European Journal of Operational Research 267 (2018) 96–106

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing and Logistics

The multivariate bullwhip effect[☆]

Chaitra H. Nagaraja[a,*], Tucker McElroy[b]

[a] Strategy and Statistics Area, Gabelli School of Business, Fordham University, Martino Hall, 45 Columbus Ave., RM 516, New York, NY 10023, USA
[b] Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA

ARTICLE INFO

Article history:
Received 11 November 2016
Accepted 8 November 2017
Available online 16 November 2017

Keywords:
Supply chain management
Multivariate time series
Cross-correlated demand

ABSTRACT

A multivariate bullwhip expression for *m* products with an order-up-to inventory policy is developed. The demand models under consideration are differenced stationary vector time series with a Wold representation for which general forecasting formulas are available, resulting in a large class of possible models (including nonstationary ones). Examples are provided for common demand models and implemented on sales data. It is found that the multivariate approach gives rise to mechanisms for understanding and reducing the bullwhip effect through horizontal information sharing, particularly for the nonstationary demand case. In the stationary setting, a more nuanced approach to bullwhip reduction can be achieved by managing the relationship between cross-correlations and lead-times. A method of determining whether a multivariate or univariate approach generates a lower bullwhip effect is proposed.

1. Introduction

The bullwhip effect occurs when the variability of orders increases as they travel up a supply chain. This volatility leads to difficulty in managing inventory and consequently is an undesired, but often observed, phenomenon (Lee, Padmanabhan, & Whang, 1997a,b). We will examine this effect from the perspective of a retailer who employs an order-up-to inventory policy, placing orders to a supplier based on observed customer demand.

Sharing demand information across stages of a supply chain – in effect, *vertical* information sharing – is a common recommendation for reducing the bullwhip effect when demand is stationary (Chen & Lee, 2009; Gill & Abend, 1997; Lee, So, & Tang, 2000). However, it requires collaboration across firms which may be difficult to achieve in practice. Challenges include timing of information (Xu, Dong, & Xia, 2015), quality of information (Kwak & Gavirneni, 2015; Mishra, Raghunathan, & Yue, 2007), and trust across firms (Kong, Rajagopalan, & Zhang, 2013; Voigt & Inderfurth, 2012). Moreover, Graves (1999) showed that for ARIMA(0,1,1) demand, a nonstationary demand model, vertical information sharing does not reduce the bullwhip effect.

We propose an alternate, practical solution: *horizontal* information sharing. By harnessing demand information from the sales of other products by fitting multivariate demand models – information that retailers already have available to them – retailers possess a bullwhip reduction mechanism which (a) does not rely on coordinating with other firms, (b) improves demand forecasts, and (c) can be applied to products with nonstationary demand where vertical information sharing is not beneficial.

In the supply chain literature, Chen and Blue (2010) provide empirical evidence for the utility of multivariate demand models in manufacturing and Boute, Dejonckheere, Disney, and Van de Velde (2012) do the same in the chocolate industry. Both use VAR(1) models. We provide further examples – stationary and nonstationary – in the paper products industry (e.g., paper towels) here.

To illustrate the concept of horizontal information sharing, we develop a multivariate demand framework from the retailer perspective and derive a generalized bullwhip measure with three key components. First, we allow for a flexible supply chain structure with *m* series. Our approach can be applied when *m* products are bought from a single suppler (i.e., two-stage supply chain); the same product is modeled separately for *m* retail branches; or even when *m* products are purchased from *m* suppliers with a common lead-time (i.e, *m* two-stage supply chains). Second, we consider a comprehensive class of demand models: those that can be represented, once the data is appropriately differenced, as a multivariate infinite order moving average. This includes Vector Autoregressive (VAR), Vector Moving Average (VMA), seasonal, and integrated (i.e., nonstationary models) demand processes. Finally, we develop a multivariate bullwhip measure which can be applied to all of these demand models, stationary and nonstationary alike.

☆ **Disclaimer:** This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the author and not necessarily those of the U.S. Census Bureau.

* Corresponding author.
E-mail addresses: cnagaraja@fordham.edu (C.H. Nagaraja), tucker.s.mcelroy@census.gov (T. McElroy).

https://doi.org/10.1016/j.ejor.2017.11.015
0377-2217/

We then use our framework to analyze the complex relationship between multivariate demand and the bullwhip effect. These technical insights lead to bullwhip control mechanisms including superior demand forecasts and lead-time adjustments. We also provide a method to determine whether the univariate or multivariate approach generates a lower bullwhip effect for a given stationary demand model.

Our paper is organized as follows. We begin with a literature review in Section 2. In Section 3, we develop our multivariate bullwhip measure and discuss empirical and parametric estimation. In Section 4, we study the technical properties of our measure. We then apply it to four standard demand models in Section 5 and to VAR(3) and integrated VAR(5) data from the Dominick's Database (Kilts Center for Marketing, University of Chicago) in Section 6. We conclude with a discussion in Section 7. Further details regarding the data and our R code are provided in the supplementary files, "Data and demand model summaries" and "multivariate_bullwhip_code.R".

2. Literature review

Maintaining proper control of inventories is difficult when the bullwhip effect is present. Consequently, this phenomenon has been studied mathematically for a variety of settings, with a demonstration in each case of the impact of the underlying dynamics of the demand process, as well as the effect of the lead-time horizon (Disney, Maltz, Wang, & Warburton, 2016; Gaalman, 2006; Ma & Wu, 2014). Wang and Disney (2016) provide an excellent overview.

The primary recommendation to reduce the univariate bullwhip effect is to share demand information across stages of the supply chain. This can be thought of as vertical information sharing. Wal-Mart's Retail Link system is an example of such an approach (Gill & Abend, 1997). Lee et al. (2000) investigate this idea for a two-stage supply chain with AR(1) demand, whereas Chen, Drezner, Ryan, and Simchi-Levi (2000a) consider multi-stage chains as well. As highlighted in Section 1, information sharing can be difficult to implement in practice and so alternative bullwhip reduction methods must be explored. Furthermore, Graves (1999) noted that in the nonstationary case, vertical information sharing is not useful to control the bullwhip effect.

A second, more technical, mechanism to manage this phenomenon is to improve modeling and estimation methods. Gilbert (2005) provides structure to this discussion by categorizing changes in demand forecasts from one inventory review period to the next: Type 1 corresponds to actual changes in demand; Type 2 refers to errors in forecasts. Both can cause the bullwhip effect. Type 2 is especially damaging as it can needlessly produce this unwanted phenomenon.

We can further split Type 2 errors into two categories: choice of model and choice of forecasting method. Better models lead to better forecasts, allowing for improved bullwhip effect management and a reduction in Type 2 issues. Taking a multivariate approach, as we do in this paper, helps address the former variety of Type 2 challenges. For the latter case, Chen et al. (2000a); Chen, Ryan, and Simchi-Levi (2000b); Zhang (2004) all find that the choice of forecasting method can increase the bullwhip effect magnitude. Zhang (2004, p. 25), however, advocates the use of minimum mean-squared error forecasts for demand, provided the demand model is relatively stable over time. We apply Zhang's advice to the multivariate case here.

Recently, there has been a shift in focus from univariate to multivariate demand. There are four main themes in the literature: (a) structure of the demand model, (b) the number of products and structure of the supply chain, (c) the formula for the bullwhip measure, and (d) whether a manufacturing or retail perspective is taken. This last point requires some additional explanation. Retail sales are of completed products, needing no explicit combination of raw materials, whereas manufacturing processes require maintaining inventory levels of component materials to produce the final products. Therefore, the insights generated from one perspective are not necessarily applicable to the other.

Chaharsooghi and Sadeghi (2008) begin work on this topic by examining the stationary VAR(1) demand for a two-stage supply chain with two products. They extend the univariate bullwhip measure to the multivariate case, discarding the off-diagonal terms of the variance matrices. Variants on the VAR(1) case include Zhang and Burke (2011) who incorporate price information into their model. They also propose a bullwhip measure that takes the difference between the quantity ordered covariance matrix and the demand covariance matrix instead of the ratio. Boute et al. (2012) take a manufacturing perspective and examine a more limited VAR(1) setting where the covariance matrix of the error term is diagonal (i.e., correlation across error series is not permitted). They focus on the production side of a supply chain and the restriction on the error term allows them to frame their insights in terms of transfer functions (i.e., Granger causality). While they study the two-product case, they note that their work can apply to a "triadic" structure; that is, one manufacturer with two retailers. Raghunathan, Tang, and Yue (2017) also take the manufacturer's perspective and allow for more than two products; however, they require the autoregressive coefficient matrix in the VAR(1) model to have a restrictive, symmetric structure. They examine both the ratio and differencing bullwhip measures and their restrictions on the demand model allow them to examine the cross-diagonal terms of the bullwhip matrix and the effect of the number of products modeled simultaneously on the bullwhip magnitude. We present a general way to interpret the cross-diagonal terms here.

Finally, Bray and Mendelson (2015) build on Chen and Lee (2009), who expressed univariate demand as a martingale model of forecast evolution using a manufacturing perspective. They skip specifying a demand structure – thereby allowing for a wide range of possibilities – and focus on examining the bullwhip effect based on demand forecasts only. Therefore, they do not explicitly discuss the impact of demand on the bullwhip effect, which is one of our goals here. Furthermore, their method generates a bullwhip formula which is applicable only to the stationary demand model case.

As noted in the introduction, we take a more general approach and place few restrictions on the structure of the demand model, encompassing both stationary and nonstationary demand. We find in our analysis of products in the Dominick's Database that taking such an approach is more appropriate for modeling retail demand. Our unifying framework also generalizes the "triadic structure" as defined by Boute et al. (2012) to m products with a range of possible supply chain settings. Furthermore, we extend Zhang and Burke's (2011) definition of the bullwhip effect because, as we will show, it is flexible enough to integrate both stationary and nonstationary demand and allows us to interpret the off-diagonal elements of the bullwhip matrix. Finally, as we take a retailer's perspective, our insights are focused on understanding the relationship between multivariate demand and bullwhip magnitude, which can then be used to manage the magnitude of the bullwhip effect.

3. Methods

We introduce a general class of demand models and formulas for calculating the quantity ordered series, which are then used to derive the multivariate bullwhip measure. We conclude by discussing estimation methods.

3.1. Demand models

Let $\{\mathbf{d}_t\}$ denote a multivariate demand series corresponding to m products, with demand observed from the retailer's perspective. We assume that $\{\mathbf{d}_t\}$ can be expressed as a differenced stationary vector time series with a causal moving average representation for which general forecasting formulas are available (McElroy & McCracken, 2015). This general structure includes VAR models, co-integrated structural time series models, SARIMA models, etc.

This broad class of multivariate time series has a degree ℓ differencing matrix polynomial $\boldsymbol{\delta}(B) = \sum_{j=0}^{\ell} \boldsymbol{\delta}_j B^j$ where the coefficients $\boldsymbol{\delta}_j$ are matrices, B is the backshift operator (i.e., $B^j \mathbf{d}_t = \mathbf{d}_{t-j}$), and $\boldsymbol{\delta}_0$ is the m-dimensional identity matrix, denoted as $\mathbf{1}_m$. The presumption is that $\det \boldsymbol{\delta}$, where det denotes the determinant of a matrix, has some (or all) of its roots on the unit circle of the complex plane, thereby corresponding to a nonstationary effect in the data. (We use the $\boldsymbol{\delta}(B)$ notation, in part, so that the multivariate bullwhip expression is clearer, as it requires $\boldsymbol{\delta}^{-1}(B)$; see Sections 5.3–5.5 for examples.)

The polynomial $\boldsymbol{\delta}$ is typically identified by exploratory analysis together with unit root testing. Its action on the data reduces it to strict stationarity, and the result $\partial \mathbf{d}_t = \boldsymbol{\delta}(B)\mathbf{d}_t$ (i.e., the differenced, and now stationary, series) can be expressed with a multivariate infinite order moving average representation:

$$\partial \mathbf{d}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}(B)\boldsymbol{\varepsilon}_t, \tag{1}$$

which is also referred to as the Wold representation. (See Lütkephol, 2005 for more details.) (Note that if the demand series is already stationary without any differencing, you can ignore $\boldsymbol{\delta}(B)$.) In (1), the mean vector is $\boldsymbol{\mu}$, $\boldsymbol{\Psi}(B) = \sum_{j\geq 0} \boldsymbol{\Psi}_j B^j$ is a causal matrix power series, and $\{\boldsymbol{\varepsilon}_t\}$ is a multivariate white noise series with variance matrix $\boldsymbol{\Sigma}$. In our examples in Section 6, we will assume that $\boldsymbol{\varepsilon}_t$ has a multivariate normal distribution.

Differencing converts the nonstationary series into a stationary one. However, that step does not imply that the subsequent analysis is equivalent for both cases. Rather, the bullwhip effect is a study of the variability in a supply chain, and the concept of variability is profoundly different for stationary versus nonstationary series. Consequently, the treatment and interpretation of the bullwhip measure differ in nontrivial ways. Forecast errors are computed using the (original) nonstationary series (Section 3.3); adjustments are made to the bullwhip derivation to incorporate the evolving variability of nonstationary series (Section 3.3); and, estimation differs as well (Sections 3.4 and 6). These are some examples of how stationary and nonstationary demand series are handled differently, even though we can express our results using a general framework.

3.2. Supply chain

To describe the supply chain, we adapt the notation from Zhang (2004) to handle m products sold by a retailer who employs an order-up-to inventory policy. Each product – if considered separately from the others – is assumed to be part of a two-stage supply chain. This covers a wide range of possible configurations. Some examples include: m products bought from a single suppler (i.e., two-stage supply chain); m products bought from j different suppliers where $j \leq m$ (see Section 6.1); or the same product modeled separately for m retail branches (see Section 6.2).

We assume that the lead-time horizon, L, is fixed and common to all m products and is defined as the length of time between placing and receiving an order. In a given inventory review period t, the sequence of events is as follows: the order made at time $t-L$ arrives, demand is observed, and a new order is placed for all m products to be delivered at time $t+L$.

Assume that demand $\{\mathbf{d}_t\}$ can be represented by the class of models outlined in Section 3.1. This class includes both stationary and nonstationary models. Gilbert (2005) uses the order-up-to inventory policy for both settings in the univariate case when he derives a bullwhip measure for ARIMA(p, d, q) models, noting that the policy is typical for Material Requirements Planning algorithms (p. 306); his argument applies to the multivariate case as well.

In the multivariate setting, current demand for any one product is allowed to depend upon its past demand, as well as the past demand of other products. This history of demand is denoted as $\mathbf{H}_t$ and is defined as the σ-algebra generated by $\{\mathbf{d}_s;\ s \leq t\}$ (e.g., realized demand from all m products up to and including time t).

The quantity ordered of the m products at time t is denoted as $\mathbf{q}_t$. It is equal to the difference between the order-up-to level between time periods t and $t-1$ plus the realized demand at time t. Let the aggregate demand that accrues during the lead-time period, as we wait for the order placed at time t to arrive at time $t+L$, be denoted as $\mathbf{D}_t^L = \sum_{s=1}^{L} \mathbf{d}_{t+s}$. Then, as the inventory position includes the safety stock – i.e., a buffer quantity of each product – assumed to be the same across all time periods, a simplified expression for $\mathbf{q}_t$ is given by

$$\mathbf{q}_t = \mathbf{D}_{t|\mathbf{H}_t}^L - \mathbf{D}_{t-1|\mathbf{H}_{t-1}}^L + \mathbf{d}_t, \tag{2}$$

where $\mathbf{D}_{t|\mathbf{H}_t}^L$ denotes the forecasted aggregate demand (based upon demand history and some model of the demand process). We observe that $\mathbf{q}_t \in \mathbf{H}_t$. That is, $\mathbf{q}_t$ is measurable with respect to the σ-algebra $\mathbf{H}_t$, so that, in particular, $\mathbb{E}[\mathbf{q}_t|\mathbf{H}_t] = \mathbf{q}_t$ (where $\mathbb{E}[\cdot]$ denotes expected value). It is possible that individual orders can be negative. As in Chen and Lee (2009), we interpret this to mean the retailer can return extra stock without penalty. One can also incorporate holding, shortage penalty, and ordering costs through calculating the safety stock level to obtain the optimal base-stock level (see proof of Theorem 1 on p. 550 in Lee et al., 1997a); however, they cancel out in the simplified version of (2) since our focus here is on the quantity ordered series.

In this paper, we utilize linear forecasting formulas, so that in order to forecast aggregate lead-time demand it is sufficient to forecast future demand at each time period. If we let $\mathbf{d}_{t+s|\mathbf{H}_t}$ denote the s-step ahead forecast of demand computed at time t (with realized demand up to time t), then

$$\mathbf{D}_{t|\mathbf{H}_t}^L = \sum_{s=1}^{L} \mathbf{d}_{t+s|\mathbf{H}_t}. \tag{3}$$

3.3. Multivariate bullwhip measure

The presence of the bullwhip effect indicates that the variability of the quantity ordered outpaces the variability of demand; hence, we must be able to write down a relationship of variances to characterize that phenomenon. The expression in (2) is convenient for logistics, but an equivalent form written in terms of forecast errors is more convenient for statistical analysis. It will allow us to more easily discern the relevant decomposition of variances.

Let $\mathbf{E}_{t|\mathbf{H}_t}^L = \mathbf{D}_{t|\mathbf{H}_t}^L - \mathbf{D}_t^L$ be the aggregate demand forecast error. Then, it follows from (3) that $\mathbf{D}_{t|\mathbf{H}_t}^L - \mathbf{D}_{t-1|\mathbf{H}_{t-1}}^L = \left(\mathbf{E}_{t|\mathbf{H}_t}^L - \mathbf{E}_{t-1|\mathbf{H}_{t-1}}^L\right) + (\mathbf{d}_{t+L} - \mathbf{d}_t)$, and therefore quantity ordered can be rewritten as:

$$\mathbf{q}_t = \left(\mathbf{E}_{t|\mathbf{H}_t}^L - \mathbf{E}_{t-1|\mathbf{H}_{t-1}}^L\right) + \mathbf{d}_{t+L}. \tag{4}$$

When the forecasting model is correctly specified, the forecasting formulas are conditional expectations of future values and the forecast errors are orthogonal (i.e., uncorrelated) to the data used for prediction. That is, $\mathbf{E}_{t|\mathbf{H}_t}^L$ is orthogonal to all variables in $\mathbf{H}_t$. Since $\mathbf{q}_t \in \mathbf{H}_t$, it follows that $\mathbf{q}_t$ is orthogonal to $\mathbf{E}_{t|\mathbf{H}_t}^L$, although it need not be orthogonal to $\mathbf{E}_{t-1|\mathbf{H}_{t-1}}^L$.

Our next step is to derive operational formulas for the forecast error. For any matrix power series $\mathbf{A}(B) = \sum_{j\leq 0} \mathbf{A}_j B^j$ (where B is the backshift operator and $\mathbf{A}$ is a matrix), we utilize the notation $[\mathbf{A}(B)]_h^k = \sum_{j=h}^k \mathbf{A}_j B^j$ where $\mathbf{A}_j$ is simply the jth coefficient of $\mathbf{A}(B)$. We also assume that the demand model has a differencing operator $\boldsymbol{\delta}(B)$ such that $\{\partial \mathbf{d}_t\}$ satisfies (1).

The power series $\boldsymbol{\delta}^{-1}(B)$ is explosive, but can be computed to a finite number of terms. Therefore, any coefficient of $\boldsymbol{\delta}^{-1}(B)\boldsymbol{\Psi}(B)$ can be determined; see McElroy and McCracken (2015) for more explanation. Letting $\boldsymbol{\Xi}(B) = \boldsymbol{\delta}^{-1}(B)\boldsymbol{\Psi}(B)$, the forecast error equation is

$$\mathbf{d}_{t+s|\mathbf{H}_t} - \mathbf{d}_{t+s} = -[\boldsymbol{\Xi}(B)]_0^{s-1} \boldsymbol{\Psi}^{-1}(B)\,(\partial \mathbf{d}_{t+s} - \boldsymbol{\mu}) \tag{5}$$

for $s \geq 1$. We can rewrite this formula as follows:

$$\mathbf{d}_{t+s|\mathbf{H}_t} - \mathbf{d}_{t+s} = -[\boldsymbol{\Xi}(B)]_0^{s-1} \boldsymbol{\varepsilon}_{t+s}, \tag{6}$$

which only depends on $\boldsymbol{\varepsilon}_{t+1}, \ldots, \boldsymbol{\varepsilon}_{t+s}$ and hence is orthogonal to $\mathbf{H}_t$. Although (5) gives the general formula for forecast error from a fitted model, (6) is only accurate when the model is correctly specified (and the parameter values are correct as well). We presume this setting holds here as the treatment of misspecified models in the bullwhip context (e.g., fitting the wrong demand model) is beyond the scope of this paper.

Using (6), the forecast errors for aggregate demand are $\mathbf{E}^L_{t|\mathbf{H}_t} = -\sum_{s=1}^L [\boldsymbol{\Xi}(B)]_0^{s-1} \boldsymbol{\varepsilon}_{t+s}$, which we can substitute into the difference of forecasted aggregate demand to yield (recall $\boldsymbol{\Xi}_s$ is the sth coefficient of $\boldsymbol{\Xi}(B)$)

$$\begin{aligned}
\mathbf{E}^L_{t|\mathbf{H}_t} - \mathbf{E}^L_{t-1|\mathbf{H}_{t-1}} &= -\sum_{s=1}^{L} [\boldsymbol{\Xi}(B)]_0^{s-1} (\boldsymbol{\varepsilon}_{t+s} - \boldsymbol{\varepsilon}_{t+s-1}) \\
&= -\sum_{s=1}^{L} [\boldsymbol{\Xi}(B)]_0^{s-1} \boldsymbol{\varepsilon}_{t+s} + \sum_{s=0}^{L-1} [\boldsymbol{\Xi}(B)]_0^{s-1} \boldsymbol{\varepsilon}_{t+s} \\
&= -[\boldsymbol{\Xi}(B)]_0^{L-1} \boldsymbol{\varepsilon}_{t+L} + \left(\sum_{s=0}^{L-1} \boldsymbol{\Xi}_s \right) \boldsymbol{\varepsilon}_t.
\end{aligned} \tag{7}$$

Now, if we substitute (7) into (4), we obtain

$$\mathbf{q}_t = \mathbf{d}_{t+L} + \left(\sum_{s=0}^{L-1} \boldsymbol{\Xi}_s \right) \boldsymbol{\varepsilon}_t - [\boldsymbol{\Xi}(B)]_0^{L-1} \boldsymbol{\varepsilon}_{t+L}. \tag{8}$$

Prior literature has focused on the stationary case, where one uses the ratio between the variances of $\mathbf{q}_t$ and $\mathbf{d}_t$, rather than $\mathbf{d}_{t+L}$ to compute the bullwhip measure. However, these variances are identical for stationary processes (i.e., $\mathbb{V}ar[\mathbf{d}_t] = \mathbb{V}ar[\mathbf{d}_{t+L}] = \sum_{s\geq 0} \boldsymbol{\Psi}_s \boldsymbol{\Sigma} \boldsymbol{\Psi}_s'$). In the nonstationary case, our analysis demonstrates that it is more appropriate to consider $\mathbf{d}_{t+L}$.

We can recast the expression for $\mathbf{q}_t$ by first adding and subtracting $\boldsymbol{\Xi}_L \boldsymbol{\varepsilon}_t$ to (8). Then, noting that $[\boldsymbol{\Xi}(B)]_0^L \boldsymbol{\varepsilon}_{t+L}$ equals $\mathbf{d}_{t+L} - \mathbf{d}_{t+L|\mathbf{H}_{t-1}}$, we can write

$$\mathbf{q}_t = \mathbf{d}_{t+L|\mathbf{H}_{t-1}} + \sum_{s=0}^{L} \boldsymbol{\Xi}_s \boldsymbol{\varepsilon}_t, \tag{9}$$

where the two summands are orthogonal. Therefore, the variance of $\mathbf{q}_t$ equals the sum of the component variances; furthermore, because of the orthogonal decomposition of $\mathbf{d}_{t+L}$ into the sum of $\mathbf{d}_{t+L|\mathbf{H}_{t-1}}$ and the forecast error $\mathbf{d}_{t+L} - \mathbf{d}_{t+L|\mathbf{H}_{t-1}}$, we have

$$\mathbb{V}ar[\mathbf{d}_{t+L}] = \mathbb{V}ar[\mathbf{d}_{t+L|\mathbf{H}_{t-1}}] + \sum_{s=0}^{L} \boldsymbol{\Xi}_s \boldsymbol{\Sigma} \boldsymbol{\Xi}_s'. \tag{10}$$

Using the fact that the summands of (9) are orthogonal along with (10), we find

$$\mathbb{V}ar[\mathbf{q}_t] = \mathbb{V}ar[\mathbf{d}_{t+L}] + \sum_{s=0}^{L} \boldsymbol{\Xi}_s \boldsymbol{\Sigma} \sum_{t=0}^{L} \boldsymbol{\Xi}_t' - \sum_{s=0}^{L} \boldsymbol{\Xi}_s \boldsymbol{\Sigma} \boldsymbol{\Xi}_s'. \tag{11}$$

In the univariate case, it is natural to define the bullwhip measure $\mathcal{B}$ from (11) as the ratio of order variability $\mathbb{V}ar[\mathbf{q}_t]$ to the future demand variability $\mathbb{V}ar[\mathbf{d}_{t+L}]$, with a bullwhip effect being present whenever this ratio exceeds unity. We need to generalize this concept to the multivariate context. Accordingly, the multivariate bullwhip is defined as:

$$\mathcal{B} = \mathbb{V}ar[\mathbf{q}_t] - \mathbb{V}ar[\mathbf{d}_{t+L}] = \sum_{s=0}^{L} \boldsymbol{\Xi}_s \boldsymbol{\Sigma} \sum_{t=0}^{L} \boldsymbol{\Xi}_t' - \sum_{s=0}^{L} \boldsymbol{\Xi}_s \boldsymbol{\Sigma} \boldsymbol{\Xi}_s'. \tag{12}$$

Clearly $\mathcal{B}$ is a matrix, and we reference individual entries jk via $\mathcal{B}_{jk}$. The diagonal entries are of particular interest, because $\mathcal{B}_{kk}$ pertains to the kth product or retail branch – a positive bullwhip effect is present if $\mathcal{B}_{kk} > 0$. Equivalently, using (11) and (12), $\mathcal{B}_{kk} > 0$ if and only if

$$\left(\sum_{s=0}^{L} \boldsymbol{\Xi}_s \boldsymbol{\Sigma} \sum_{t=0}^{L} \boldsymbol{\Xi}_t' \right)_{kk} > \left(\sum_{s=0}^{L} \boldsymbol{\Xi}_s \boldsymbol{\Sigma} \boldsymbol{\Xi}_s' \right)_{kk}. \tag{13}$$

Similarly, a negative bullwhip effect would be present for product k when $\mathcal{B}_{kk} < 0$, requiring the reversal of the inequality sign in (13).

Zhang and Burke (2011) developed a bullwhip measure from the perspective of a supplier handling orders from two retailers where demand follows a modified VAR(1) model. They also extract the diagonal elements of the bullwhip matrix. However, their compound bullwhip measure only applies if both demand streams are for the same product, a constraint we are not requiring here.

3.4. Estimation

The multivariate bullwhip effect is straightforward to estimate for stationary models, as $\mathbb{V}ar[\mathbf{d}_t] = \mathbb{V}ar[\mathbf{d}_{t+L}]$. Only the observed demand and order quantity series are required; the lead-time and a demand model are not. The covariance matrices for both series can be computed directly from the observed values and the difference between the two is the estimated multivariate bullwhip effect. This is an empirical estimate of $\mathcal{B}$.

A second approach is to model the observed demand series and, when combined with the lead-time L, estimate the bullwhip effect using (12). This is a parametric estimate of $\mathcal{B}$. The advantage of such estimates is that it allows for an understanding of why the bullwhip effect occurs in a given case – allowing one to construct a plan to reduce the effect – as opposed to simply knowing that the effect exists, which is all one can elicit from the empirical approach.

For nonstationary demand, however, the variability of the demand process changes with time (i.e., $\mathbb{V}ar[\mathbf{d}_t] \neq \mathbb{V}ar[\mathbf{d}_{t+L}]$). Therefore, simply computing the covariance matrix of the differenced demand series and proceeding as in the stationary case is not appropriate. Furthermore, knowledge of just the demand forecasts is not sufficient either as the order quantity series is also nonstationary. (Graves, 1999 showed this for the ARIMA(0,1,1) case.) Consequently, the parametric approach is the only option for the nonstationary case because the fitted model parameters are required.

4. Properties of the multivariate bullwhip measure

Having derived our bullwhip measure, we now turn to exploring its properties, primarily focusing on stationary demand models. The final section illustrates some of these insights with an example.

4.1. Univariate versus multivariate demand models

What happens if we use a univariate demand model instead of a multivariate one? To start, we find that the multivariate bullwhip effect for the kth product (out of m) can be larger or smaller than the corresponding univariate bullwhip effect for product k, and this behavior occurs in spite of the superior forecast performance of the multivariate structure. To see why this is true, observe that (12) can be written as

$$\mathcal{B}^{(L)} = f^{(L)}(0) - \frac{1}{2\pi}\int_{-\pi}^{\pi} f^{(L)}(\lambda)\, d\lambda, \tag{14}$$

where L is explicitly denoted above, and $f^{(L)}$ is the spectral density matrix of the VMA(L) process defined such that its sth coefficient is $\mathbf{\Xi}_s$ for $0 \le s \le L$, and with error variance matrix $\mathbf{\Sigma}$. The formula for $f^{(L)}$ is $f^{(L)}(\lambda) = \sum_{s=0}^{L} \mathbf{\Xi}_s e^{-i\lambda s}\, \mathbf{\Sigma} \sum_{t=0}^{L} \mathbf{\Xi}_t' e^{i\lambda t}$, from which (14) follows. If the demand process is stationary with a Wold representation (1), then its spectral density matrix is

$$f(\lambda) = \sum_{s=0}^{\infty} \mathbf{\Psi}_s e^{-i\lambda s}\, \mathbf{\Sigma} \sum_{t=0}^{\infty} \mathbf{\Psi}_t' e^{i\lambda t}, \tag{15}$$

and in such a case $f^{(L)}$ is a truncated version of f, in the sense that the corresponding VMA(L) process is a truncation to lag L of the true VMA(∞) demand process. Therefore, it follows that the asymptotic bullwhip effect given by $\lim_{L\to\infty}\mathcal{B}^{(L)}$ is $\mathcal{B}^{(\infty)} = f(0) - \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\lambda)\, d\lambda$. If the demand process is a VMA(q) for some $q < \infty$, then the bullwhip effect is identical with its asymptotic limit reached once L exceeded q (i.e., $\mathcal{B}^{(L)} = \mathcal{B}^{(\infty)}$ for $L \ge q$). Similar results have been found in the univariate case (Nagaraja, Thavaneswaran, & Appadoo, 2015). In contrast, a VAR(p) demand process with a non-truncating Wold representation would not have this property; $\mathcal{B}^{(L)}$ converges to the asymptotic bullwhip effect, but will never equal it.

Returning to the general case of a stationary demand process with spectral density matrix (15), the spectral density of the kth component time series of demand is just the kth diagonal of f, or f_{kk}. The Wold representation of this univariate time series can be determined via spectral factorization (Hannan & Deistler, 1988), whereby a variance σ^2 and a power series $\psi(B) = \sum_{s=0}^{\infty}\psi_s B^s$ are identified, having the following relationship to the given f_{kk}: $f_{kk}(\lambda) = \left|\sum_{s=0}^{\infty}\psi_s e^{-is\lambda}\right|^2 \sigma^2$. It is known that $\sigma^2 \ge \mathbf{\Sigma}_{kk}$, which indicates that mean squared forecast error is always weakly improved over a univariate structure by including additional series; the inequality is strong when some of the additional demand components Granger cause (Lütkephol, 2005) the kth demand component.

The univariate bullwhip effect is computed in an analogous manner as the multivariate case described above: with the truncated spectral density $f_{kk}^{(L)}(\lambda) = \left|\sum_{s=0}^{L}\psi_s e^{-is\lambda}\right|^2 \sigma^2$, we have in analogy with (14),

$$b^{(L)} = f_{kk}^{(L)}(0) - \frac{1}{2\pi}\int_{-\pi}^{\pi} f_{kk}^{(L)}(\lambda)\, d\lambda. \tag{16}$$

In general this univariate bullwhip for the kth series can differ from the kth component of the multivariate bullwhip, i.e., $b^{(L)} \ne \mathcal{B}_{kk}^{(L)}$, but they are equal asymptotically (as $L \to \infty$). This is because $\mathcal{B}_{kk}^{(\infty)} = f_{kk}(0) - \frac{1}{2\pi}\int_{-\pi}^{\pi} f_{kk}(\lambda)\, d\lambda = b^{(\infty)}$. Moreover, note that for a VMA(q) demand process where $q < \infty$, the univariate and bivariate bullwhip measures are equal whenever $L \ge q$ (because for such an L, the lead L bullwhip has already attained its asymptotic limit).

The question which still remains is whether to measure the bullwhip effect with a multivariate model or m (individual) univariate models. Given the range of models and settings considered here, it is not possible to provide a general heuristic to answer this practical question. However, for any stationary model which can be expressed as a VAR(p) model, there exists an implied univariate model. Specifically, the implied univariate model for a VAR(p) model with m series is an ARMA(mp, $(m-1)p$) for each of those m series. The parameter estimates for these univariate models can be calculated directly from the fitted multivariate one (Tsay, 2014; Section 2.4.4). By comparing the bullwhip effect for the multivariate and the implied univariate models, we have a general technique to choose the approach with the lower bullwhip effect without having to fit multiple demand models. A single, multivariate model is sufficient. We provide an example of this strategy in Section 4.3.

4.2. Effects of lead-time

For the univariate setting (stationary and nonstationary), a key recommendation for diminishing the bullwhip effect is to reduce L (e.g., Gilbert, 2005). However, Luong and Phien (2007) show that the existence of negative autoregressive parameters can reverse this advice (see their Table 1a on page 204). An analogous analysis in the multivariate case would be to study the diagonal elements of the autoregressive coefficient matrices. We find the univariate insights generally apply here.

Another, more technical, approach is to compare the order of the demand model with L. Nagaraja et al. (2015) showed that for univariate models, setting L below the seasonal lag eliminates the impact of the seasonal effect. In the multivariate case, this means that the $\mathbf{\Xi}_s$ terms when $s > L$ do not contribute to the bullwhip effect magnitude.

A special case of this property produces an interesting interpretation of the off-diagonal terms of $\mathcal{B}$. At one extreme, if all cross-series terms are zero (i.e., $\mathcal{B}$ is a diagonal matrix), then there is a separate bullwhip effect for each of the m individual products. Modeling m independent products would produce such a matrix. Now, let us loosen the independence requirement and see what happens. Instead, say that $\mathcal{B}$ can be partitioned into a block diagonal matrix. (Let b be the number of blocks where $b \le m$; therefore, when $b = m$, $\mathcal{B}$ is diagonal.) Each block, then, has a separate multivariate bullwhip effect.

From (12), we can determine the sufficient conditions under which a block diagonal $\mathcal{B}$ is generated. Specifically, one must be able to express the following matrices using an identical block diagonal structure: (a) $\mathbf{\Xi}_s$ where $0 \le s \le L$ and (b) $\mathbf{\Sigma}$. (For stationary models, we can replace $\mathbf{\Xi}_s$ with $\mathbf{\Psi}_s$.)

Note from (a) above that the demand model structure alone does not define these blocks. For example, the block diagonal structure can be violated for lags exceeding L, but still form blocks in $\mathcal{B}$ since those terms do not contribute to the bullwhip effect. This indicates that products across blocks can be unrelated in terms of measuring the bullwhip effect, even if they are dependent in the overall multivariate demand model. This property is similar to how econometric variables may be partitioned into groups according to the criterion of Granger causality (Lütkephol, 2005) – but here, the groups are determined by bullwhip variability rather than forecast variability.

4.3. Example

To visualize the insights from this section, consider a VAR(1) demand process with coefficient matrix $\mathbf{\Phi}_1$ and where $m = 2$. If we were to use separate models for each product, we can derive the corresponding implied univariate models by computing $(\mathbf{1}_m - \mathbf{\Phi}_1 B)^{-1}$, which can be determined from the adjoint (adj) and the determinant (det) of $\mathbf{\Phi_1}$. Setting $\varphi(B)$ equal to the determinant yields $\varphi(B) = 1 - \text{tr}\,\mathbf{\Phi}_1 B + \det \mathbf{\Phi}_1 B^2$, where tr is the trace of a matrix. The adjoint is $\text{adj}[\mathbf{\Phi_1}(B)] = \begin{bmatrix} 1-\phi_{22}B & \phi_{12}B \\ \phi_{21}B & 1-\phi_{11}B \end{bmatrix}$.

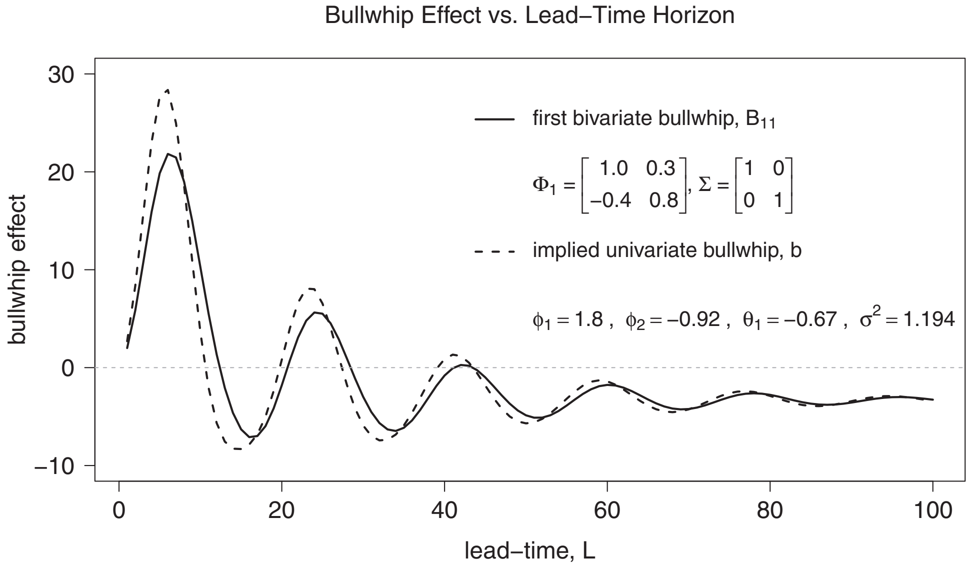

Fig. 1. We examine the relationship between the bullwhip effect and lead-time horizon in the VAR(1) demand case, comparing the bullwhip effect for the first product using a multivariate demand model (solid line) versus the implied univariate demand model which is an ARMA(2,1) (dotted line). (Note: Both demand models included in the plot are stationary.).

This equation specifies the autoregressive components of the implied univariate model, an ARMA(2,1), which are extracted from the VAR(1) parameters. However, we still need to derive the value of the moving average component (ψ_1) and the variance of the error term, σ^2. We can do this as as follows.

The ARMA(2,1) process described above is a 1-dependent process, with autocovariances at lags 0 and 1 given by $\gamma(0) = (1 + \phi_{22}^2)\,\mathbf{\Sigma}_{11} + \phi_{12}^2\,\mathbf{\Sigma}_{22} - 2\,\phi_{12}\,\phi_{22}\,\mathbf{\Sigma}_{12}$ and $\gamma(1) = -\phi_{22}\,\mathbf{\Sigma}_{11} + \phi_{12}\,\mathbf{\Sigma}_{12}$. Therefore, the implied spectral density for the first component process is $f_{11}(\lambda) = \gamma(0) + 2\,\gamma(1)\cos(\lambda)$. Spectral factorization provides ψ_1 and σ (note: $\psi_0 = 1$) such that $(1+\psi_1^2)\,\sigma^2 = \gamma(0)$ and $\psi_1\,\sigma^2 = \gamma(1)$, as these are the formulas for the autocovariances of an MA(1) with polynomial $1 + \psi_1 B$. Although for some special cases analytical formulas for ψ_1 and σ are explicitly available, typically solutions are computed using polynomial root-finding algorithms.

Now, let us set $\mathbf{\Phi}_1$ with $\phi_{11} = 1$, $\phi_{12} = 0.3$, $\phi_{21} = -0.4$, and $\phi_{22} = 0.8$, and $\mathbf{\Sigma} = \mathbf{1}_2$. Then, $\varphi(B) = 1 - 1.8\,B + 0.92\,B^2$ and the solutions are $\psi_1 = -0.67$ and $\sigma^2 = 1.194$ – which exceeds the corresponding bivariate forecast mean squared error of 1 (e.g., $\mathbf{\Sigma} = \mathbf{1}_2$), demonstrating the superior forecast performance of the bivariate demand model. Next, we compute both $\mathcal{B}_{11}^{(L)}$ from (14) and $b^{(L)}$ from (16) for varying L, which we plot in Fig. 1.

In this graph, the convergence towards the asymptotic value $\mathcal{B}_{11}^{(\infty)}\big(\text{and } b^{(\infty)}\big)$ is observed at the right of the plot, whereas for lower values of L there is a discrepancy between the two measures. Both approaches generate a cyclical effect of lead-time on the bullwhip effect magnitude. Furthermore, we can see that the univariate bullwhip has a greater amplitude to its oscillations, indicating a more volatile relationship between b and L compared to the multivariate measure. Furthermore b also anticipates the bivariate bullwhip slightly.

In this example, nearly 60% of the lead-times in the graph result in a less volatile bullwhip measure if the multivariate model is used; reductions range from 2% to 184% which, over time, can yield substantial cost savings for the retailer. Consequently, in most cases for this product, a multivariate model is preferable to a univariate one. This example also illustrates that the bullwhip can be negative (and non-monotone), being a case where the inequality in (13) is reversed, highlighting the complex relationship between lead-time and effect magnitude.

5. Applications

To demonstrate the flexibility of our framework, we derive the components to compute (12) for five common multivariate demand models: stationary VAR(p), stationary VMA(q), integrated VAR(p) with first order-differencing, integrated VMA(q) with first-order differencing, and co-integrated VAR(1). For more details on these models, refer to Lütkephol (2005) or Tsay (2014).

5.1. VAR(p) demand

A stationary VAR(p) demand model is generally written as $\mathbf{d}_t = \mathbf{\Phi}_0 + \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{d}_{t-i} + \boldsymbol{\varepsilon}_t$ where $\mathbf{\Phi}_i$ are the autoregressive parameters. We can express the model above as a moving average to match the form in (1): $\mathbf{d}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \mathbf{\Psi}_i \boldsymbol{\varepsilon}_{t-i}$ where $\boldsymbol{\mu} = (\mathbf{1}_m - \sum_{i=1}^{p} \mathbf{\Phi}_i)^{-1}\mathbf{\Phi}_0$, $\mathbf{\Psi}_0 = \mathbf{1}_m$, and $\mathbf{\Psi}_i = \sum_{j=1}^{min(i,p)} \mathbf{\Phi}_j \mathbf{\Psi}_{i-j}$. The symbol $\boldsymbol{\mu}$ can be interpreted as the overall average demand and $\mathbf{\Phi}_i$ as the innovation weights. As this is a stationary model, there is no differencing polynomial, $\delta(B)$. Therefore, $\mathbf{\Xi}_s = \mathbf{\Psi}_s$. (Note that conditions are placed on $\mathbf{\Phi}_i$ to produce a stationary model.)

5.2. VMA(q) demand

We can express a stationary VMA(q) demand model as $\mathbf{d}_t = \boldsymbol{\mu} + \sum_{i=0}^{q} \mathbf{\Theta}_i \boldsymbol{\varepsilon}_{t-i}$ where $\mathbf{\Theta}_0 = \mathbf{\Psi}_0 = \mathbf{1}_m$, $\mathbf{\Theta}_i = \mathbf{\Psi}_i$ for $1 \le i \le q$, and $\mathbf{\Theta}_i = \mathbf{\Psi}_i = 0$ for $i > q$. Certain conditions on $\mathbf{\Theta}$ must hold for this model to be invertible. Since this model is also stationary, $\mathbf{\Xi}_s = \mathbf{\Psi}_s$.

5.3. Integrated VAR(p) demand

For this nonstationary process, the demand model is $(\mathbf{d}_t - \mathbf{d}_{t-1}) = \mathbf{\Phi}_0 + \sum_{i=1}^{p} \mathbf{\Phi}_i(\mathbf{d}_{t-i} - \mathbf{d}_{t-1-i}) + \boldsymbol{\varepsilon}_t$ and can be expressed as $(\mathbf{d}_t - \mathbf{d}_{t-1}) = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \mathbf{\Psi}_i \boldsymbol{\varepsilon}_{t-i}$ where $\boldsymbol{\mu} = (\mathbf{1}_m - \sum_{i=1}^{p} \mathbf{\Phi}_i)^{-1}\mathbf{\Phi}_0$, $\mathbf{\Psi}_0 = \mathbf{1}_m$, and $\mathbf{\Psi}_i = \sum_{j=1}^{min(i,p)} \mathbf{\Phi}_j \mathbf{\Psi}_{i-j}$. The second expression for demand is the moving average representation of the autoregressive process as given in (1). As before, $\mathbf{\Phi}_i$ must satisfy certain conditions to produce an invertible model.

For a first-order difference (i.e., $d = 1$), the coefficients of $\delta(B)$ are $\boldsymbol{\delta}_0 = \mathbf{1}_m$ and $\boldsymbol{\delta}_1 = -\mathbf{1}_m$. Then, the coefficients of $\delta^{-1}(B)$ are $\boldsymbol{\delta}_j^{-1} = \mathbf{1}_m$ for $j \ge 0$. (Note: in the univariate case $\delta(B) = 1 - B$, which corresponds to the polynomial $f(x) = 1 - x$; its inverse is

$1/(1-x)=\sum_{i=0}^{\infty}x^i$. This is simply the multivariate version extension.) Finally, $\boldsymbol{\Xi}_s=\sum_{i=0}^{s}\boldsymbol{\Psi}_i$.

5.4. Integrated VMA(q) demand

This nonstationary process is expressed as $(\mathbf{d}_t-\mathbf{d}_{t-1})=\boldsymbol{\mu}+\sum_{i=0}^{q}\boldsymbol{\Theta}_i\boldsymbol{\varepsilon}_{t-i}$ where $\boldsymbol{\Theta}_0=\boldsymbol{\Psi}_0=\mathbf{1}_m$ and $\boldsymbol{\Theta}_i=\boldsymbol{\Psi}_i$ for $i>0$. Certain conditions on $\boldsymbol{\Theta}$ must hold for the differenced model to be both stationary and invertible. We have the same differencing polynomial $\boldsymbol{\delta}(B)$ (and inverse) as in the previous model, since both models involve first-order differencing. To compute the bullwhip effect, set $\boldsymbol{\Xi}_s$ equal to $\sum_{i=0}^{s}\boldsymbol{\Psi}_i$ if $s<q$ or to $\sum_{i=0}^{q}\boldsymbol{\Psi}_i$ if $s\geq q$.

5.5. Co-integrated VAR(1) demand

This nonstationary process is written $\mathbf{d}_t=\boldsymbol{\Phi}\,\mathbf{d}_{t-1}+\boldsymbol{\varepsilon}_t$, where $\boldsymbol{\Phi}$ has r unit eigenvalues and $m-r$ stable eigenvalues with modulus less than one. Following Gómez (2016, Section 5.7), there exist $m\times r$ dimensional matrices α and β such that $\boldsymbol{\Phi}-\mathbf{1}_m=\alpha\,\beta'$. Set $\boldsymbol{\Upsilon}=\alpha\,\beta'+\beta\,(\beta'\,\beta)^{-1}\,\beta'$ and $\mathbf{U}=\beta_\perp\,(\beta_\perp'\,\beta_\perp)^{-1}\,\beta_\perp'$, where the columns of $\beta_\perp$ span the nullspace of β' (where $\beta_\perp$ is the orthogonal complement of β). Then, we have the decomposition $\mathbf{1}_m-\boldsymbol{\Phi}\,B=(\mathbf{1}_m-\boldsymbol{\Upsilon}\,B)\,(\mathbf{1}_m-\mathbf{U}\,B)$. Here, $\boldsymbol{Y}$ has stable eigenvalues, but $\mathbf{U}$ has eigenvalues that are either zero or one. This means that $\boldsymbol{\delta}(B)=\mathbf{1}_m-\mathbf{U}\,B$ is the differencing operator, and $\partial\mathbf{d}_t=\sum_{i=0}^{\infty}\boldsymbol{\Upsilon}^i\boldsymbol{\varepsilon}_{t-i}$ so that $\boldsymbol{\Psi}_i=\boldsymbol{\Upsilon}^i$. Moreover, $\boldsymbol{\Xi}(B)=(\mathbf{1}_m-\boldsymbol{\Phi}\,B)^{-1}$, which has in its expansion the ith coefficient $\boldsymbol{\Phi}^i$.

6. Data examples

In this section, we show empirical evidence of multivariate demand beyond the VAR(1) case with a simulation-based approach. We use the publicly available Dominick's Database published by the Kilts Center for Marketing, The University of Chicago Booth School for Business. This database contains weekly store-level sales from Dominick's Finer Foods from 1989 through 1994 around Chicago, Illinois in the U.S.A. The "Category Specific Files" contain data for individual products, identified by UPC, and (a) week of sale (numbered with no calendar date), (b) description of the product, (c) store code, and (d) quantity sold. Information regarding orders placed or lead-time information is available. Details regarding imputing missing data, summary statistics, and model fitting can be found in the supplementary file titled "Data and demand model summaries." Bullwhip computation code written in R is provided in the supplementary file "multivariate_bullwhip_code.R".

We focus on two examples: a stationary VAR(3) model and an integrated VAR(5) model. In choosing these cases, we have selected products which are low-cost, high use, and nonperishable – products that are most likely to follow an order-up-to-inventory policy. We assume that there is sufficient safety stock to satisfy observed demand. This is a strong but necessary assumption since we have no information about actual inventory levels. We also suppose that the supplier always fulfills the retailer's orders completely, a standard (often implicit) assumption in the bullwhip literature (Chen, Luo, & Shang, 2017). Finally, we assume $\{\boldsymbol{\varepsilon}_t\}$ is a normal multivariate white noise series.

6.1. Stationary demand

We consider sales of toilet paper (DOM WHITE BATH TISSUE; UPC = 3828111217) and paper towels (BOUNTY WHITE/DESIGNER; UPC = 3700063527) at store 112 (Buffalo Grove: 1160 W. Lake Cook Road, Buffalo Grove, IL 60090). Let $\{d_{1,\,t}\}$ denote the demand series for toilet paper, and $\{d_{2,\,t}\}$ for paper towels. Fig. 2 shows the pair of demand series, indicating imputed values. While toilet paper sales are more volatile than paper towel sales, both products appear to be stationary, roughly following a similar pattern across time suggesting the application of the multivariate approach.

We fit a VAR(3) model to the observed demand and define the model in terms of the notation given in (1), incorporating the formulas from Section 5.1. The mean demand, then, is $\hat{\boldsymbol{\mu}}=$ [64.6023 units/week 60.3646 units/week]. To obtain $\boldsymbol{\Psi}$, we express the model in the MA representation by defining $\hat{\boldsymbol{\Phi}}_i$:

$$\hat{\boldsymbol{\Phi}}_1=\begin{bmatrix}0.1941 & 0.0357\\ 0.0597 & 0.2061\end{bmatrix},\ \hat{\boldsymbol{\Phi}}_2=\begin{bmatrix}0.2079 & 0.1741\\ 0.0616 & 0.1447\end{bmatrix},$$
$$\hat{\boldsymbol{\Phi}}_3=\begin{bmatrix}0.1357 & 0.0711\\ -0.0002 & 0.1754\end{bmatrix}. \tag{17}$$

Finally, the estimate for the covariance matrix is $\hat{\boldsymbol{\Sigma}}=\begin{bmatrix}756 & 117\\ 117 & 434\end{bmatrix}$.

Since we fit a stationary model, we can compare the empirical and parametric bullwhip estimates (see Section 3.4). To compute the parametric estimate, we calculate (12) using the estimates $\hat{\boldsymbol{\Phi}}_i$ and $\hat{\boldsymbol{\Sigma}}$. For the empirical estimate, we take a simulation based approach as order quantities are not included in the database. Therefore, for this example, we fit a multivariate time series model to the demand data and simulate the quantity ordered series corresponding to the demand forecasts. (This approach is similar to one used in Dejonckheere, et al, 2003, p. 577.) We can then compute the difference between the variance of the simulated orders and the observed demand to obtain our empirical bullwhip estimate. In Fig. 3, we show the results of our experiment for a range of lead-times, $L=1,\ldots,10$ weeks. In both plots, L is on the horizontal axis and the relevant diagonal element of $\mathcal{B}$ is on the vertical axis.

The parametric and empirical estimates graphed in Fig. 3 are close, in part indicating that the estimated VAR(3) demand model was a good fit for the data, validating the use of the parametric approach. However, the estimates diverge slightly as the lead-time L increases and the parametric estimate seems to be consistently higher than the empirical one. We find that the discrepancies between estimates tend to appear after $L>p$ for VAR(p) demand models, although neither the magnitude (i.e., growing or shrinking $\mathcal{B}$ as L increases, given $L>p$) nor the direction (i.e., which estimate of $\mathcal{B}$ is higher) is consistent across demand models. (If we relax our assumptions on having sufficient safety stock and perfectly filled orders, this analysis could be expanded to study differences between material and information flow bullwhip estimates. Such an analysis would extend Chen et al. (2017) from AR(1) models to the general, stationary multivariate demand case. Our empirical estimate of $\mathcal{B}$ can be connected to their information flow concept.)

6.2. Nonstationary demand

In this example, we look at one product, JOB SQUAD DECORATIVE (UPC 5400014190) paper towels, which is sold at two individual Dominick's stores: Store 112 (1160 W. Lake Cook Road and Buffalo Grove, IL 60090) and Store 115 (1300 S. Naper Blvd, Naperville, IL 60540). Here, $m=2$, but are differentiated by facility rather than product. In the database, each product has a code called "nitem," the last digit of which indicates whether the item was sent directly to the store or to a Dominick's warehouse (0 versus 1). For this product, the last digit is 1, signifying that shipments were delivered to the warehouse first. We assume in this example that these stores share a common warehouse.

It is tempting to model a single series here by aggregating demand from both branches since we are examining the same product. However, given that merchandise must eventually be shipped separately to stores, we would still need suitable demand forecasts for each branch.

From Fig. 4A and C, it is clear that the demand in both stores is nonstationary. However, if we take the first-order

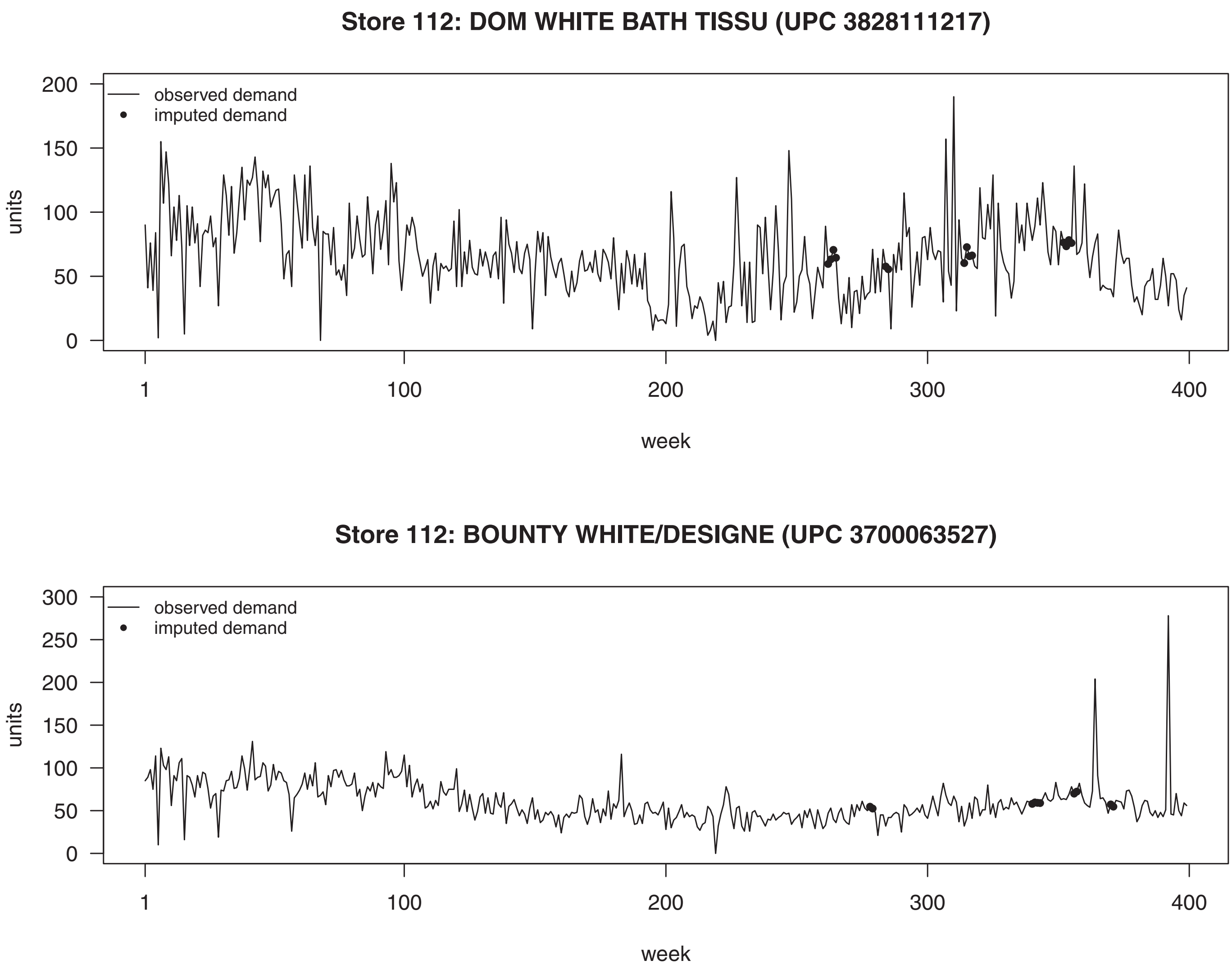

Fig. 2. Observed demand for the toilet paper and paper towel sales from Example 6.1.

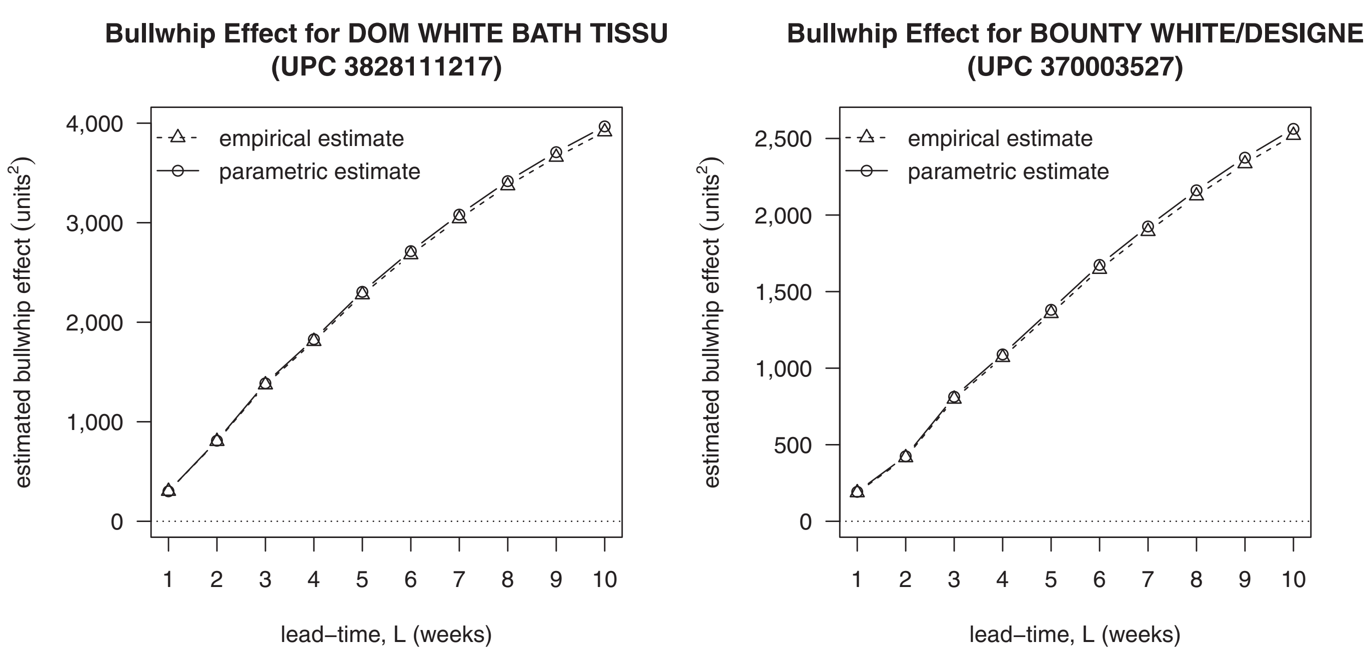

Fig. 3. Comparing parametric and empirical approaches to estimating the multivariate bullwhip effect for the stationary VAR(3) demand model fit in Example 6.1.

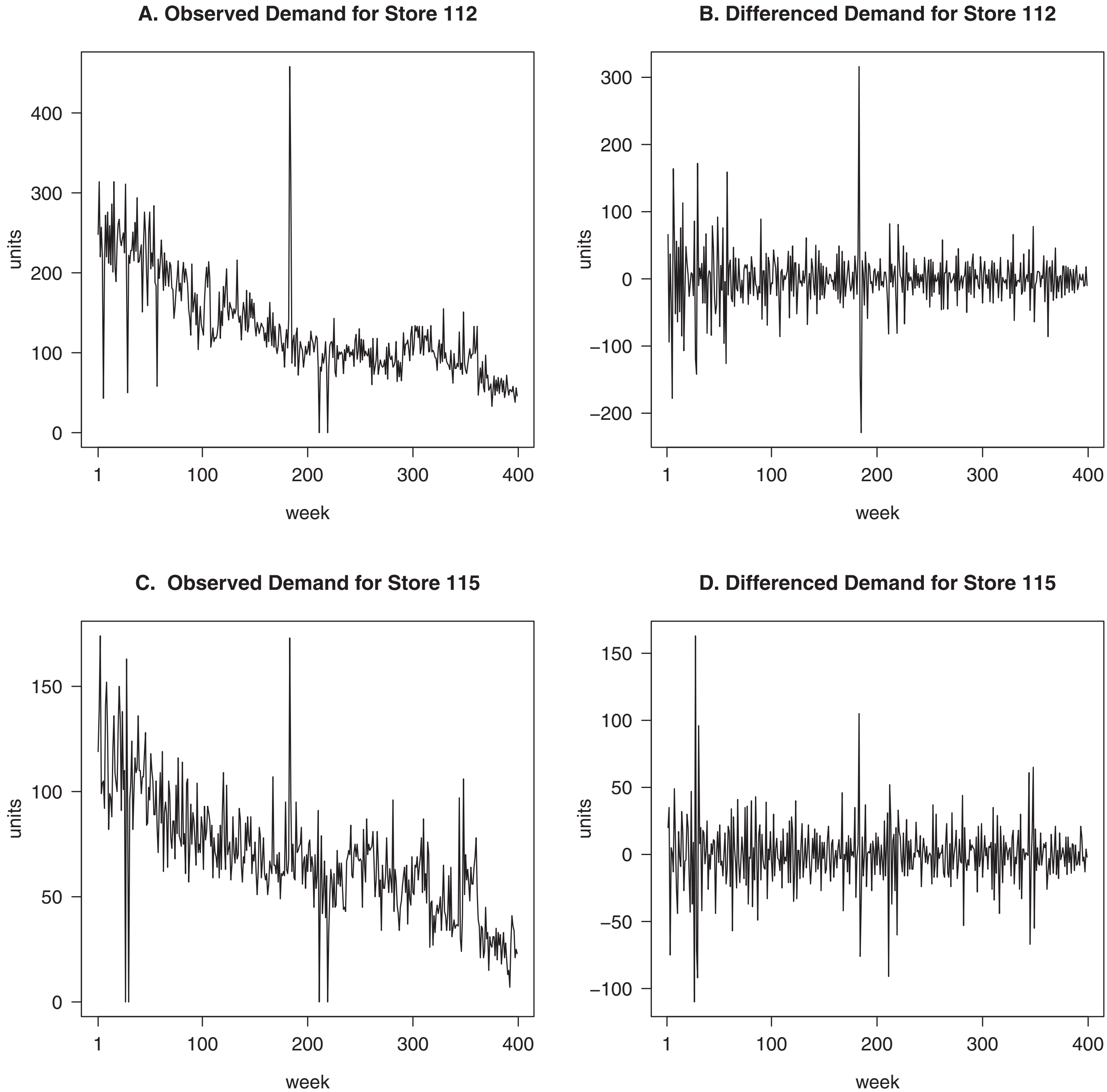

Fig. 4. For the paper towel sales in Example 6.2 (JOB SQUAD DECORATIVE with UPC 5400014190), demand is nonstationary in both stores. Plots A and C show the original demand series (including imputed values) and plots B and D show the corresponding differenced series for all weeks.

difference – as graphed in Fig. 4B and D – the resulting series is not. We will fit a VAR(5) model on this differenced series (i.e., an integrated VAR(5) model on the original multivariate series). To start, let $\{d_{1,\,t}\}$ denote sales from store 112 and $\{d_{2,\,t}\}$ from store 115. Then, $\partial d_{1,t} = d_{1,t} - d_{1,t-1}$ and $\partial d_{2,t} = d_{2,t} - d_{2,t-1}$. The fitted model can be expressed as in (1) following the steps outlined in Section 5.3, beginning with generating the differencing polynomial. The mean vector (for the differenced data) is $\boldsymbol{\mu} = [-0.3667 \text{ units/week} \quad -0.2503 \text{ units/week}]$. To obtain Ψ_i, let

$$\hat{\boldsymbol{\Phi}}_1 = \begin{bmatrix} -0.6122 & 0.0959 \\ 0.1533 & -0.9677 \end{bmatrix}, \hat{\boldsymbol{\Phi}}_2 = \begin{bmatrix} -0.5584 & 0.1820 \\ 0.1232 & -0.8333 \end{bmatrix},$$

$$\hat{\boldsymbol{\Phi}}_3 = \begin{bmatrix} -0.4247 & 0.1107 \\ 0.0270 & -0.5056 \end{bmatrix}, \hat{\boldsymbol{\Phi}}_4 = \begin{bmatrix} -0.2731 & 0.2070 \\ -0.0345 & -0.1849 \end{bmatrix},$$
$$\hat{\boldsymbol{\Phi}}_5 = \begin{bmatrix} -0.1901 & 0.0865 \\ -0.0539 & -0.0733 \end{bmatrix} \tag{18}$$

with $\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1,251 & 195 \\ 195 & 308 \end{bmatrix}$.

In this example, the demand model parameters in (18) show negative autocorrelation and negative cross-correlation for product 1. This results in a positive bullwhip effect in both stores; for instance, when $L = 3$, $\mathcal{B} = \begin{bmatrix} 2,902.5 & 684.7 \\ 684.7 & 445.2 \end{bmatrix}$. As with the stationary example, the bullwhip effect is present in both stores and increases with lead-time L. (All cross-diagonal elements of $\mathcal{B}$ are positive and grow with L too.)

7. Discussion

It has been established that information sharing across supply chain levels is a means to reduce the bullwhip effect (e.g., Chen et al., 2000a; Lee et al., 2000). This can be thought of as a vertical approach to information sharing. In this paper, we propose a second form: *Our multivariate bullwhip methodology – which incorporates sales data across a flexible supply chain structure – can be interpreted as a horizontal information sharing strategy.* A vertical approach requires firms to work collaboratively, which can be difficult in practice, whereas a horizontal one does not. Furthermore, Graves (1999) notes that vertical information sharing is not useful for reducing the bullwhip effect for ARIMA(0,1,1) demand models. Consequently, our *multivariate approach is critical for reducing the bullwhip effect for nonstationary demand settings*.

Gilbert (2005) calls forecast error a Type 2 source of the bullwhip effect and it can be reduced with better – possibly multivariate – demand models. *By expanding the number of related products modeled simultaneously* (*i.e.,* $m > 1$), *the tradeoff for more complexity can be a lower bullwhip effect.* Furthermore, we show *empirical evidence for our multivariate approach* with our analysis of the Dominick's Database.

Our approach produces a *single formula for the bullwhip estimate which works for both stationary and nonstationary cases*. For simple demand models, which require fewer parameters to be estimated, we show that the *parametric bullwhip estimates are reasonably accurate*. The advantage of the parametric estimate over the empirical one is the ability to gain insight into the drivers of the bullwhip effect, which is important for managing its impact, especially when nonstationary demand is observed.

As in the univariate setting, *reducing the lead-time can reduce the multivariate bullwhip effect*. Furthermore, we find that *demand model terms* $\boldsymbol{\Xi}_s$ *where* $s > L$ *do not impact the bullwhip effect magnitude, a property which can be used to reduce the phenomenon.* There is a special case of this property when, given a lead-time L, the multivariate bullwhip matrix is block diagonal. This occurs when one can express the following matrices using an identical block diagonal structure: (a) $\boldsymbol{\Xi}_s$ where $0 \le s \le L$ and (b) $\boldsymbol{\Sigma}$. With such a structure, products between blocks are unrelated. The key here is that *it is possible that two products operate under separate bullwhip effects even if their demand is dependent, provided that dependence occurs after lag L. These are examples of how the demand structure itself can be used by a firm to select a suitable lead-time.*

Finally, we provide a technique firms can use to determine whether the multivariate or univariate model is preferred in the stationary case: *Fit a single demand model – a multivariate one – and then derive the implied univariate model to decide which option generates a lower bullwhip effect.* This requires only one model to be fit, from which optimal ordering decisions can be made.

By considering an extensive range of demand models in this paper, including nonstationary ones, we have presented a general framework for studying the bullwhip phenomenon. Taking such an approach allows us to view demand with a more realistic lens: not as individual processes in isolation, but rather as an interdependent set of series. From our analysis, we can see that horizontal information sharing, through a multivariate demand model, can cast a new light on understanding and managing the bullwhip phenomenon.

Acknowledgments

We thank the editor and the anonymous referees for their helpful comments and suggestions for improving this paper. We also thank the Gabelli School of Business at Fordham Universityfor providing summer research support (SS15-39) for this work.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ejor.2017.11.015

References

Boute, R. N., Dejonckheere, J., Disney, S. M., & Van de Velde, W. (2012). Bullwhip in a multi-product production setting. *Review of Business and Economic Literature, 57*(3), 354–381.

Bray, R. L., & Mendelson, H. (2015). Production smoothing and the bullwhip effect. *Manufacturing & Service Operations Management, 17*(2), 208–220.

Chaharsooghi, S. K., & Sadeghi, A. (2008). On the bullwhip effect measure in supply chains with VAR(1) demand process. *International Journal of Industrial Engineering & Production Research, 19*(4), 9–19.

Chen, A., & Blue, J. (2010). Performance analysis of demand planning approaches for aggregating, forecasting, and disaggregating interrelated demands. *International Journal of Production Economics, 128*(2), 586–602.

Chen, F., Drezner, Z., Ryan, J. K., & Simchi-Levi, D. (2000). Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, lead times, and information. *Management Science, 46*, 436–443.

Chen, F., Ryan, J. K., & Simchi-Levi, D. (2000). The impact of exponential smoothing forecasts on the bullwhip effect. *Naval Research Logistics, 47*, 269–286.

Chen, L., & Lee, H. L. (2009). Information sharing and order variability control under a generalized demand model. *Management Science, 55*(5), 781–797.

Chen, L., Luo, W., & Shang, K. (2017). Measuring the bullwhip effect: Discrepancy alignment between information and material flows. *Manufacturing & Service Operations Management, 19*(1), 36–51.

Dejonckheere, J., Disney, S. M., Lambrecht, M. R., & Towill, D. R. (2003). Measuring and avoiding the bullwhip effect: A control theoretic approach. *European Journal of Operational Research, 147*, 567–590.

Disney, S. M., Maltz, A., Wang, X., & Warburton, R. D. H. (2016). Inventory management for stochastic lead times with order crossovers. *European Journal of Operational Research, 248*, 473–486.

Gaalman, G. (2006). Bullwhip reduction for ARMA demand: The proportional order-up-to-policy versus the full-state-feedback policy. *Automatica, 42*, 1283–1290.

Gilbert, K. (2005). An ARIMA supply chain model. *Management Science, 51*(2), 305–310.

Gill, P., & Abend, J. (1997). Wal-mart: The supply chain heavyweight champ. *Supply Chain Management Review, 1*, 8–16.

Gómez, V. (2016). *Multivariate time series with linear state space structure*. New York: Springer.

Graves, S. C. (1999). A single-item inventory model for a nonstationary demand process. *Manufacturing & Service Operations Management, 1*(1), 150–161.

Hannan, E. J., & Deistler, M. (1988). *The statistical theory of linear systems*. New York: John Wiley and Sons.

Kilts Center for Marketing, The University of Chicago Booth School for Business Dominick's database. https://research.chicagobooth.edu/kilts/marketing-databases/dominicks.

Kong, G., Rajagopalan, S., & Zhang, H. (2013). Revenue sharing and information leakage in a supply chain. *Management Science, 59*(3), 556–572.

Kwak, J. K., & Gavirneni, S. (2015). Impact of information errors on supply chain performance. *Journal of the Operational Research Society, 66*(2), 288–298.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997a). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*, 546–558.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997b). The bullwhip effect in supply chains. *Sloan Management Review, 38*, 93–102.

Lee, H. L., So, K. C., & Tang, C. S. (2000). The value of information sharing in a two-level supply chain. *Management Science, 46*, 626–643.

Luong, H. T., & Phien, N. H. (2007). Measure of bullwhip effect in supply chains: The case of high order autoregressive demand process. *European Journal of Operational Research, 183*, 197–209.

Lütkephol, H. (2005). *New introduction to multiple time series analysis.*. Germany: Springer-Verlag.

Ma, J., & Wu, W. (2014). A kind of multi-stage supply chain bullwhip effect and its application research based on the complex system theory. *Journal of Convergence Information Technology, 9*, 72–83.

McElroy, T., & McCracken, M. W. (2015). Multistep ahead forecasting of vector time series. *Econometric Reviews*, 1–26. Published online.

Mishra, B. K., Raghunathan, S., & Yue, X. (2007). Information sharing in supply chains: Incentives for information distortion. *IIE Transactions, 39*(9), 863–877.

Nagaraja, C. H., Thavaneswaran, T., & Appadoo, S. S. (2015). Measuring the bullwhip effect for supply chains with seasonal demand components. *European Journal of Operational Research, 242*, 445–454.

Raghunathan, S., Tang, C. S., & Yue, X. (2017). Analysis of the bullwhip effect in a multiproduct setting with interdependent demands. *Operations Research, 65*(2), 424–432.

Tsay, R. S. (2014). *Multivariate time series analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Voigt, G., & Inderfurth, K. (2012). Supply chain coordination with information sharing in the presence of trust and trustworthiness. *IIE Transactions, 44*(8), 637–654.

Wang, X., & Disney, S. M. (2016). The bullwhip effect: Progress, trends and directions. *European Journal of Operational Research, 250*, 691–701.

Xu, K., Dong, Y., & Xia, Y. (2015). 'to little' or 'too late": The timing of supply chain demand collaboration. *European Journal of Operational Research, 241*, 370–380.

Zhang, X. (2004). The impact of forecasting methods on the bullwhip effect. *International Journal of Production Economics, 88*, 15–27.

Zhang, X., & Burke, G. J. (2011). Analysis of compound bullwhip effect causes. *European Journal of Operational Research, 210*, 514–526.