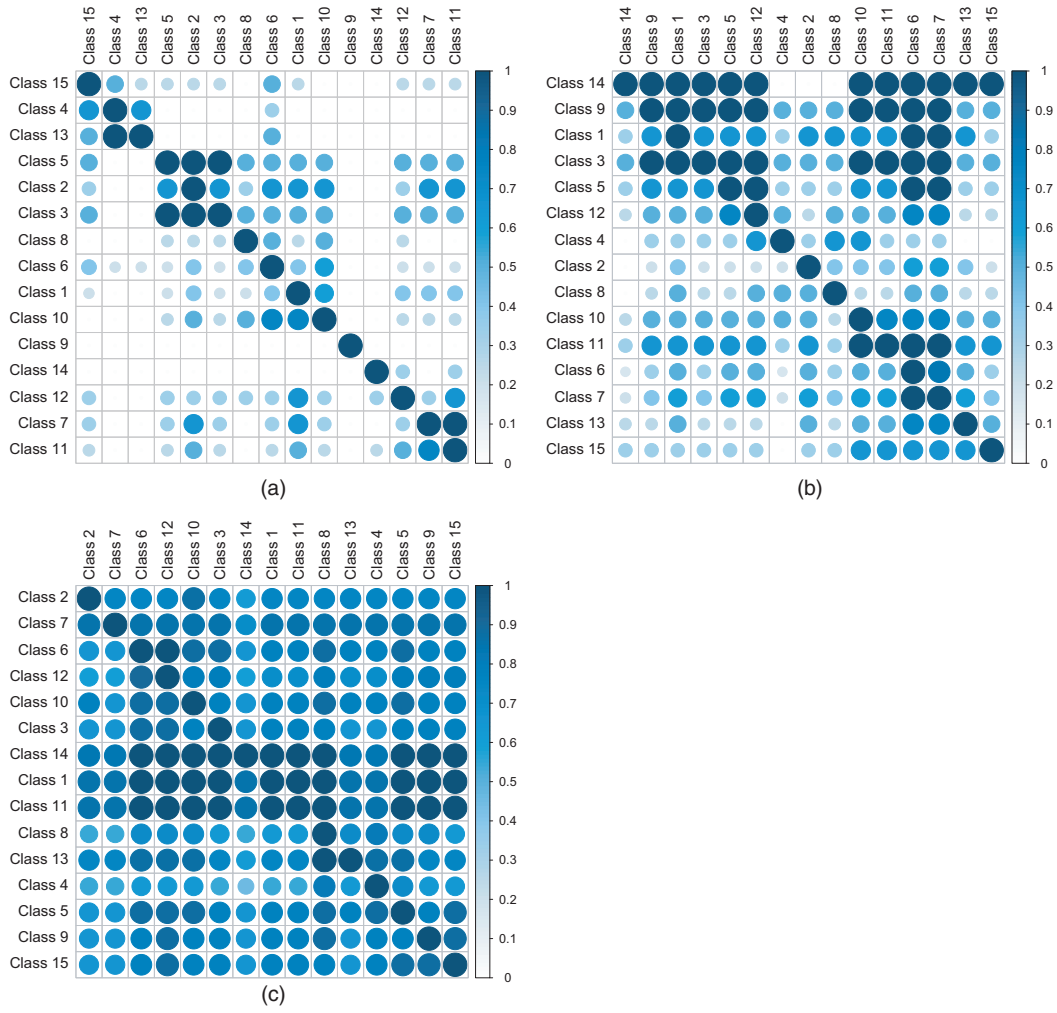


**Fig. 3.** Similarity matrices: each cell indicates the proportion of within- and cross-category effects of (a) prices on sales, (b) promotion on sales and (c) sales on sales for store *i* (row) that are also present for store *j* (column) (the darker and larger the circle, the higher the proportion)

stores. Dominick's promotional strategy, in contrast, is set more uniformly across stores, hence explaining the higher similarity of promotions-on-sales effects between stores.

Looking at the shared prices-on-sales effects of each pair of stores in Fig. 3(a), we find some results that can be explained by common market area demographics. For store 13, for instance, 100% of its prices-on-sales effects are shared with store 4. In terms of geographical proximity, store 13 is close to store 4. Both stores operate in an area that is occupied by small households with large homes (low values of hsizeavg and high values of hvalmean in Table 2).

## 6. Extensions

We discuss three possible extensions of the multiclass VAR approach of Section 2. We apply them to the multistore sales data set. Similar results to those in Section 5 are obtained. Detailed

results are available from the authors on request. The code to obtain these three extensions is available from `http://wileyonlinelibrary.com/journal/rss-datasets`.

### *6.1. Alternative penalty function $P_2(\cdot)$*

Other types of penalty functions $P_2(\cdot)$ such as group penalties (Yuan and Lin, 2006) or bilevel penalties (Huang *et al.*, 2012) can be useful for certain applications. For the multistore sales application, for instance, we have $15 = 3 \times 5$ series for each store. Taking sales, prices and promotion of each product category as a group yields groups of size 3. One could exploit this grouping structure by using an adaptive group lasso penalty (Wang and Leng, 2008) instead of equation (5). The product category networks of prices, promotion and sales on sales are then encouraged to have the same edges for each store.

Our algorithm can be easily extended to other choices of penalty function $P_2(\cdot)$. Changing $P_2(\cdot)$ requires changing only the proximal operator function in the fast iterative shrinkage thresholding algorithm (Section 2.2). All other steps of the algorithm remain unchanged.

### *6.2. Incorporating additional clustering information*

Sometimes, additional class information is available. Consider, for instance, the five demographic variables from Table 2 for the multistore sales application. If such information is available, the regularization parameter $\lambda_1$ can be adjusted.

Instead of using one single $\lambda_1$, an adaptive regularization parameter

$$\lambda_1^{(k,k')} = \frac{\lambda_1 \max_{k'}(d^{(k,k')})}{d^{(k,k')}} \tag{11}$$

can be used for each combination of two classes in equation (4). Here, $d^{(k,k')}$ is a distance measure between class $k$ and $k'$ obtained from the additional clustering information, for instance the Euclidean distances between stores computed from the standardized demographic variables in Table 2. The smaller $d^{(k,k')}$, the more similar both classes are. As a consequence, the regularization parameter $\lambda_1^{(k,k')}$ will be adjusted upwards, thereby encouraging more similarity in the corresponding effects of class $k$ and $k'$. The lower the distance between two classes, the more their corresponding effects are encouraged to be similar.

Incorporating additional clustering information requires adjusting only the $C$-matrix representing the pairs of coefficients that are coupled across classes. Instead of using the same coupling $(\lambda_1, -\lambda_1)$ in equation (9) for each pair of classes, we now use a distinct coupling $(\lambda_1^{(k,k')}, -\lambda_1^{(k,k')})$. All other steps of the algorithm remain unchanged.

### *6.3. Allowing for error correlations across classes*

The same omitted factor may affect the values of the time series for all classes. For the multistore sales application, weather conditions, for instance, may affect the weekly sales in all stores, since they are all in the Chicago area. As a consequence, we might expect the errors to be correlated across the different classes. In Appendix B, we show how to adjust the multiclass VAR estimator for correlated errors across classes. It turns out that about 93% of the partial correlations between error terms are estimated as 0, and the average absolute value of the remaining correlations is also rather low.

## 7. Conclusion

This paper proposes a method for the joint estimation of multiple VAR models corresponding

to distinct but related classes. By this joint estimation, we borrow strength across classes to estimate multiple VAR models that share certain characteristics. Our simulations show that this estimation approach results in a higher estimation accuracy. The multiclass estimator proposed outperforms other estimators that do not encourage corresponding parameters across classes to be estimated identically.

We apply the multiclass VAR model to a multistore sales data set. The shared sales dynamics across stores allow retailers to design a chainwide strategy that reflects the chain's image. The store-specific findings allow retailers to understand how each particular store responds to changes in its marketing mix. We provide visual tools helping to interpret the results of the multiclass VAR model. They show

(a) the store clustering,
(b) the product category networks and
(c) the similarity matrices of cross-category effects among stores.

The product category networks visualize the lagged effects that are captured in the autoregressive coefficient matrix. Alternatively, one could draw the product category networks based on the estimated impulse responses. The impulse response functions give the response of a certain time series to a unitary impulse in the error of another time series as a function of the lag. The network-based analysis can then be extended by looking at, for instance, cumulative impulse responses.

Our multiclass VAR modelling approach is easily applicable to a variety of other settings. In biostatistics, the methodology proposed might be employed to analyse genetic data (Abegaz and Wit, 2013). The time series contain gene expression measurements that are collected over time for a large number of genes. The classes are the treated patients and the controls. The joint estimation could result in a more precise estimation of the gene regulatory networks. In finance, one could study the differences and/or similarities in stock market dynamics among a set of connected financial institutions. The time series are stock market returns; the classes are the different financial institutions (Diebold and Yilmaz, 2015). Another relevant application is the study of the dynamic relationships between different air pollutant levels: in, for instance, Kumar and Prakash (2009) only one geographical area was considered. The multiclass framework could extend this analysis to different areas (i.e. classes) for which the measurements are available.

## Acknowledgements

We thank the Joint Editor, the Associate Editor and referees for their constructive and insightful remarks that substantially improved our paper. We gratefully acknowledge support from Research Foundation Flanders, contract 12M8217N, and from the GOA/12/014 project of the Research Fund, KU Leuven.

## Appendix A: Standard errors for parameter estimates

We proceed as follows.

(a) Estimate model (2) and compute the centred residuals $\hat{\mathbf{e}}_t$ by subtracting $\bar{e}_i^{(k)} = \Sigma_{t=1}^{N} \hat{e}_{t,i}^{(k)}/N$ from the residual of time series $i$ in class $k$ at each time point $t$. Centring was recommended by Chatterjee and Lahiri (2011) and Kreiss and Lahiri (2012).
(b) Let $B = 1000$ be the number of bootstraps. For $b = 1, \ldots, B$ perform the following steps.
    (i) Construct the residual bootstrap time series (e.g. Kreiss and Lahiri (2012)) from model (2) with the parameter estimates from step (a) and with bootstrap errors $\mathbf{e}_t^* = \hat{\mathbf{e}}_{\mathcal{U}_t}$ with $\mathcal{U}_t$, $t = 1, \ldots, N$,

an independently and identically distributed sequence of discrete random variables uniformly distributed on $\{1,\ldots,N\}$.

(ii) Check whether the bootstrap errors are multivariate white noise by using the multivariate portmanteau statistic (Tsay (2014), page 72). If so, continue to the next step; otherwise redraw from the centred residuals. Note that this step is performed for stability, but it could be removed.

(iii) Apply the multiclass estimator of equation (6) to the bootstrap sample. Denote the bootstrap estimate by $\hat{\boldsymbol{\beta}}^{*(b)}$.

(c) For each $i$th element of $\hat{\boldsymbol{\beta}}$, for $1 \leqslant i \leqslant \dim(\hat{\boldsymbol{\beta}})$, compute

$$\widehat{\mathrm{sd}}_i = \left\{ \sum_{b=1}^{B} (\hat{\beta}_i^{*(b)} - \hat{\beta}_i^{*})^2/(B-1) \right\}^{1/2}, \qquad \hat{\beta}_i^{*} = \sum_{b=1}^{B} \hat{\beta}_i^{*(b)}/B.$$

If $|\hat{\beta}_i| > 1.96\,\widehat{\mathrm{sd}}_i$, $\hat{\beta}_i$ is said to be statistically significant (at significance level 5%).

## Appendix B: Multiclass vector auto-regressive estimator allowing for cross-error correlations

To allow for error correlations across different classes, consider

$$(\hat{\boldsymbol{\beta}}, \hat{\bar{\Omega}}) = \arg\min_{\boldsymbol{\beta}, \bar{\Omega}} (\mathbf{y} - X\boldsymbol{\beta})' \bar{\Omega} (\mathbf{y} - X\boldsymbol{\beta}) - \log|\bar{\Omega}| + \lambda_1 P_1(\boldsymbol{\beta}) + \lambda_2 P_2(\boldsymbol{\beta}) + \gamma_1 P_1(\bar{\Omega}) + \gamma_2 P_2(\bar{\Omega}) \tag{12}$$

instead of equation (6). Here, $\mathbf{y}$ is an $NJK$-vector stacking all $J$ time series of all $K$ classes, $X$ is an $NJK \times KJ^2P$ block diagonal matrix with $X^{(1)},\ldots,X^{(K)}$ on the main diagonal and $\bar{\Omega} = \bar{\Omega}_0 \otimes I_N$ with

$$\bar{\Omega}_0 = \begin{pmatrix} \Omega^{(1)} & \ldots & \Omega^{(1,K)} \\ \vdots & \ddots & \vdots \\ \Omega^{(K,1)} & \ldots & \Omega^{(K)} \end{pmatrix}$$

a symmetric $KJ \times KJ$ matrix. The off-diagonal matrices $\Omega^{(k,k')}$ are $J \times J$ matrices representing the partial correlations between the errors of class $k$ and $k'$. As penalty functions, we take

$$P_1(\bar{\Omega}) = \sum_{k<k'}^{K} \sum_{i,j=1}^{J} |\omega_{ij}^{(k,k)} - \omega_{ij}^{(k',k')}|$$

and

$$P_2(\bar{\Omega}) = \sum_{k,k'=1}^{K} \sum_{i,j=1}^{J} |\boldsymbol{\omega}_{ij}^{(k,k')}|,$$

where $\omega_{ij}^{(k,k')}$ is the $ij$th element of $\Omega^{(k,k')}$. As in Section 2.1, $P_1(\bar{\Omega})$ encourages similarity in $\hat{\Omega}^{(1)},\ldots,\hat{\Omega}^{(K)}$. $P_2(\bar{\Omega})$ ensures that estimation remains feasible. The estimator from equation (6) is obtained when taking $\bar{\Omega}_0$ to be block diagonal, i.e. $\Omega^{(k,k')} = \mathbf{0}$ for $k \neq k'$. To approximate the objective function in equation (12), we extend the alternating directions method-of-multipliers algorithm used when solving for $\bar{\Omega}$ conditional on $\boldsymbol{\beta}$, in a similar way to that in Pircalabelu *et al*. (2016).

## References


Abegaz, F. and Wit, E. (2013) Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, **14**, 586–599.

Basu, S., Shojaie, A. and Michailidis, G. (2015) Network Granger causality with inherent grouping structure. *J. Mach. Learn. Res.*, **16**, 417–453.

Beck, A. and Teboulle, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imgng Sci.*, **2**, 183–202.

Briesch, R., Dillon, W. and Fox, E. (2013) Category positioning and store choice: the role of destination categories. *Marktng Sci.*, **32**, 488–509.

Chatterjee, A. and Lahiri, S. (2011) Bootstrapping lasso estimators. *J. Am. Statist. Ass.*, **106**, 608–625.

Chen, X., Lin, Q., Kim, S., Carbonell, J. and Xing, E. (2012) Smoothing proximal gradient method for general structured sparse regression. *Ann. Appl. Statist.*, **6**, 719–752.
Danaher, P., Wang, P. and Witten, D. M. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc.* B, **76**, 373–397.
Davis, R., Zang, P. and Zheng, T. (2016) Sparse vector autoregressive modeling. *J. Computnl Graph. Statist.*, **25**, 1077–1096.
Dekimpe, M. and Hanssens, D. (1995) The persistence of marketing effects on sales. *Marktng Sci.*, **14**, 1–21.
Diebold, F. and Yilmaz, K. (2015) *Financial and Macroeconomics Connectedness: a Network Approach to Measurement and Monitoring*. Oxford: Oxford University Press.
Gelper, S., Wilms, I. and Croux, C. (2016) Identifying demand effects in a large network of product categories. *J. Retail.*, **92**, 25–39.
Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.
Hsu, N.-J., Hung, H.-L. and Chang, Y.-M. (2008) Subset selection for vector autoregressive processes using lasso. *Computnl Statist. Data Anal.*, **52**, 3645–3657.
Huang, J., Breheny, P. and Ma, S. (2012) A selective review of group selection in high-dimensional models. *Statist. Sci.*, **27**, 481–499.
Hurvich, C. M. and Tsai, C. L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
Kamkura, W. and Kang, W. (2007) Chain-wide and store-level analysis for cross-category management. *J. Retail.*, **83**, 159–170.
Kim, S. and Xing, E. (2009) Statistical estimation of correlated genome associations to a quantitative trait network. *PLOS Genet.*, **5**, no. 8, article e1000587.
Kreiss, J. and Lahiri, S. (2012) Bootstrap methods for time series. In *Time Series Analysis: Methods and Applications* (eds T. Rao, S. Rao and C. Rao). Amsterdam: North-Holland.
Kumar, U. and Prakash, A. (2009) A multivariate time series approach to study the interdependence among $O_3$, $NO_x$ and VOCs in ambient urban atmosphere. *Environ. Modlng Assessmnt*, **29**, 631–643.
Lang, S., Steiner, W., Weber, A. and Wechselberger, P. (2015) Accommodating heterogeneity and nonlinearity in price effects for predicting brand sales and profits. *Eur. J. Oper. Res.*, **246**, 232–241.
Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016) Exact post-selection inference, with application to the lasso. *Ann. Statist.*, **44**, 907–927.
Leeflang, P. and Selva, J. (2012) Cross-demand effects of price promotions. *J. Acad. Marktng Sci.*, **40**, 572–586.
Milligan, G. W. and Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.
Nesterov, Y. (2005) Smooth minimization of non-smooth functions. *Math. Programm.* A, **103**, 127–152.
Pircalabelu, E., Claeskens, G. and Waldorp, L. J. (2016) Mixed scale joint graphical lasso. *Biostatistics*, **17**, 793–806.
She, Y. (2010) Sparse regression with exact clustering. *Electron. J. Statist.*, **4**, 1055–1096.
Song, I. and Chintagunta, P. (2006) Measuring cross-category price effects with aggregate store data. *Mangmnt Sci.*, **52**, 1594–1609.
Srinivasan, S., Pauwels, K., Hanssens, D. and Dekimpe, M. (2004) Do promotions benefit manufacturers, retailers, or both? *Mangmnt Sci.,* **50,** 617–629.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc.* B, **67**, 91–108.
Tsay, R. (2014) *Multivariate Time Series Analysis: With R and Financial Applications*. Hoboken: Wiley.
Wang, H. and Leng, C. (2008) A note on adaptive group lasso. *Computnl Statist. Data Anal.*, **52**, 5277–5286.
Wedel, M. and Zhang, J. (2004) Analyzing brand competition across subcategories. *Marktng Sci.*, **41**, 448–456.
Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.
Yuan, M. and Lin, Y. (2007) Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35.
Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.

*Supporting information*

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary file "Multi-class vector autoregressive models for multi-store sales data"'.