Identifying paths-to-purchase segments via Clustered Vector Autoregression

Yicheng Song, Nachiketa Sahoo, Shuba Srinivasan, Chrysanthos Dellarocas

School of Management, Boston University

Abstract

Vector Autoregression Models (VAR) are widely used by researchers to capture the linear interdependencies among multiple time series. We propose a novel method called Clustered VAR (CVAR) to identify components of the data generated by a mixture of *K* VAR processes. By applying a CVAR model to a consumer-level time series dataset on shopping behavior at a retailer, we segment consumers based on their path-to-purchase. We estimate the CVAR model using the EM (expectation–maximization) algorithm that assigns each consumer into a segment that maximizes the likelihood and optimizes the VAR parameters for each segment given the membership assignments. We verify the effectiveness of the Clustered VAR model on a simulated dataset. Following successful evaluation, we apply the Clustered VAR model to a retail dataset from a major multi-channel, multi-brand North American Retailer. Our study could segment 2,000 randomly selected consumers into 4 clusters and offers insights on two issues: 1. Potential interdependencies among online marketing, offline marketing and their effects for each group, 2. Differences in the above effects across consumer segments. As a result, the consumer clusters in our study will guide managers in tailoring the marketing mix for different customer segments to help them move forward on the path-to-purchase.

Identifying paths-to-purchase segments via Clustered Vector Autoregression

Yicheng Song, Nachiketa Sahoo, Shuba Srinivasan, Chrysanthos Dellarocas

School of Management, Boston University

1 Introduction

Vector Autoregression models [1] are widely adopted by researchers to capture dynamic linear interdependencies among multiple time series to analyze the dynamic interaction of different marketing channels. By applying VAR model on the aggregate data, we can get the macro-level dynamic interaction between marketing variables, such as the overall marketing to the overall revenue [1]. On the other hand, by applying VAR model on the individual level consumer data on shopping behavior [2] [6], we can get the individual-level path-to-purchase. In the current paper we extend such an individual level VAR approach to segment customers based on their paths-to-purchase. We call the proposed approach Clustered VAR (CVAR) model. It accomplishes following two goals simultaneously:

1. **Discover customer heterogeneity in their path to purchase:** Customers are likely to have different shopping behavior and react differently to the marketing campaigns. The proposed CVAR model detects these different groups.
2. **Estimate group-level dynamic interactions**: It estimates segment-level path-to-purchase that indicates a segment-member's unique transition between browsing, searching, online and offline shopping. It also estimates a segment-member's response to marketing campaigns and other exogenous factors.

2 Model Development

2.1 VAR Model

A typical $p\ th$ order VAR can be expressed as:

$$X_i^T = c + A^1 X_i^{T-1} + \cdots + A^p X_i^{T-p} + A^{p+1} Y_i^T + e \tag{1}$$

where X_i^T is endogenous vector of the $i\ th$ consumer at $T\ th$ period and $X_i^{T-1}, \dots, X_i^{T-p}$ are lag vectors of X_i^T. Y_i^T is the exogenous vector of the $i\ th$ consumer at $T\ th$ period[1]. Due to the assumption of $e \sim N(0, \Omega)$, we can apply OLS to estimate A^i. One problem is that when working with sparse count data the Normal assumption doesn't hold, which requires a modification of the standard VAR approach. In [7], researchers utilize Zero-Inflated-Poisson (ZIP) [5] to model the endogenous variables. Specifically,

$$P(X_i^T(j)|X_i^T(j) = 0) = p + (1-p)Poisson(X_i^T(j)|\lambda)$$
$$P(X_i^T(j)|X_i^T(j) > 0) = (1-p)Poisson(X_i^T(j)|\lambda) \tag{2}$$

where $Log(\lambda) = B\gamma\ and\ Logit(p) = G\gamma$, in which γ is the $X_i^{T-1}, \dots, X_i^{T-p}$ and exogenous variable Y_i^T. Thus, B and G that capture the dynamic interactions.

2.2 Clustered VAR Model

We now extend VAR to VAR mixture models. The data generating process of VAR mixture is:

1. Each consumer is drawn from one of *K* different groups with certain probability.
2. The time series data for the consumer is generated by VAR model of the group.

[1] The descriptions of endogenous and exogenous variables from retail dataset are listed in Table 1.

CVAR model could determine the membership of VAR mixture and meanwhile estimate the coefficients of group-level VAR parameters. The bayesian network of CVAR model is described in Figure 1.

We adopt EM algorithm [3] to solve the following two problems: 1) determine the membership of each consumer to the cluster via E step, 2) estimate group-level VAR parameters via M step.

In E step, we estimate the membership of each consumer using Bayes' theorem:

$$P(z_k|X_i, B_k, G_k) = \frac{P(z_k)P(X_i|B_k, G_k, z_k)}{\sum_{u=1}^{K} P(z_u)P(X_i|B_u, G_u, z_u)} = \frac{\pi_k P(X_i|B_k, G_k, z_k)}{\sum_{u=1}^{K} \pi_u P(X_i|B_u, G_u, z_u)} \quad (3)$$

In M step, we maximize the expectation of the complete log likelihood:

$$\sum_{i=1}^{N}\sum_{k=1}^{K} P(z_k|X_i, B_k, G_k) InP(X_i|z_k, B_k, G_k) + \sum_{i=1}^{N}\sum_{k=1}^{K} P(z_k|X_i, B_k, G_k) InP(z_k) \quad (4)$$

where (B_k, G_k) can be derived from maximizing $\sum_{i=1}^{N}\sum_{k=1}^{K} P(z_k|X_i, B_k, G_k) InP(X_i|z_k, B_k, G_k)$, which is equal to estimating the weighted Zero-Inflated-Poisson regression. Moreover, $\pi_k = P(z_k) = \frac{\sum_{i=1}^{N} P(z_k|X_i, B_k, G_k)}{N}$.

3 Experiments on Simulation Data

In order to validate the effectiveness of the CVAR model, we need a dataset with known cluster membership so that we can compare the result of CVAR classification with it. To obtain this, we simulate the data based on mixture VAR, and then we use CVAR model to estimate the membership of the simulation data.

The 2,000 consumer data are generated based on mixture VAR with 4 clusters, all of them have the same number of endogenous variables=5, number of exogenous variables=4, lag=2 and time span=T. For each cluster, we randomly set the transition matrixes B_i, G_i by a Normal Distribution (mean=0, variance=0.15). For each cluster, there are 500 consumers. Then, for each consumer, we generate the first two periods of endogenous data with a uniform distribution (min=0, max=2). We also generate exogenous data for each consumer based on the same uniform distribution. Based on the first two periods endogenous, exogenous data and the corresponding transition matrixes B_i, G_i we generate the endogenous data of remaining T-2 periods for that consumer.

We use CVAR to determine the membership of the simulation data and compare its performance with K-means (estimated individual-level VAR coefficients and cluster consumers based on their VAR coefficients). We launch 5 runs for each case and average the results. We employ purity [4], *fraction of the consumers assigned to the cluster containing majority of their original group members,* to measure the clustering accuracy. Table 2 shows the results when we set T=100, 200 and 500. We find that: K-means doesn't performance well when T=100 or 200. The reason is: the estimation of individual-level VAR coefficients is not accurate due to limited amount of individual-level data and the large number of parameters (150) that need to be estimated. In some situations with extreme sparseness, we are unable to estimate the zero-inflated-poisson regression using data of individual consumers. When the length of time-series data is longer, the performance of K-means improves and approaches that of CVAR. In contrast, CVAR performs well even with shorter time-series data: when set T=100, the purity of CVAR is good at 0.967. The reason CVAR could handle the shorter time-series data is that when estimating the coefficients, CVAR uses weighted data of all consumers rather than individual data. Overall, CVAR outperforms for different durations of time-series data.

4 Experiments on Retailer Data

We collect a dataset from a major multi-channel, multi-brand North American retailer, which includes data on marketing activity, customers' website activities, customers' purchases, as well as their post purchase behavior, such as product returns and product reviews. The dataset spans from the 26th week, 2010 to the 26th week, 2012. We segment the two-year data into weekly data. The dataset used in this study include randomly selected 2,000 consumers from one brand of the retailer. Both online and offline

purchase are quite sparse, for 23,891 consumers in this brand, 98.7% of the weekly online purchase and 94.66% of the weekly offline purchase data are zero. Therefore, as mentioned in section 2.1, we will use ZIP rather than OLS to estimate the transition matrixes.

4.1 Cluster Result

For CVAR model, the number of clusters K is determined based on the AIC, BIC criteria. Here we tried $K \in \{2,3,4,5\}$and their AIC and BIC comparison is illustrated in Figure 2. Due to small difference between $K = 4, K = 5$ and for the convenience of description, we choose $K = 4$. The relative size of four clusters are $\{57\%, 5\%, 24\%, 12\%\}$. Because the transition matrixes B_k, G_k are hard to interpret, we use impulse response to analysis the reactions of 4 clusters when they are facing with the impulse from different exogenous variables, the result is illustrated in Figure 3, we have following observations:

Cluster 1. This is the only one with an immediate positive response to email on both offline and online purchases; consumers in cluster 1 consumers are email-centric shoppers.

Cluster 2. Email has a negative effect on both purchases for consumers in cluster 2. However, there is immediate spike around holidays for both online and offline conversion. Although all 4 clusters have a positive conversion to offline from catalogs; the effect is strongest for cluster 2. These clues indicate cluster 2 consumers are holiday shoppers with their offline purchases driven by catalog.

Cluster 3. Consumers in this cluster have a negative effect of email on online purchase and also have negative conversion around holidays. Further they seem to substitute to offline sales (i.e. experience negative online sales) through catalog. Except cluster 3, other three clusters respond with higher online conversion in response to catalog. Thus, cluster 3 seems like catalog driven offline shopper.

Cluster 4. There is also an immediate spike around holidays for both online and offline conversion. But the effect to the online purchase is not that strong. Moreover, cluster 4 respond by higher online sales to catalog comparing to other clusters. These clues also indicate cluster 4 is a holiday shopper (similar to cluster 2), but the online purchase is highly driven by catalog.

4.2 Validation by out-of-sample prediction

To validate the CVAR model we measure its out-of-sample prediction ability and compare it to a VAR model that uses aggregate unsegmented data. We divide the 106 weeks data of all consumers into training data (first 96 weeks) and testing data (last 10 weeks), we train a VAR model[2] and a CVAR model based on the training data. Then we predict the number of online purchase and offline purchase in the testing data by the trained models and exogenous variables in the same period. The Mean Square Error (MSE) of prediction via VAR model is 1.31015, while it is 0.82646 for CVAR model, which shows the CVAR model predicts much better than VAR model.

In summary, the approach proposed in this research identifies heterogeneity in consumers' path to purchase and segments them using the proposed CVAR model. The results of the study will guide the retailer in tailoring the marketing mix for different customer segments on the path-to-purchase to drive sales conversion.

[2] Individual level data still modeled using ZIP, but all the users are assumed to belong to the same group.

References

[1] ZIVOT, E. AND WANG, J. 2006. Vector autoregressive models for multivariate time series. In Modeling Financial Time Series with S-PLUS. Springer New York, 385¨C429.

[2] WIESEL, T., PAUWELS, K., AND ARTS, J. 2011. Marketing¡¯s Profit Impact: Quantifying Online and Off-line Funnel Progression. Marketing Science 30, 4, 604¨C611.

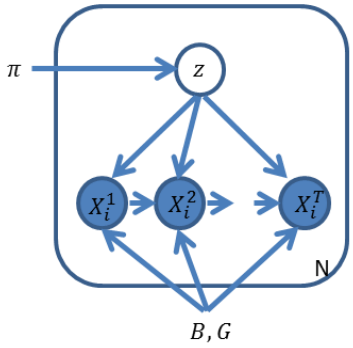
[3] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society, Series B 39 (1): 1–38.

[4] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: Introduction to information retrieval. Cambridge University Press 2008.

[5] Lambert, Diane (1992). "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing". Technometrics 34 (1): 1–14.

[6] A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty. Modeling online browsing and path analysis using clickstream data.Marketing Science, 23:579–595, September 2004.

[7] Stephen, Andrew T. and Galak, Jeff, The Effects of Traditional and Social Earned Media on Sales: A Study of a Microlending Marketplace (April 4, 2012). Journal of Marketing Research, 49 (October).



π is the relatively weight of clusters

z is the latent variable that determine the membership of X_i

B and G are transition matrixes of zero-inflated-poisson.

X_i^t is time-series endogenous variables contains the observed online and offline behavior of the customer for the t th week, where $i = 1, \dots, N$ and t= 1, … , T

Figure 1, The Bayesian network of Cluster VAR model

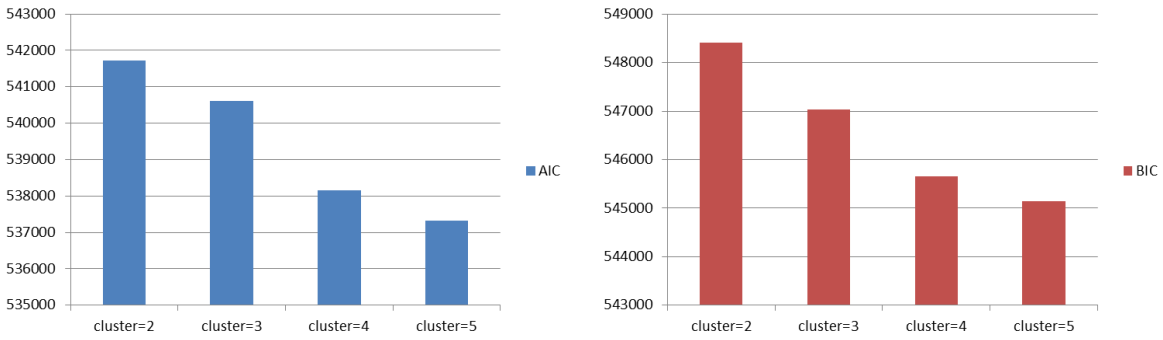


Figure 2, AIC and BIC comparison of choosing different clusters.

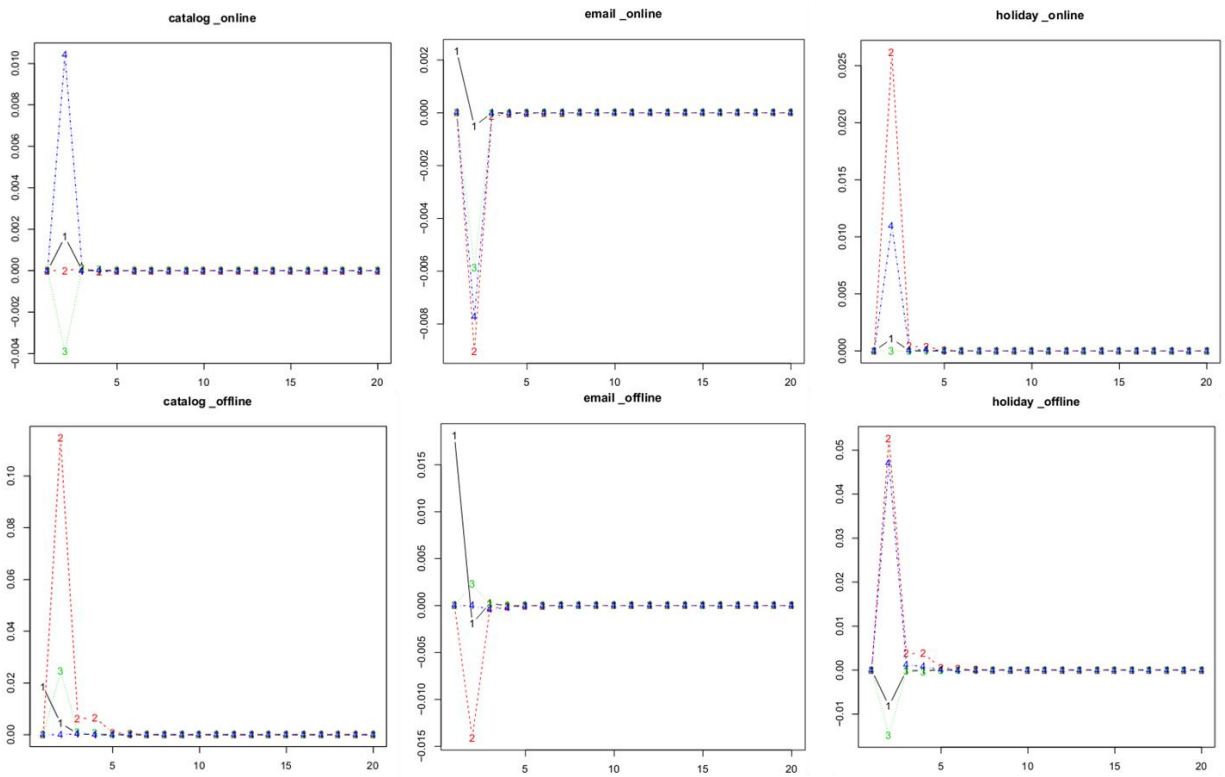


Figure 3, Impulse response of online and offline purchase based on catalog, email and holiday. Each figure illustrates the response of 4 clusters, which can be identified by the number in the response curve.

Endogenous	Description	Exogenous	Description
Browse	number of product browsed	Catalog	number of marketing campaigns received via catalog
Search	number of searches within the web site	Email	number of marketing campaigns received via email
Online	number of purchase via online channel	Holiday	Holiday dummy
Offline	number of purchase via offline channel	Promotion	the number of available promotions
Rating	number of online ratings		
Return	number of returns		

Table 1, Description of endogenous and exogenous variables in the retailer dataset.

	T=100	T=200	T=500
CVAR	0.967	0.996	0.998
K-Means	0.728	0.802	0.974

Table 2, Clustering performance comparison between CVAR and K-Means.