2017

A Cross-Cohort Changepoint Model for Customer-Base Analysis

Arun Gopalakrishnan

Eric T. Bradlow
University of Pennsylvania

Peter S. Fader
University of Pennsylvania

Recommended Citation

Gopalakrishnan, A., Bradlow, E. T., & Fader, P. S. (2017). A Cross-Cohort Changepoint Model for Customer-Base Analysis. *Marketing Science, 36* (2), 195-213. http://dx.doi.org/10.1287/mksc.2016.1007

A Cross-Cohort Changepoint Model for Customer-Base Analysis

Abstract
We introduce a new methodology that can capture and explain differences across a series of cohorts of new customers in a repeat-transaction setting. More specifically, this new framework, which we call a *vector changepoint model*, exploits the underlying regime structure in a sequence of acquired customer cohorts to make predictive statements about new cohorts for which the firm has little or no longitudinal transaction data. To accomplish this, we develop our model within a hierarchical Bayesian framework to uncover evidence of (latent) regime changes for each cohort-level parameter separately, while disentangling cross-cohort changes from calendar-time changes. Calibrating the model using multicohort donation data from a nonprofit organization, we find that holdout predictions for new cohorts using this model have greater accuracy—and greater diagnostic value—compared to a variety of strong benchmarks. Our modeling approach also highlights the perils of pooling data across cohorts without accounting for cross-cohort shifts, thus enabling managers to quantify their uncertainty about potential regime changes and avoid "old data" aggregation bias.

Keywords
changepoint, cross-cohort, hierarchical Bayesian, forecasting, customer-base analysis, customer lifetime value, reversible-jump MCMC

Disciplines
Business | Business Analytics | Business Intelligence | Management Sciences and Quantitative Methods | Marketing

A Cross-Cohort Changepoint Model for Customer-Base Analysis

Arun Gopalakrishnan
Eric T. Bradlow
Peter S. Fader*

August 1, 2016

Abstract

We introduce a new methodology that can capture and explain differences across a series of cohorts of new customers in a repeat-transaction setting. More specifically, this new framework, which we call a *vector changepoint model*, exploits the underlying regime structure in a sequence of acquired customer cohorts to make predictive statements about new cohorts for which the firm has little or no longitudinal transaction data. To accomplish this, we develop our model within a Hierarchical Bayesian framework to uncover evidence of (latent) regime changes for each cohort-level parameter separately, while disentangling cross-cohort changes from calendar-time changes. Calibrating the model using multi-cohort donation data from a non-profit organization, we find that holdout predictions for new cohorts using this model have greater accuracy – and greater diagnostic value – compared to a variety of strong benchmarks. Our modeling approach also highlights the perils of pooling data across cohorts without accounting for cross-cohort shifts, thus enabling managers to quantify their uncertainty about potential regime changes and avoid "old data" aggregation bias.

Key words: Changepoint; Cross-Cohort; Hierarchical Bayesian; Forecasting; Customer-base analysis; Customer Lifetime Value; Reversible-Jump MCMC

* Arun Gopalakrishnan is an Assistant Professor at Olin Business School, Washington University in St. Louis; Eric T. Bradlow is the K.P. Chao Professor, Professor of Marketing, Statistics, and Education, Vice-Dean and Director, Wharton Doctoral Programs, and Co-Director of the Wharton Customer Analytics Initiative; and Peter S. Fader is the Frances and Pei-Yuan Chia Professor, Professor of Marketing, and Co-Director of the Wharton Customer Analytics Initiative, at the Wharton School of the University of Pennsylvania. All correspondence concerning this manuscript should be addressed to Arun Gopalakrishnan, agopala@wustl.edu, Olin Business School, 1 Brookings Dr, Campus Box 1156, Saint Louis, MO 63130. The authors thank Yupeng Chen, Elea Feit, Bruce Hardie, Shane Jensen, Raghu Iyengar, Eric Schwartz, and Katie Yang for their feedback and suggestions, Hugh Macmullan and Paul Amos for computing and data support, as well as the non-profit organization that generously allowed us to use its data. This work was supported by an AWS in Education Grant award.

Introduction

"In times of rapid change, experience could be your worst enemy" – Jean Paul Getty

Managers in industries focused on repeat transactions such as Internet retailing, non-profit fundraising, catalog marketing, and hospitality/travel services rely on customer-base analysis to track acquisition and retention performance, and to make resource allocation decisions based on customer lifetime value (Jain and Singh 2002). Many data sets in such settings feature a time series of cohorts, a sequence in which each cohort contains a group of individuals who share a common feature (such as being acquired in the same year), and each successive cohort has a smaller number of time observations. Customer-base models have been successful in modeling *within-cohort* behaviors of interest such as customers' interpurchase times and churn propensities (Schmittlein, Morrison, and Colombo 1987), usage patterns of financial services (Allenby, Leone, and Jen 1999), and donor giving and dropout propensities (Fader, Hardie, and Shang 2010; Netzer, Lattin, and Srinivasan 2008), providing managers with tools to understand and forecast the heterogeneous behavior of any given cohort with a sufficiently large number of time observations. However, existing customer-base models remain largely silent on how to develop forecasts for new cohorts for which data are sparse, and how to allow for "regime changes" that differentially alter some cohorts' behaviors that make it inappropriate (and statistically "dangerous") to pool data across all cohorts. We illustrate the managerial questions that remain unaddressed as a result with an example.

Consider Anita, sales manager for an Internet retailer which has been in business since 2001. Each year, Anita is tasked with acquiring new customers and managing existing ones. Anita has come to view the set of customers acquired in a given year as a cohort that she can track over time, and she uses sophisticated models to understand and forecast the behavioral patterns of each cohort. However, she has little to no data on brand-new customers and is therefore unable to forecast their behavior so she turns to several colleagues for advice. Bob, the data analyst, suggests pooling data from all cohorts to make an educated guess about the behavioral patterns of the newest cohort. In contrast, Colleen, the business intelligence manager, has the intuition that a competitor who entered the market in 2006 has been

"skimming the best new customer prospects" such that the quality of newer cohorts may be significantly different from pre-2006 ones. Colleen thinks it is important to understand what changes may have occurred and when. Bob is unsure whether and how to incorporate Colleen's insights to address Anita's needs. His conundrum is accentuated by the declining number of time observations across cohorts; while he could run a stand-alone model on older cohorts observed over a long period of time, new cohorts may feature a paucity of time observations which are critical to identify any reasonable model.

From the above scenario, three managerial questions arise. First, is there evidence for discrete shifts or "regime changes" in a sequence of cohorts? Second, what aspects of cohort behavior drive these shifts and can multiple potential drivers of aggregate-level changes be teased apart, including separating cohort-level shifts from temporal effects that impact all cohorts? Third, how do we model and predict the behavior of new customer cohorts for which time observations are sparse?

In this paper, we present two cross-cohort vector changepoint ("VC") models that can provide guidance to these important managerial questions and address the limitations of existing customer-base models which implicitly assume that individuals across all cohorts can be pooled to estimate a population-level model; an assumption that ignores the potentially changing behavioral patterns across cohorts. Importantly, we highlight the challenge posed in a multi-cohort setting of robustly estimating parameters for newer cohorts and demonstrate how accurate predictions of future behavior for such cohorts can be generated using the VC models.

Using a multi-cohort data set from a U.S. non-profit organization, we compare our VC models to a number of strong benchmarks: (1) a single fully pooled model that assumes all cohorts share a base set of common parameters but allowing for heterogeneity using parametric cross-cohort effects, (2) a Hierarchical Bayesian model without changepoints (HB-0) that assumes each cohort's parameters are drawn from a common hierarchical distribution, and (3) a classical changepoint model that places restrictions (i.e. a single changepoint location

as discussed below) on regime change patterns across *all* parameters (HB-CC). The results indicate that our proposed VC models outperform the benchmarks, with improved out-of-sample prediction for new cohort donations (the problem that Anita, our hypothetical sales manager, is facing), while providing new insights into the nature of the regime changes (i.e., what aspects of behavior are changing) that would be missed by these benchmarks, while also accounting for calendar-time effects common to all cohorts. Moreover, the "HB-0" and "HB-CC" benchmarks are themselves novel to the analysis of multi-cohort customer transactional data, and our work suggests that the VC models can be a superior choice for data of this kind.

We introduce two "flavors" of VC models in this paper. The first assumes cohort parameters are shared across cohorts but only within parameter-specific regimes that break up the sequence of cohorts using changepoints ("B-VC" model). The second VC model jointly estimates parameter-specific changepoints and cohort-specific parameters using an HB framework ("HB-VC" model). The key difference between the models is within-regime heterogeneity (which exists in the HB-VC and not within the B-VC) and we discuss what each model brings to the table in empirical settings such as the one we study in this paper. We emphasize that these VC models provide a new modeling twist on the conventional approach to modeling changepoints in a multi-dimensional parameter space. In particular, the VC models allow each parameter in a cohort-level model to be drawn from a regime-specific hierarchical distribution and the number of regimes to possibly differ across parameters. In contrast, the classical changepoint paradigm requires all cohort parameters to shift simultaneously across regimes (Bai 1997; Bhattacharya 1987). Furthermore, the classical changepoint model is inefficient if only a subset of parameters is affected by regime changes.

We also note that our VC framework is agnostic to the type of behavior being modeled and is compatible with any setting in which a sequence of customer cohorts exists, individual-level data is available, and a parametric cohort-level likelihood function can be defined. At the cohort level, we use a general discrete-time model of donor attrition and transaction behavior that allows for correlation between attrition and transaction propensities and time-varying

covariates that nests the Beta-Geometric Beta-Bernoulli (BG/BB) model developed by Fader, Hardie and Shang (2010). A different cohort-level model can easily be accommodated in our framework, which can be applied to other settings such as doctor visits for age-cohorts of patients (Winkelmann 2004), online music purchasing (Fader, Hardie, and Lee 2005), effects of aging on product consumption (Rentz and Reynolds 1991), and interpurchase times for financial products (Allenby, Leone, and Jen 1999).

Our work is related to four strands of literature. First, our Bayesian approach to detecting changepoints when the number of cohort regimes is unknown builds upon the Reversible-Jump MCMC (R-J) method introduced by Green (1995), which (i) allows for considerable flexibility in defining prior distributions for the number and location of changepoints; (ii) enables the researcher to specify a broad range of allowed moves from one model to the next; and (iii) uses a generalization of the Metropolis-Hastings algorithm to estimate the posterior densities. We note that other approaches to changepoint detection, notably the product-partition method of Barry and Hartigan (1995) and Bayesian model selection approach of Carlin and Chib (1995) are popular alternatives but do not allow for the flexibility in specifying priors and model moves that the R-J method enables[1]. To date, the R-J method has had limited applications in marketing (Kim, Menzefricke, and Feinberg 2007; Ebbes, Liechty, and Grewal 2013; Narayanan 2013) due to concerns about generating efficient proposal distributions for Bayesian computation. We show that appropriate choices of model moves can result in efficient traversal of large model spaces that can result from even a short sequence (e.g., 10) of cohorts.

A second literature stream uses mixture models (Richardson and Green 1997; Allenby, Leone and Jen 1999) and semi-parametric approaches (Ansari and Iyengar 2006) to model complex distributional shapes. However, since these models do not impose structure in the

[1] The product-partition approach outlined by Barry and Hartigan (1993) is popular due to the simplicity of using data augmentation (Tanner and Wong 1987) for the presence or absence of changepoints in a sequence. However, priors are specified in a non-intuitive form of "prior cohesions" and there is limited flexibility in traversing the model space.

form of contiguous regimes of cohorts, they are not suitable for detecting regime changes, even if they are able to capture multi-modal distributions in a fully pooled model.

A third stream of literature accounts for behavioral changes across cohorts by either explicitly specifying covariates that capture systematic differences across cohorts (Schweidel, Fader and Bradlow 2008; Yang and Allenby 2003; Zhang 2008) or by introducing hierarchical priors that enable "information sharing" across spatially segmented cohorts (Hofstede, Wedel and Steenkamp 2002; Hui and Bradlow 2012). Our approach is most closely related to these hierarchical prior approaches – however, while spatial segments are on a "level playing field" in terms of observations, the temporal structure of a sequence of cohorts ordered by time of acquisition results in successive cohorts having fewer time observations, which is an additional challenge in our setting. The main issues with parametric cross-cohort effects are over-fitting due to noisy data within-sample to the detriment of out-of-sample fit, and the inability to capture potentially multiple *sharp* transitions in cohort parameters parsimoniously. Our VC model flexibly allows for arbitrary cross-cohort patterns without imposing high-degree polynomial effects (as per one of our benchmarks above) that may predict implausible parameters for future cohorts.

Finally, we use a general discrete-time model of donor attrition and transaction behavior that nests the BG/BB model (Fader, Hardie and Shang 2010) to characterize cohort-level donation behavior, linking our work to the literature on "Buy 'til you die" (BTYD) models in non-contractual settings. Our approach provides a natural extension to such models by avoiding the extreme options of either estimating one model by fully pooling multi-cohort data (as Fader, Hardie, and Shang 2010 did) or estimating cohort-specific models in isolation (which is virtually impossible for newer cohorts), while allowing for correlation between attrition and transaction propensities and time-varying covariates (e.g. calendar time effects).

The remainder of this paper is organized as follows. In Section 1, we present the model and the "technology" to estimate the posterior distributions of interest. We then discuss model

identification in Section 2, and empirical results in Section 3. We conclude in Section 4 with managerial implications and suggestions for future work. Technical appendices contain details and choices made to implement our computational approach that will allow for replication by other researchers interested in these methods.

1.0 Model Development

"You never change things by fighting the existing reality. To change something, build a new model that makes the existing model obsolete" – Richard Buckminister Fuller

In this section, we specify the vector changepoint model starting with the cohort-level discrete-time model of donor attrition and transaction behavior in Section 1.1, followed by the changepoint model that drives the assignment of cohorts to regimes in Section 1.2. Since the R-J method is less well-known compared to the Metropolis-Hastings and Gibbs samplers, we provide a brief description of how it is used to estimate the model in Technical Appendices A and B.

1.1 Cohort-level model

Our non-profit data set contains records of donations made over the history of each donor's association with the organization. In addition, we were able to obtain donor zip code data that may provide relevant information regarding donation behavior (e.g., distance from the organization's headquarters, median household income and household size of zip code) and demonstrate how our cohort-level model (described below) can allow for time-invariant covariates. In the absence of covariates, the panel data of donation incidence across donors in the same cohort (started giving in the same year) can be described by a standard BTYD model such as the BG/BB, which captures donor transaction and latent attrition propensities while allowing for heterogeneity in this non-contractual setting.

The BG/BB model, however, has the following three limitations that make it less than fully appropriate for our problem: it (1) does not allow for correlation between transaction and attrition propensities; (2) does not allow for time-invariant covariates; and (3) does not allow for time-varying covariates. A hierarchical Bayesian BTYD model akin to Abe (2009) or Singh, Borle and Jain (2009) defined for a discrete-time setting would relax the first two limitations but not the third. We therefore propose a more general discrete-time BTYD model that allows for the probability of transaction and attrition to vary by individual and time (cf. Schweidel and Knox 2013).

For customer *i* in cohort *c* observed over T_c repeat transaction opportunities (the initial transaction which identifies the cohort an individual belongs to is excluded), we define the attrition and transaction probabilities as follows, where y_{ict} is the observed transaction incidence, and z_{ict} is a latent indicator of attrition ($z_{ict} = 0$ connotes being alive, and $z_{ict} = 1$ leads to the donor entering an absorbing state of no transactions).

$$\Pr\left(z_{ict} = 1 | z_{ic,t-1} = 0, X_{ict}, \alpha_{icz}, \beta_z\right) = \Phi(\alpha_{icz} + X_{ict}\beta_z)$$

$$\Pr\left(z_{ict} = 1 | z_{ic,t-1} = 1\right) = 1$$

$$\Pr(y_{ict} = 1 | z_{ict} = 0, X_{ict}, \alpha_{icy}, \beta_y) = \Phi(\alpha_{icy} + X_{ict}\beta_y)$$

$$\Pr(y_{ict} = 1 | z_{ict} = 1) = 0$$

The individual-specific random intercepts α_{icz} and α_{icy} are distributed multivariate normal with mean μ_c and covariance matrix Σ_c while the slopes β_z and β_y are common across individuals. We call this the mixed Correlated Probit-Normal/Probit-Normal (CPN/PN) model given that the intercepts are heterogeneous and the slopes are common (hence a mixed model), which allows for the attrition and transaction intercepts to be correlated. Further, X_{ict} can contain a combination of individual-specific time-invariant covariates (e.g., demographics), cohort-specific time-invariant covariates (i.e., to model cross-cohort effects), calendar-time effects (i.e., modeling systematic changes over time affecting all cohorts) in which case we

denote the models as HB-VCT and B-VCT (T for time effects), and other time-varying covariates, as we demonstrate in our application.

For an individual with observed incidences $Y_{ic} = \{y_{ict}\} \forall\, t$ and recency (time of most recent transaction), rec_{ic}, the likelihood function is obtained by marginalizing the joint likelihood of Y_{ic} and unobserved lifetime L_{ic} as shown in equation (1). Lifetime cannot be lower than recency since a BTYD model does not allow transactions after attrition. $L_{ic} = T_c$ indicates that the individual survived till the end of the observed time periods without attrition. $L_{ic} = 0$ indicates that the individual dropped out even before facing a single repeat transaction opportunity. For a given lifetime L_{ic}, the joint likelihood is composed of three terms: the probability of attrition (not included if we observe survival), the probability of observing the transactional history over this lifetime, and the probability of *not having* dropped out prior to the duration of this lifetime.

$$p\left(Y_{ic}\middle|\alpha_{icz}, \alpha_{icy}, \beta_z, \beta_y, \{X_{ict}\}\right)$$

$$= \sum_{L_{ic}=rec_{ic}}^{T_c} \Phi\left(\alpha_{icz} + X_{ic,L_i+1}\beta_z\right)^{I(L_{ic}<T_c)} \cdot \left[\prod_{t=\min(1,L_{ic})}^{L_{ic}} \left(\Phi(\alpha_{icy} + X_{ict}\beta_y\right)^{y_{ict}} \cdot \left(1 - \Phi\left(\alpha_{icy} + X_{ict}\beta_y\right)\right)^{1-y_{ict}} \cdot \left(1 - \Phi(\alpha_{icz} + X_{ict}\beta_z)\right)\right]^{I(L_{ic}>0)} \quad (1)$$

By setting all the elements of β_z and β_y to zero, and constraining the attrition and transaction intercepts to be uncorrelated, the likelihood is analogous to the BG/BB model re-parameterized with a Probit-link function in place of Bernoulli probabilities. The benefits of the mixed CPN/PN model are three-fold. First, α_{icz} and α_{icy} can be correlated through the full covariance matrix Σ_c in the cohort-level prior, which provides the model with greater flexibility to fit data patterns where such relationships may exist. Second, time-invariant covariates can be incorporated in X_{ict}, which would then apply across each time period for an individual. Finally, X_{ict} can include time-varying covariates such as common temporal shocks affecting all cohorts over calendar time. These covariate effects can differentially impact attrition and transaction propensities through slopes β_z and β_y, and we discuss identification restrictions relating to these in Section 2.1.

The likelihood function for all the individuals in the cohort (total of I_c) as shown in equation (2), is simply the product of individual-level likelihoods.

$$p\left(Y_c \middle| \{\alpha_{icz}\}, \{\alpha_{icy}\}, \beta_z, \beta_y, \{X_{ict}\}\right) = \prod_{i=1}^{I_c} p\left(Y_{ic} \middle| \alpha_{icz}, \alpha_{icy}, \beta_z, \beta_y, \{X_{ict}\}\right) \quad (2)$$

The prior for the individual-level intercepts is the multivariate normal cohort-level distribution. The prior for the cohort-level distribution is obtained from the vector changepoint model (whose parameters we denote as M_{VC}) that we define in Section 1.2, which allows for parameter and regime-specific shrinkage based on imputed cross-cohort effects. The conjugate prior for the covariate slopes (β_z, β_y) is also multivariate normal and we use standard hyperparameters that enable a diffuse prior. We can now write down the joint posterior distribution of the mixed CPN/PN model parameters conditional on M_{VC}, Y_c and X_c.

$$p\left(\{\alpha_{icz}\}, \{\alpha_{icy}\}, \beta_z, \beta_y, \mu_c, \Sigma_c \middle| M_{VC}, Y_c, X_c\right)$$

$$\propto p\left(Y_c \middle| \{\alpha_{icz}\}, \{\alpha_{icy}\}, \beta_z, \beta_y, \{X_{ict}\}\right) \cdot p(\{\alpha_{icz}\}, \{\alpha_{icy}\} | \mu_c, \Sigma_c) \cdot p(\mu_c, \Sigma_c | M_{VC}) \cdot p(\beta_z) \cdot p(\beta_y)$$

For our setting, the main cohort-level parameters of interest are mean μ_c and covariance matrix Σ_c. To assign a multivariate normal prior to these parameters (that allows for greater flexibility than the commonly used Normal-Inverse Wishart prior), we take the log transform of the two variance terms in Σ_c, and a logit transformation of $\frac{1}{2}(\rho_{zy} + 1)$, where ρ_{zy} is the correlation coefficient.

The following parameter vector η_c: $\{\mu_{cz}, \mu_{cy}, \log(\sigma_{cz}^2), \log(\sigma_{cy}^2), logit\left(\frac{1}{2}(\rho_{zy} + 1)\right)\}$ then contains the same information as the cohort-level mean and covariance matrix, and is a convenient form which we exploit in the VC model.

1.2 Vector Changepoint model

The crux of this paper lies in the definition of $\mathrm{p}(\eta_c | \mathrm{M_{VC}})$, where $\mathrm{M_{VC}}$ represents the set of parameters in the VC models. As mentioned previously, η_c is a five-dimensional cohort-specific parameter vector that completely specifies the CPN/PN model for each cohort. We start with

how these priors are defined for two benchmark models (B-PE and HB-0), to build intuition for how the VC models differ from these.

Let η_c have a degenerate probability distribution with all its mass at μ_g. This corresponds to a special case of the fully pooled B-PE model with cross-cohort effects absent, where $\mathrm{M}_{\mathrm{BPE}} = \{\mu_g\}$. This constrains all cohorts to be represented by an identical CPN/PN model. In the HB-0 model, $\eta_c \sim MVN(\mu_g, \Sigma_g)$, where Σ_g is the covariance matrix capturing cross-cohort heterogeneity while assuming that all cohort-specific parameter vectors η_c are draws from the same distribution. This is the standard HB model that would be fit using "Bayesian shrinkage 101".

The intent of the VC models is to allow each parameter *d* in η_c to have its own regime structure and therefore pool with "relevant peers" – i.e., parameter *d*'s prior will depend on the number of changepoints $K_d \in \{0, \dots, N-1\}$ in a sequence of *N* cohorts, the *starting locations* of any changepoints Q_d and the set of parameters Ω_d associated with each of the $(K_d + 1)$ regimes (hereafter interchangeably referred to as a block). The set $\{K_d, Q_d, \Omega_d\}$ then defines the changepoint model for the d^{th} parameter of the cohort sequence. When Ω_d is a set of univariate normal distribution parameters (i.e., cohort parameters are i.i.d draws from the regime-specific distribution), we refer to the model as HB-VC to reflect Bayesian shrinkage at the cohort-level. When Ω_d is a set of parameters which are assumed to be constant within a regime, we refer to the model as B-VC to reflect that there is no cohort-specific shrinkage. Both are vector changepoint models, with the B-VC model being an important special case of the HB-VC model that may perform well under sparse-data situations where the identification of regime changes are possible, but the exact identification of heterogeneous parameters within a block is less so.

In the HB-VC model, we allow for dependence between cohort-level parameters in η_c through a correlation matrix *R* that is shared across regimes. Two points are important to note regarding this correlation matrix: (1) allowing the correlation matrix to be regime-specific will require aligning $K_d, Q_d \equiv K, Q \; \forall d$, which translates to the HB-CC classical changepoint model which forces synchronized regime changes across parameters, and (2) this correlation captures

the relationship *between* pairs of cohort parameters such as the mean attrition and transaction propensity in a cohort, and *is* entirely different from the individual-level correlation of attrition and transaction intercepts *within* a cohort described in Section 1.1. The HB-VC model may therefore be relatively parsimonious compared to the classical changepoint model as estimating a large set of correlation coefficients reliably *by regime* is challenging even for a moderate-sized cohort sequence.

It is straightforward to see that $K_d = 0 \; \forall \; d, \Omega_d = \{\mu_{gd}, \Sigma_{gd}\}$ collapses the HB-VC to the HB-0 model. Unlike the HB-0 model, the multivariate normal prior for η_c is constructed element-by-element, with mean and variance of each cohort parameter η_{cd} depending on the regime membership b_{cd} of this parameter. Similar to the HB-0 model, the correlation matrix R denoting the relationships between cohort parameters is common across regimes.

In the HB-VC model, $\eta_c \sim MVN\left(\begin{bmatrix} \mu_{g_1,b_{c1}} \\ \mu_{g_2,b_{c2}} \\ \mu_{g_3,b_{c3}} \\ \mu_{g_4,b_{c4}} \\ \mu_{g_5,b_{c5}} \end{bmatrix}, S_c R S_c\right)$, where S_c is a diagonal matrix whose diagonal elements are the standard deviations of normal distributions that serve as priors for cohort *c*'s parameters and $\mu_{g_d,b_{cd}}$ is the mean of the b_{cd}-th block for parameter *d*. $\mathrm{M_{VC}} \equiv \{\{K_d, Q_d, \Omega_d\} \forall d, R\}$ comprises the vector of changepoint models and the correlation matrix R. We illustrate with an example of a data generating process with differing regime structures by parameter in Figure 1. As can be observed, no two cohorts need have the same multivariate normal prior unless all of their parameter-specific block memberships are exactly the same. Further, even though cohorts have different priors, the elements corresponding to parameters with just one regime are the same across these priors. If a changepoint is found at cohort 6, as illustrated for parameter 1, cohorts are partitioned into two regimes (cohorts 1 through 5, and cohorts 6 through 10). Inference for all cohorts pre- and post-changepoint are affected by multiple regimes since newer cohorts do not share information with older cohorts and vice versa. Importantly, the changepoint locations induce multiple regimes as opposed to either a cohort-only model or a single-regime hierarchical model (HB-0).

[INSERT FIGURE 1 HERE]

We now define hyperpriors for the parameters contained in $\mathrm{M_{VC}}$. We choose a uniform prior over K_d's support (from 0 to *N-1*) and locations of changepoints Q_d are also assumed non-informative conditional on K_d. In this manner, we let the data "speak" as to the locations and number of changepoints; however, in situations where informative priors exist they are easily incorporated. Ω_d consists of $K_d + 1$ normal distributions with corresponding means and variances. Each mean-variance pair is assumed to be Normal-Inverse-Chisquared (NIX) distributed for conjugacy, with relatively non-informative yet proper hyperpriors (Gelman 2006). The prior on correlation matrix R is defined such that the marginal distribution for each of the correlation coefficients is uniform (Barnard et al 2000). It follows that

$$p(\mathrm{M_{VC}}) \propto p(R) \prod_{d=1}^{D} p(\Omega_\mathrm{d}|K_d, Q_d) p(Q_d|\, K_d) p(K_d),$$

where $p(R)$ is the Barnard marginally uniform prior on the correlation coefficients, $p(\Omega_\mathrm{d}|K_d, Q_d)$ is the product of NIX distributions depending on the number and location of changepoints, $p(Q_d|K_d) = \binom{N-1}{K_d}^{-1}$, and $p(K_d) = \frac{1}{N}$. For our empirical setting, we focus on a more parsimonious nested version of the full HB-VC model in which R is set to the identity matrix, and also estimate the B-VC model for comparison[2]. Alternative specifications that lie in between the HB-VC and HB-CC models are possible where synchronization of regime changes in subsets (e.g., pairs) of the parameter space are specified; this offers an important direction for future research.

In equation (3), we specify the joint posterior distribution of the HB-VC and cohort-level model parameters that we seek to estimate.

[2] The B-VC model requires a more restrictive prior to avoid the boundary condition of fitting each cohort in its own regime (since parameters are constant within a regime). Hence, we use a prior that provides an upper bound on the number of changepoints and a lower bound on regime length, as discussed in Technical Appendix B. The HB-VC fits each cohort with its own unique parameter vector (by design) and does not require these binding constraints.

$$p(\{\{\alpha_{icz}\},\{\alpha_{icy}\},\eta_c\}\ \forall c,\beta_z,\beta_y,\mathrm{M_{VC}}|\{Y_c,X_c\}\ \forall c)$$

$$\propto \prod_{c=1}^{N} p\left(Y_c\middle|\{\alpha_{icz}\},\{\alpha_{icy}\},\beta_z,\beta_y,\{X_{ict}\}\right)\cdot p(\{\alpha_{icz}\},\{\alpha_{icy}\}|\eta_c)\cdot p(\eta_c|M_{VC})\cdot p(\mathrm{M_{VC}})\cdot p(\beta_z)\cdot p(\beta_y)$$

(3)

There are four layers for estimation inherent in the above specification: (i) the hierarchical changepoint model and its parameters, which we draw using Reversible-Jump MCMC methods, (ii) the cohort-specific parameter vector η_c, which we draw using a random-walk Metropolis-Hastings sampler for each cohort, (iii) individual-level intercepts within each cohort, which we draw using a partially collapsed Gibbs sampler (Park and van Dyk 2009) that augments additional random variables for computational efficiency and better mixing, and (iv) the common slopes for covariates, drawn using a Gibbs sampler. Technical Appendix A contains computational details of how to implement our HB-VC model (including the full model that entails an additional layer to estimate the correlation matrix *R*).

1.3 Summary of Models

A summary of the models we feature in our empirical analysis is given in Table 1. The HB-VCT and B-VCT models allow for regime changes in *cohort* time as well as calendar-time effects. The HB-VC model is equivalent to HB-VCT with calendar-time effects turned off. B-VC is a special case of HB-VC with cohort parameters constant within regimes. HB-0 is equivalent to HB-VC with regime changes in *cohort* time turned off. HB-CC allows for changepoints in cohort time *but* constrains them to be the same across cohort-level parameters (i.e. a classic changepoint model). B-PET allows for parametric cohort and calendar-time effects, and B-PE is identical to B-PET with calendar-time effects turned off. However, B-PET and B-PE only allow for these effects on the attrition and transaction mean propensities (see equation 1). This provides a rich set of models to understand which model features improve performance.

All models include demographic covariates affecting attrition and transaction propensities. While we estimate our models using common slopes for the covariates across

cohorts, cohort-specific slopes could be identified, if for instance, individual-level marketing variables were available that provided variation within a cohort above and beyond that driven by parameter heterogeneity. In this case, cohort-specific slopes could be straightforwardly incorporated into η_c and the changepoint model framework, and can be explored in future research with data sets including such covariates.

[INSERT TABLE 1 HERE]

It is important to note that the HB-VCT and B-VCT models are able to separate regime changes in *cohort time* from calendar-time effects, which capture an alternate source of non-stationarity that affects customers in all cohorts over time. When all or a subset of newer cohorts' parameters are drawn from a different distribution than older cohorts, the VC models will estimate the number and locations of these regime changes, along with the distributions that best describe each regime. Older cohorts' parameters do not experience changes due to a regime shift in cohort time. These parameters (governed by a CPN/PN model) already capture within-cohort non-stationary behavior through the absorbing attrition state and calendar-time effects (which we capture using linear, quadratic and log-time terms similar to Khan et al. 2009). An interesting avenue for future research is to also allow for changepoints within cohorts over time (e.g., Fader et al. 2004), but in most cases would require much longer time series data.

What types of data characteristics would indicate the need for the VC models? Observed aggregate cohort behaviors over time (e.g., number of transactions, recency-frequency table) can be compared across cohorts after appropriate normalization for cohort size. Should there be considerable variation in these patterns across cohorts, the VC models would be the ideal approach to parse out what underlying regime changes may account for these differences. Other models such as HB-0 do not allow for sufficient flexibility to capture richer cross-cohort changes, and models such as B-PE suffer from out-of-sample generalization issues due to the parametric form of cross-cohort effects. Even if aggregate patterns do not indicate cross-cohort changes, there may be underlying regime changes due to heterogeneity within a cohort that "cancel out" in aggregate data. Thus, visual inspection is not a sufficient

condition to fit a simpler model. Further, the VC models are parsimonious in that changepoints are not fit if there is insufficient signal in the data. If no changepoints are found across all parameters, the HB-VC collapses to the HB-0 model, and the B-VC model collapses to a fully pooled model (i.e., B-PE without parametric cross-cohort effects).

2.0 Model Identification

We now discuss identification of the cohort-level CPN/PN model with common slopes (Section 2.1), and the hierarchical vector changepoint model (Section 2.2) given our interest in predicting the behavior of cohorts with very limited longitudinal data or a new cohort with no observed repeat transactional data at all. We then discuss the empirical recovery of the HB-VCT model using simulated data in Section 2.3.

2.1 Identification of CPN/PN model with common slopes

We separate the identification into two parts – first the identification of the CPN/PN model without covariates, for which the recency-frequency (RF) table serves as sufficient statistics (Fader, Hardie, and Shang 2010). We then discuss the identification of covariate coefficients (which then relies upon the relationship between individual-level parameters and covariates).

For a given cohort observed over T_c repeat observations, $J_c = \frac{T_c(T_c+1)}{2} + 1$ unique recency-frequency combinations exist. Identification is established if model parameters μ_c, Σ_c have a one-to-one mapping with the probability distribution over these J_c recency-frequency patterns; in other words, $\Delta_{\mathrm{j}}(\mu_1, \Sigma_1) = \Delta_{\mathrm{j}}(\mu_2, \Sigma_2) \rightarrow \mu_1 = \mu_2, \Sigma_1 = \Sigma_2$, where Δ_{j} is the transformation mapping μ_c, Σ_c to $p\left(RF_j \middle| \mu_c, \Sigma_c\right)$ and is defined as

$$\Delta_{\mathrm{j}}(\mu_c, \Sigma_c) = \int p\left(RF_j \middle| \alpha_{icz}, \alpha_{icy}, \beta_z = 0, \beta_y = 0, \{X_{ict}\}\right) \cdot p(\alpha_{icz}, \alpha_{icz} | \mu_c, \Sigma_c) \cdot d\alpha_{icz} \cdot d\alpha_{icz}$$

We first note that identification requires that $J_c \geq 5$, since we have to estimate five parameters for the CPN/PN model without covariates, which implies that $T_c \geq 3$ is required. As an illustration take $T_c = 1$ which implies that there is only a single transaction period and thus

the only outcomes are binary (0 or 1 repeat transactions). By turning off attrition (let the attrition mean approach negative infinity) and transaction-rate heterogeneity (let transaction-rate variance approach zero), we can exactly explain the data with the transaction-rate mean alone. Adding these additional parameters leads to severe under-identification.

Suppose $\Delta_j(\mu_1, \Sigma_1) = \Delta_j(\mu_2, \Sigma_2)\ \forall\ j$ for two sets of model parameters and $J_c \geq 5$. This would then imply that the multivariate normal mixing distributions $p(\alpha_{icz}, \alpha_{icz}|\mu_1, \Sigma_1)$ and $p(\alpha_{icz}, \alpha_{icz}|\mu_2, \Sigma_2)$ are identical, which is only possible if $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$. Thus, each set of CPN/PN model parameters generates a unique probability distribution of RF patterns and a given set of observed RF moments coincides with a unique set of model parameters.

Adding covariates means that the recency-frequency table of counts no longer serves as sufficient statistics. Slopes for time-invariant covariates such as demographics would allow for changes to an individual's propensities and are identified through the covariation between observed covariates and imputed behavior. For instance, if customers with higher incomes were less likely to drop out, the data should reflect such patterns. Because we allow for differential effects on transaction and attrition propensities, it is possible that controlling for these covariates may reduce variances and correlation of the unobserved heterogeneity in intercepts.

Identification of slopes of time-varying covariates is considerably more nuanced in the case where these covariates do not vary across individuals in a cohort. For instance, suppose we wish to add (as we do in HB-VCT and B-VCT) a parametric calendar time effect to capture "common shocks" over time to all customers' propensities in that cohort. The absence of cross-sectional variation within a cohort makes it challenging (but not impossible) to pin down calendar time effects compared to the (already latent) attrition propensity. The inclusion of multiple cohorts to estimate common calendar time effects does provide variation due to differing cohort birth times, and attrition rates and helps identify such effects. Hence, we use common slopes across cohorts.

We discuss empirical recovery of the CPN/PN model with common slopes (essentially the B-PE model) using simulated data in Technical Appendix C. The results show that cross-cohort and calendar-time effects can be teased apart in a sequence of ten cohorts (akin to our data), along with the CPN/PN parameters.

2.2 Identification of regime changes using the changepoint model

Our VC algorithm searches (in our case probabilistically samples) over the space of possible models ($\{K_d, Q_d, \Omega_d\} \forall d$) to identify the ones which best fit the data. The factors driving identification are conceptually similar to those that can distinguish models in a likelihood ratio test for model selection (Vuong 1989). For example, consider the null hypothesis as $K = 0$ (a single regime) and what in the data would drive the selection of models where $K > 0$. If we postulate two regimes, a large difference between the means of the two regimes combined with low within-regime "noise" or variance would increase the odds of correctly rejecting the null hypothesis, akin to classical signal-to-noise ratio (SNR) research (e.g., Box 1988; Pelli and Farell 1999). In addition, the length of the second regime and within-regime variance can influence whether the cohort parameter values in the new regime are interpreted as a one-off statistical blip or a discrete shift.

We demonstrate empirical parameter recovery by simulating 10-cohort data sets in which we manipulate the number of regimes by parameter (either 1 or 2), length of second regime, within-regime variance (same in both regimes for convenience) and size of the mean shift, and estimate the HB-VC parameters from the generated data. The cohort-level parameters are held fixed for the MCMC estimation, in which we take 50,000 draws using the R-J method discarding the first 10,000 as burn-in. Only the first parameter has a regime shift, and the other four parameters are drawn from single regimes to contrast the recovery of true versus possibly spurious changepoints. We then compute the average posterior probabilities across 25 runs based on the true model parameters.

From Figures 2a and 2e, we find that the SNR (defined here as the ratio of across-regime mean-shift to within-regime standard deviation) needs to be high enough to recover a true changepoint. Above an SNR of roughly 4, a true changepoint is detected with high probability

even for short new regimes. The patterns are similar for low (0.1) and high (0.75) mean shift, indicating that SNR is the primary driver for empirical recovery. For SNRs below 3, the probability of a true changepoint is attenuated and therefore downwardly biased. Importantly, this conservatism helps avoid false positives when there is no changepoint. In Figures 2d and 2h, we show that for a parameter with no changepoints, the probability of a single regime is recovered cleanly (> 90%) even for low SNRs. SNR also impacts recovery of changepoint location – at high SNRs, a true changepoint location is practically perfectly recovered, and this gets attenuated as SNR decreases, as shown in Figures 2c and 2g.

[INSERT FIGURE 2 HERE]

These results indicate that the model is able to pick up changes even for short new regimes, as long as the SNR is high enough, which bodes well for "early detection" of regime changes in actual datasets.

Finally, correctly identifying the number and locations of changepoints leads to recovery of the normal distribution parameters in each cohort-regime, since the correct parameters will be pooled to estimate the mean and variance of each distribution.

2.3 Empirical recovery of HB-VCT model using simulated data

Section 2.1 establishes the conditions to identify cohort-level parameters and discusses that cohort- and calendar-time effects can be teased apart in a sequence of cohorts. Section 2.2 demonstrates that regime changes underlying a sequence of cohorts can be accurately recovered as long as SNR is high enough, by holding the cohort-level parameters as knowns in the analysis. In this section, our focus is to demonstrate empirical recovery of the HB-VCT model using simulated data.

We simulate an eleven-cohort sequence with a changepoint for one of the parameters at the seventh cohort. We then run the HB-VCT model on this data (detailed in Technical Appendix D), holding out the final cohort and making out-of-sample transaction predictions for each cohort (similar to our empirical setting). For cohorts included in model estimation, the HB-VCT model computes posterior distributions for each cohort's parameters, which are used to

compute the posterior predictive distributions for the number of transactions. Mean Absolute Percentage Error (MAPE) calculations were performed at each iteration of the MCMC sampler and averaged across iterations to yield model performance measures for each cohort. Note that the posterior distribution for the held-out cohort is identical to the hierarchical prior distribution since its data are not used in model estimation.

As we show in Technical Appendix D, the HB-VCT model recovers the parameters of older and newer cohorts while separating out calendar-time effects, and has accurate in-sample and out-of-sample predictions for all cohorts, including the held-out eleventh cohort.

3.0 Empirical Analysis

In this section, we describe our data set (Section 3.1), discuss the criteria used for model convergence and fit assessment (Section 3.2) and present the results (Section 3.3).

3.1 Dataset

The dataset consists of donation records of individual donors from a U.S. non-profit public television broadcaster. This organization regularly runs television campaigns in which a special show is broadcast (e.g., Andrea Bocelli's Love in Portofino), which contains many breaks for the station to request donations that can yield a souvenir item related to the show. These campaigns support a significant portion of the operating budget, in contrast to competing cable channels that do not solicit for funds to operate the channel in the same way.

In addition, we were able to obtain donor zip code data that we were able to match with 1990 and 2000 census data to define distance from the organization's headquarters, median household income and household size as potentially relevant covariates. We track the longitudinal giving history of donors who made their first donation between Jan 1, 1990 and Dec 31, 2000, where the history is available from the time of acquisition through Dec 31, 2006. Due to the annual cycle of fundraising programs, we divide donors into a sequence of 11 yearly cohorts from 1990 through 2000. Each individual within a cohort is then observed for $T_c =$

(2006 – *YearAcquired*) repeat observations, where each observation is annual donation incidence. We define a donor to have given if they made at least one donation in a calendar year. Since this is a non-contractual setting, we do not observe a donor's decision to churn, and only have access to transactional patterns. We randomly sampled 33% of the donors in each cohort for estimation, yielding a total of 56,423 donors across all cohorts. The breakdown of their cohort membership is given in Table 2.

[INSERT TABLE 2 HERE]

For the remainder of this analysis, we focus on the observed data at the end of year 2002, at which point the final (year 2000) cohort would have had just two repeat transaction opportunities. This presents identification challenges as discussed in Section 2 and severe forecasting challenges for Anita, the hypothetical sales manager described in the Introduction. In Figure 3, we show the percentage of active donors (those who gave in a given year) by cohort (normalized by cohort size for ease of comparison) up through 2002 which indicates the differing amounts of longitudinal data available for each cohort and the differing drop-offs in active donors across cohorts. For instance, less than 16% of the original cohort is still active for all but the 2000 cohort. As discussed in Section 2.1, variation in cohort birth times enables the identification of any possible calendar-time effects.

[INSERT FIGURE 3 HERE]

In Figure 4, we provide a different view of the same information, comparing cohorts since their year of acquisition, which more clearly highlights cross-cohort differences. Newer cohorts appear to have a steeper fall-off in donation incidence as compared to older cohorts. Using the classification tree approach of Schwartz et al. (2014), the BG/BB model (which we extend to include covariates in the CPN/PN model) best fits the characteristics of each cohort's data as compared to other options such as Hidden Markov Models suggesting forward transitions to a non-giving ("death") state without periods of returning activity. One aspect of the cohort transaction curves, however, is worth noting as unusual – for many cohorts, the number of transactions in the most recent time period (2002) is flat or even slightly higher than the previous time period (2001). Latent attrition models such as CPN/PN or BG/BB would not fit

this particular pattern because the number of active donors within a cohort must necessarily decrease over time. We point this feature out because of the implications if an analyst used data up to 2001 to forecast future time periods of existing and new cohorts – i.e., any latent attrition model will under forecast the actual transaction activity as it would never predict a flat or slight increase in this measure. We study the effect this would have on model estimation as a robustness check in Section 3.3.2.

[INSERT FIGURE 4 HERE]

Whether the overall patterns in Figure 4 can be attributed to multiple cohort regimes, a single regime with "noise," or a deterministic trend (e.g., a calendar-time effect) is unclear from visual observation. Our data set spans a duration in which the Telecommunications Act of 1996 was passed, which increased access to a larger number of television channels for consumers, thereby potentially affecting donation behavior to the public television broadcaster. In addition, prime time household ratings for public television (percentage of US households viewing) fell from 2.3% in 1990 to 1.4% in 2006, a drop of 39%[3]. It is therefore of particular interest to understand *in what ways* cohort behaviors may have changed, both from cohort to cohort as well as over the passage of calendar time for all cohorts.

3.2 Criteria for assessing model convergence and fit

We assess model convergence for the changepoint model using the non-parametric chi-squared test (Brooks, Giudici and Philippe 2003) based on the counts of number of changepoints (K) and changepoint locations (Q) from two over-dispersed MCMC chains. The null hypothesis is that both chains' counts come from the same underlying distribution, and a lack of evidence to reject the null is taken as the metric for showing convergence (i.e., p-value > 0.05).

[3] Source: http://tvbythenumbers.zap2it.com/2010/04/12/where-did-the-primetime-broadcast-tv-audience-go/47976/

We assess the convergence of cohort-level parameters (for all models) using the Gelman-Rubin test statistic (Gelman et al 2004). The draws from two over-dispersed chains are used to compute this statistic, and values less than 1.1 were used to reflect convergence.

As noted by Shirley et al (2010), model selection criteria such as DIC are well known to be problematic for complicated hierarchical models and we therefore compare the performance of the five models described in Table 1 using in-sample and out-of-sample MAPE. We compute MAPE by first drawing individual-level binary transaction patterns over time for each iteration. These are then aggregated to the cohort level to obtain the number of cohort transactions over time for a given iteration. MAPE is then calculated iteration by iteration for each cohort[4] and we report the average MAPE across iterations in accordance with prior literature (e.g., Netzer et al. 2008; Ascarza and Hardie 2013). For each cohort, we use in-sample MAPE as a "sanity check" for model fit but rely on holdout MAPE (for years 2003 – 2006) to gauge each model's predictive validity. In particular, we hold out the 2000 cohort from estimation and use years 2003 – 2006 to compare the forecasts arising from the VC models and benchmarks. Should there be evidence for changepoints, we expect the VC models to outperform the benchmarks in terms of increased accuracy for the held-out 2000 cohort.

In addition, we also conduct a "rolling time window" analysis, which is especially relevant for a multi-cohort data set. This analysis starts with the minimum number of cohorts and time observations to identify all models, and holds out the next cohort at that point in time. For example, let us assume we have data up till year 1995. The 1992 cohort will have three repeat observations, which makes it the latest cohort that can be used in model estimation (see Section 2.1 for why this is necessary for identification). We then compute MAPE using a four-period-ahead forecast for the 1993 cohort as a holdout cohort. The rolling time window arises from repeating this analysis by adding one more cohort and time period incrementally. This analysis allows the researcher to examine model performance under limited number of cohorts and time observations, and to study the effects of any data outliers on

[4] For a given MCMC iteration and cohort *c*, the MAPE is the average percentage error between the aggregated actual and predicted time series.

model performance. Given some of the data features in Figure 4, this analysis can help shed light on "what would happen" if the data set ended at 2001, instead of 2002, for example.

3.3 Results

We obtain posterior inferences using 1,000 draws obtained from the MCMC procedure outlined in Technical Appendix A by running each model for 400,000 iterations, treating the first 200,000 iterations as burn-in and thinning every 200 iterations to reduce autocorrelation in cohort parameter draws. The models' parameters converge per the tests outlined in Section 3.2.

3.3.1 Model comparison (when 1990 through 1999 cohorts are included in the data set up till time period 2002, holding out the 2000 cohort)

In Table 3, we report the in-sample MAPEs for each cohort and model. The HB models generally fit better than the non-HB models in-sample since they allow for a more flexible cohort-level model. In other words, the B-PE model allows cohorts to be different based on a parametric functional form while the HB models balance the cohort-level likelihood with the hierarchical prior. However, between the HB models, there is little difference in the model fit. All of these HB models seemingly provide sufficient flexibility in capturing in-sample data patterns for this data set. The B-VC model, as it does not allow for heterogeneity in cohort parameters within a regime, features higher in-sample MAPE than the HB models.

[INSERT TABLES 3 AND 4 HERE]

In Table 4, we compare out-of-sample MAPEs to understand how well each model generalizes for the 1990 through 1999 cohorts, and importantly for the held out 2000 cohort whose prediction depends critically on the hierarchical assumptions of how cohorts "share" information under each model. We find that the B-PE model has higher error for the 1998 and 1999 cohorts than HB-0. This arises from the parametric cross-cohort specification in B-PE while HB-0 is able to allow for more flexibility in the parameters for these cohorts. Interestingly, HB-CC and HB-0 have similar performance as HB-CC finds little evidence for a changepoint (thus more or less collapsing to HB-0). As the specification of hierarchical priors (and hyperparameters) differs between these models, the results indicate convergent validity via

robustness to the specifications. For the held out 2000 cohort, however, B-PE, HB-0, and HB-CC all deliver high error. Under the B-PE, the 2000 cohort's parameters are obtained by extrapolating the flexible cross-cohort parametric form. Under HB-0/HB-CC, the hierarchical prior reflects a common distribution all cohorts are assumed to be drawn from. Both assumptions lead to poor predictions for the 2000 cohort.

The HB-VC model improves upon the prediction of the 2000 cohort (MAPE of 22.6% versus 38 to 40% for the B-PE, HB-0 and HB-CC models), as it finds evidence for a changepoint at the 1996 cohort, thus changing the hierarchical prior that the 2000 cohort is assumed to be drawn from. In other words, this model suggests that cohorts acquired in 1996 and later are behaviorally different from pre-1996 cohorts, and thus the 2000 cohort is better estimated by reducing the influence of those older cohorts on its hierarchical prior. It should not be surprising, however, that the 1990 through 1999 cohorts' predictions are similar in error magnitude to the HB-0 model since cohort-level parameter estimates are much more influenced by the likelihood function for these cohorts (due to increased number of time points) than the hierarchical prior. The B-VC model also improves predictive performance for the 2000 cohort but has poorer performance than HB-VC for most in-sample cohorts due to its lack of within-regime heterogeneity. The inclusion of a calendar-time effect (HB-VCT) improves performance for the newer cohorts (1997 through 2000) relative to HB-VC. HB-VCT combines the benefits of regime changes across cohorts and time effects and is the "winning model" based on 2000 cohort performance. The B-VCT model in which cohort parameters are constant within a regime (as opposed to drawn from a hierarchical prior like HB-VCT) is the "next best" model in terms of the 2000 cohort's performance. However, its poorer performance for several of the other relatively new cohorts is again due to its lack of cohort-level flexibility. B-PET does not improve performance substantially over B-PE as both model cross-cohort patterns only for the attrition and transaction mean propensities, and therefore misses potential changes in other CPN/PN parameters.

Overall, the HB-VCT model provides the most general description of cohorts of customers over time, and the results demonstrate its efficacy, particularly for the newer

cohorts in a sequence. If the analyst is focused only on holdout cohort performance and wishes to run a non-HB model, the B-VCT model offers a reasonable alternative to the HB-VCT. As a robustness check, we estimated the HB-VCT model holding out both 1999 and 2000 cohorts (while using data up to year 2002 for the 1990 through 1998 cohorts) and obtain out-of-sample MAPEs of 19.0% and 12.9% for the 1999 and 2000 cohorts, while the MAPEs of other cohorts remain virtually unchanged. Across the suite of models, HB-VCT and B-VCT remain as the best performing models for both held out cohorts. Given that data from the 1999 and 2000 cohorts are completely held out from model estimation, these prediction errors suggest the robustness of the changepoint model structure uncovered from the cohort sequence. We examine the estimates from the HB-VCT model in Section 3.3.3.

3.3.2 Rolling time window analysis

As discussed in Section 3.2, it is helpful to study HB-VCT model performance under a subset of the full data set used in Section 3.3.1 – which we term rolling time window analysis. We start by assuming we have data up till year 1995 and include the 1990 through 1992 cohorts for model estimation. We then add one more year of data and one more cohort until we reach the full data set (which simply returns us to the analysis in Section 3.3.1). In Table 5, we present the out-of-sample MAPEs (always over a four-year predictive horizon) for each cohort (including the held out one).

[INSERT TABLE 5 HERE]

The bottom row of Table 5 describes the results including the first three cohorts and forecasting the fourth cohort. We find that forecasts are reasonable considering the limited amount of data used both in number of cohorts and time observations. The forecasts continue to be reasonable (and slightly improve) with the addition of successive time observations and cohorts until the 1998 and 1999 time points when the held out cohort's forecast gets worse. This finding is consistent with a changepoint that causes post-changepoint cohorts to be different from previous ones. At the 1998 and 1999 time points, there is insufficient statistical evidence for regime change and the model therefore considers all cohorts to be in the same regime (i.e., probability of a changepoint is practically zero) – which is statistically the

appropriate inference (as discussed in Section 2.2) – but therefore does not accurately capture the future evolution of the held out cohort.

When the 2001 time point is the last observation used, the probability of a changepoint at the 1996 cohort does increase. However, the MAPEs for all cohorts are substantially worse than other windows. The reason for this surprising uptick in MAPE is related to the data features discussed in Section 3.1 as relating to Figure 4. With several cohorts exhibiting an uptick in the number of transactions after 2001, the model estimated on data up to 2001 underforecasts future transactions since latent attrition models do not allow for "dead" customers to return – and are especially sensitive to the last period of transactions for inferring customer "death". Our initial analysis (which used data up till the 2001 point) yielded this pattern of MAPEs, which enabled us to find this data outlier by adding the 2002 time point to the cohorts (which is the full data set used in Section 3.3.1). As can be seen from the first row of Table 5 (which is identical to the HB-VCT row of Table 4), MAPEs are significantly improved for all cohorts while a high probability of a changepoint is inferred at the 1996 cohort (as we detail in Section 3.3.3).

The main lesson to draw from this analysis is that the rolling time window approach can be useful in multi-cohort data sets to parse out any unusual data patterns that may preclude accurate forecasts, especially those that occur at the end of a time series.

3.3.3 Model estimates

The HB-VCT model estimates a 73% probability of at least one changepoint for the correlation between attrition and transaction propensities, with the modal location being the 1996 cohort (see Figure 5). The pre-1996 cohorts have a negative correlation (regime prior mean of -0.20) between attrition and transaction propensities, meaning that a customer in these cohorts who has a low probability of "death" also has a high probability of transacting while "alive". Conversely, this also indicates the double whammy of attrition-prone customers who are also less likely to transact. The post-changepoint cohorts essentially have negligible correlation (regime prior mean of 0.01) between attrition and transaction propensities and contain relatively "less polarized" customers within the cohort. The other parameters do not exhibit

strong evidence for a changepoint in the cohort sequence. Figure 5c shows the prior and posterior means for each cohort parameter. The prior mean (dotted line) for the correlation between attrition and transaction propensities, consistent with the estimated changepoint, shows a shift at the 1996 cohort. The posterior mean for each cohort parameter incorporates information from the cohort-level likelihood function and can therefore deviate from the prior mean due to cohort heterogeneity.

[INSERT FIGURE 5 HERE]

We find that demographic covariate effects are directionally intuitive for the most part (see Table 6). Attrition propensity decreases with income – indicating that wealthier households are less likely to end their relationship with the organization. Transaction propensity decreases with household size and increases with income – indicating that smaller and wealthier households are also more likely to donate. Attrition propensity increases with distance (which is directionally intuitive) as does transaction propensity (which is slightly counter-intuitive, but may be indicative of a non-linear relationship). Importantly, the effect sizes for these demographic covariates are small, such that the other model parameters (and posterior predictive distributions) would be practically unchanged by their exclusion.

[INSERT TABLE 6 HERE]

From Table 6, we find that calendar-time effects are statistically significant for both the attrition and transaction propensities. In Figure 6, we show the net trend implied over time for each parameter. The attrition propensity trend increases with time and has a larger effect size than the transaction propensity trend, which exhibits less variation over time. The calendar-time effects suggest that customers from any cohort (old or new) are more likely to experience attrition over time above and beyond that predicted by the CPN/PN model.

[INSERT FIGURE 6 HERE]

Our empirical findings suggest two different types of effects observed in the data. First, the 1996 changepoint in the correlation between attrition and transaction propensities points to a change in the mix of donors obtained from 1996 onwards. In older cohorts, donors on

average seem more "polarized" – they are more likely to be long lifetime and frequent donors (both positively affecting CLV) or short lifetime infrequent donors. In newer cohorts, there is less evidence for such donors, with attrition and transaction behaviors largely uncorrelated. Second, attrition propensity increases over time for all cohorts. These findings highlight that changes in the regulatory landscape (1996 Telecommunications Act) and substitution from public television programming to specialized cable channels likely had an effect on the mix of acquired donors and the declining likelihood of being "alive" as donors over calendar time. An interesting area for future research is to study the mechanisms by which viewership of public television programs affect donation behaviors.

In summary, the suite of models we used to fit the multi-cohort donor data has yielded the HB-VCT as offering best out-of-sample performance for the 2000 cohort, as well as most other newer cohorts. As the HB-VCT is the most general model in our lineup, this suggests that having the flexibility to separate cross-cohort regime changes by parameter along with account for calendar-time effects offers promise as a method for future multi-cohort analyses.

4.0 Discussion and Future Work

4.1 Discussion

The comparison of the vector changepoint models with the benchmarks shows that utilizing our approach yields new insights about the regime structure across cohorts that were not previously accessible. In addition, the regime structure plays a significant role in enhancing the predictive accuracy of the new 2000 cohort's behavior in holdout data and our model helps tease apart data patterns driven by a cross-cohort changepoint versus a common calendar-time effect. Our findings suggest that simply using older cohorts as a proxy for predicting new cohorts without understanding any potential regime changes may lead to inaccurate predictions.

Returning to the managerial questions posed at the beginning of this paper, Bob the data analyst is now able to draw parallels from the non-profit dataset to his own context. Using the vector changepoint models, he is able to determine which cohorts would be relevant peers

for the new cohorts that Anita, the sales manager is interested in. He realizes that pooling data from all cohorts would likely decrease the accuracy of his predictions as older cohorts may be part of a different regime as compared to the newest cohorts. Thus, decision making based on heuristics of what is relevant can be augmented with data-driven evidence of regime change. Further, Bob is able to use the VC models to yield robust predictions of cohorts' future behaviors and tease apart calendar time effects as well.

Anita has also learned that she can examine the parameters that have shifted at changepoint locations to better understand the intuition of what cohort-level behavior is responsible for the shift. She may be able to confirm Colleen's intuition by analyzing what aspect of newer cohorts may have undergone a significant shift since the new competitor's entry, which she was previously unable to quantify. The bulk of holdout transactions (2003 – 2006) for each cohort are contributed by donors with high in-sample recency. While this validates the attrition process inherent in a BTYD model, there may not be much Anita can do to bring back the customers who have already ended their relationship with the firm. Instead, she should focus her retention efforts on customers in newer cohorts, who may still be persuaded to remain active for a longer period of time through judicious use of targeted marketing communications and inducements (e.g., Khan et al. 2009). In addition, the acquisition strategy for new cohorts may need updating – if existing strategies are ineffective in bringing in the desired portfolio of customers (e.g., Lewis 2006), new acquisition channels (e.g., mobile advertising) and incentives such as referral programs (e.g., Schmitt et al 2011) may be needed to attract better prospects.

Forecasts based on older cohorts' behavior, in the case of the non-profit dataset, would lead management to forecast less accurate donation figures as well. Management of startup companies whose valuations rely upon customer acquisition may portray a distorted view of future growth if they project new cohorts to behave like the "average" of all previous cohorts. This can have ramifications for allocating marketing budgets and in reporting projections to stakeholders. Supposing that Anita's firm (from our example) uses projected transactions to

provide guidance to investors, large inaccuracies can hurt credibility and valuation in the longer-run.

Further, managers may discover signs of saturation in the market they are targeting, perhaps due to "over-fishing;" successive new cohorts may be less attractive than older ones in which the better prospects may already be captured. Such cross-cohort shifts can be easily missed when conducting analysis at the aggregate level, such as tracking overall revenue across the entire customer base and our approach is a means to avoiding aggregation bias in multi-cohort settings.

For researchers, these findings suggest that investigating cross-cohort patterns can be a powerful tool when dealing with a dataset that has a multi-cohort structure. The vector changepoint models allow for more flexible and robust cross-cohort patterns than a model with parametric effects, can provide greater power to uncover parameter-specific regime changes than the classical changepoint model, and adds a low computational burden as compared to a no-changepoint Hierarchical Bayesian model. Our findings suggest that our most general vector changepoint model with a hierarchical Bayesian structure can be a good starting point for multi-cohort datasets given its ability to tease apart cross-cohort regime changes, cohort heterogeneity, and global calendar-time effects. In addition, the "family tree" of simpler models that can be obtained by turning off features of our most general model is also available for researchers looking only for a subset of modeling features. For example, a researcher who has strong reasons to believe there are no cross-cohort effects, could use the no-changepoint Hierarchical Bayesian model and also allow for calendar-time effects, which is a model that was not in our set of benchmarks, but nevertheless may be useful in other research settings.

Our robustness checks reveal a cautionary note in using the most general changepoint model we propose along with calendar-time effects. In the presence of data outliers at the last observed time period, we show in our data setting that the model can lead to higher prediction errors than may be expected. It is difficult to provide general guidance on detection of such data outliers as much of the issue can be dependent on what the cohort-level behavior being modeled is. However, we recommend the rolling time window analysis as one method to

detect if any given time end point leads to abnormally high prediction errors (relative to cutting off the data at other time points). This is not entirely different (conceptually) than the results of Van den Bulte and Lilien (1997) who show that the Bass model is highly sensitive to the observed diffusion in the last (few) data points of a time series.

More broadly, our approach provides a general framework for the model selection problem, which frequently arises in marketing. Instead of running thousands of possible models to determine the "best fitting one," our approach leverages Bayesian methods to deliver inferences that incorporate uncertainty in the model in a single framework. Other contexts in which there exist multiple models with varying number of parameters which could describe the data, could benefit from the broad steps we have presented in defining a Reversible-Jump MCMC algorithm that can traverse the model space efficiently and avoid the Davies problem (Davies 1987; Hansen 1996) that arises in classical hypothesis testing.

4.2 Future Work

We discuss fruitful directions in which this work can be enhanced. First, it is possible that cohorts are not necessarily related to each other in contiguous temporal blocks. Returning to the idea of a more general model of relationships amongst cohorts (e.g., Hui and Bradlow 2012), the model could allow for cohorts exhibiting similar behaviors to be grouped together, akin to a latent class model, even if this breaks the temporal sequencing. This may be interesting if new cohorts follow a cyclical pattern of resembling earlier cohorts (perhaps due to economic cycles), or if the segmentation is not based on period of acquisition. A more general segmentation of customers may not have a cohort sequence structure but there can still be underlying patterns which can be uncovered. Extending the vector changepoint framework to Bayesian clustering methods (François et al 2006; Green and Richardson 2002) while addressing the explosive increase in the number of possible groupings would enable such data structures to be appropriately modeled.

Second, the power of this modeling approach can be further exploited by richer models of customer behavior, which can allow for a range of behaviors such as timing, choice, and monetary amounts. A hybrid of the vector and classical changepoint models can then be used

to appropriately partition the priors on the full parameter space. We suggest that further work in exploring alternative specifications lying in the spectrum between vector and classical changepoint models can help uncover when each type of model is best suited for analysis.

Third, a limitation of our modeling approach is the assumption that a changepoint at a given cohort does not affect the behavioral parameters of preceding ones. That is, while we explored changepoints across acquired cohorts, there may well be changepoints that occur in time that also differ across cohorts. This is an avenue for further research by integrating cross-cohort and time- changepoints into a single model.

Fourth, marketing covariates could be incorporated at the individual level, if available, to better understand the effect of targeted marketing on attrition and transaction behaviors. In particular, our model presents statistical evidence for regime changes that can enable this non-profit organization to optimize their cohort-management strategies. An interesting area for future research would be to introduce time-varying marketing covariates (our model is general enough to accommodate these, but the data do not contain them) to help organizations understand the effectiveness of their implemented cohort-management strategies.

Cohort-specific slopes capturing response propensity can be added to the parameter space of the vector changepoint model to further generalize the model. Given that these marketing variables may be endogenously set with the customers' expected responses in mind, an appropriate model of firm-side decision making (e.g., Schweidel and Knox 2013) would need to be integrated with the vector changepoint model.

Fifth, it is possible that cross-cohort changepoints may occur at future cohorts not currently in the data set. Our model estimates changepoints based on patterns in cohorts included for estimation. Future work can investigate predicting the timing of future changepoints by incorporating an additional hierarchical layer above the changepoint model.

Sixth, a study of the mechanisms by which consumers of public television broadcasters across the U.S. may have changed as a function of the shifting television programming landscape is a substantive topic for exploration. As some of these changes are likely exogenous

from the viewpoint of donation and viewership behavior (1996 Telecommunications Act), and there may be regression discontinuities in the rates of change of local television markets (c.f. Hartmann et al. 2011), there is potential for causal inferences to be drawn about changes in consumer behavior.

The perils of pooling data without accounting for cross-cohort shifts have been highlighted in this work. We hope our findings encourage further research in uncovering new structural patterns in customer databases that help managers make the most effective use of their data.

5.0 References

Abe, M. (2009). “Counting Your Customers” One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model. Marketing Science, 28, 3, 541-553

Allenby, G.M., Leone, R.P., and Jen, L. (1999). A Dynamic Model of Purchase Timing with Application to Direct Marketing. Journal of the American Statistical Association, 94, 446, 365-374

Ansari, A., and Iyengar, R. (2006). Semiparametric Thurstonian Models for Recurrent Choices: A Bayesian Analysis. Psychometrika, 71, 4, 631-657

Bai, J. (1997). Estimation of a Change Point in Multiple Regression Models. The Review of Economics and Statistics, 79, 4, 551-563

Barry, D., and Hartigan, J.A. (1993). A Bayesian Analysis for Change Point Problems. Journal of the American Statistical Association, 88, 421, 309-319

Barnard, J., McCulloch, R., and Meng X-L. (2000). Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage. Statistical Sinica, 10, 1281-1311

Bhattacharya, P.K. (1987). Maximum Likelihood Estimation of a Change-Point in the Distribution of Independent Random Variables: General Multiparameter Case. Journal of Multivariate Analysis, 23, 183-208

Box, G. (1988). Signal-to-Noise Ratios, Performance Criteria, and Transformations. Technometrics, 30, 1, 1-17

Brooks, S.P., Giudici, P., and Philippe, A. (2003). Nonparametric Convergence Assessment For MCMC Model Selection. Journal of Computational and Graphical Statistics, 12, 1, 1-22

Budden, M., Hadavas, P., and Hoffman, L. (2008). On the Generation of Correlation Matrices. Applied Mathematics E-Notes, 8, 279-282

Carlin, B.P., and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. Journal of the Royal Statistical Society B, 57, 3, 473-484

Davies, R.B. (1987). Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative. Biometrika, 74, 1, 33-43

Ebbes, P., Liechty, J.C., and Grewal, R. (2015). Attribute-Level Heterogeneity. Management Science, 61, 4, 885-897

Fader, P.S., Hardie, B.G.S, and Huang, C-Y. (2004). A Dynamic Changepoint Model for New Product Sales Forecasting. Marketing Science, 23, 1, 50-65

Fader, P.S., Hardie, B.G.S, and Lee, K.L. (2005). "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. Marketing Science, 24, 2, 275-284

Fader, P.S., Hardie, B.G.S, and Shang, J. (2010). Customer-Base Analysis in a Discrete-Time Noncontractual Setting. Marketing Science, 29, 6, 1086-1108

François, O., Ancelet, S., and Guillot, G. (2006). Bayesian clustering using Hidden Markov Random Fields in Spatial Population Genetics. Genetics, 174, 2, 805-816.

Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. Bayesian Analysis, 1, 3, 515-533

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). Bayesian Data Analysis (2nd edition). Chapman and Hall

Green, P.J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika, 82, 4, 711-732

Green, P. J., and Richardson, S. (2002). Hidden Markov Models and Disease Mapping. Journal of the American Statistical Association, 97, 460, 1055-1070.

Hansen, B.E. (1996). Inference When a Nuisance Parameter is Not Identified Under the Null Hypothesis. Econometrica, 64, 2, 413-430

Hartmann, W., Nair, H. S., & Narayanan, S. (2011). Identifying Causal Marketing Mix Effects Using a Regression Discontinuity Design. Marketing Science, 30(6), 1079-1097.

Hastie, D. (2005). Towards Automatic Reversible Jump Markov Chain Monte Carlo. Doctoral Dissertation. University of Bristol

Hofstede, F.T., Wedel, M., and Steenkamp, J-B.E. (2002). Identifying Spatial Segments in International Markets. Marketing Science, 21, 2, 160-177

Hui, S.K., and Bradlow, E.T. (2012). Bayesian Multi-Resolution Spatial Analysis with Applications to Marketing. Quantitative Marketing and Economics, 10, 4, 419-452

Hürlimann, W. (2012). Positive Semi-Definite Correlation Matrices: Recursive Algorithmic Generation and Volume Measure, 1, 3, 137-149

Jain, D., and Singh, S.S. (2002). Customer Lifetime Value Research in Marketing: A Review and Future Directions. Journal of Interactive Marketing, 16, 2, 34-46

Joe, H. (2006). Generating Random Correlation Matrices Based on Partial Correlations. Journal of Multivariate Analysis, 97, 2177-2189

Khan, R., Lewis, M., and Singh, V. (2009). Dynamic Customer Management and the Value of One-to-One Marketing. Marketing Science, 28, 6, 1063-1079

Kim, J.G., Menzefricke, U., and Feinberg, F.M. (2007). Capturing Flexible Heterogeneous Utility Curves: A Bayesian Spline Approach. Management Science, 53, 2, 340-354

Lewis, M. (2006). Customer Acquisition Promotions and Customer Asset Value. Journal of Marketing Research, 43, 2, 195-203

Liechty, J.C., Liechty, M.W., and Müller, P. (2004). Bayesian Correlation Estimation. Biometrika, 91, 1, 1-14

Narayanan, S. (2013). Bayesian Estimation of Discrete Games of Complete Information. Quantitative Marketing and Economics, 11, 1, 39-81

Netzer, O., Lattin, J.M., and Srinivasan, V. (2008). A Hidden Markov Model of Customer Relationship Dynamics. Marketing Science, 27, 2, 185-204

Park, T., and van Dyk, D.A. (2009). Partially Collapsed Gibbs Samplers: Illustrations and Applications. Journal of Computational and Graphical Statistics, 18, 2, 283-305

Pelli, D.G., and Farell, B. (1999). Why Use Noise? Journal of the Optical Society of America, 16, 3, 647-653

Rentz, J. O., & Reynolds, F. D. (1991). Forecasting the effects of an aging population on product consumption: An age-period-cohort framework. Journal of Marketing Research, 355-360.

Richardson, S., and Green, P.J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. Journal of the Royal Statistical Society B, 59, 4, 731-792

Schmitt, P., Skiera, B., & Van den Bulte, C. (2011). Referral programs and customer value. Journal of Marketing, 75, 1, 46-59

Schmittlein, D.C., Morrison, D.G., and Colombo, R. (1987). Counting your customers: Who are they and what will they do next? Management Science, 33, 1, 1-24

Schweidel, D.A., Fader, P.S., and Bradlow, E.T. (2008). Understanding Service Retention Within and Across Cohorts Using Limited Information. Journal of Marketing, 72, 1, 82-94

Schweidel, D.A., and Knox, G. (2013). Incorporating Direct Marketing Activity into Latent Attrition Models. Marketing Science, 32, 3, 471-487

Shirley, K.E., Small, D.S., Lynch, K.G., Maisto, S.A., and Oslin, D.W. (2010). Hidden Markov Models for Alcoholism Treatment Trial Data. Annals of Applied Statistics, 4, 1, 366-395

Singh, S.S., Borle, S., and Jain, D.C. (2009). A Generalized Framework for Estimating Customer Lifetime Value when Customer Lifetimes Are Not Observed. Quant Mark Econ, 7, 2, 181-205

Tanner, M.A., and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, 82, 398, 528-540

Van den Bulte, C., & Lilien, G. L. (1997). Bias and systematic change in the parameter estimates of macro-level diffusion models. Marketing Science, 16, 4, 338-353.

Vuong, Q.H. (1989). Likelihood Ratio tests for Model Selection and Non-Nested Hypotheses. Econometrica, 57, 2, 307-333

Winkelmann, R. (2004). Health Care Reform and the Number of Doctor Visits – An Econometric Analysis. Journal of Applied Econometrics, 19, 455-472

Yang, S., and Allenby, G.M. (2003). Modeling Interdependent Consumer Preferences. Journal of Marketing Research, 40, 3, 282-294

Zhang, H.Y. (2008). Modeling Discrete-Time Transactions Using the BG/BB Model. Dissertation. University of Pennsylvania

Table 1: Summary of models

Model	Description
Single fully pooled model (B-PE)	Only one CPN/PN parameter vector across all cohorts (μ_g) plus a parametric cross-cohort effect using linear, quadratic and log components.
HB no-changepoint (HB-0)	All cohorts pool under a single regime but each cohort i has its own parameter vector η_i
HB classical changepoint (HB-CC)	All parameters share the same regime configuration and each cohort i has its own parameter vector η_i.
HB vector changepoint (HB-VC)	Each parameter has its own regime configuration and each cohort i has its own parameter vector η_i
Vector changepoint without HB (B-VC)	A special case of HB-VC where cohort parameters are constant within a regime.
Single fully pooled model with calendar-time effects (B-PET)	Only one CPN/PN parameter vector across all cohorts (μ_g) plus parametric cross-cohort and time effects using linear, quadratic and log components.
Vector changepoint without HB and with calendar-time effects (B-VCT)	Each parameter has its own regime configuration and each cohort's parameter vector η_i is a function of the regime configuration plus time effects using linear, quadratic and log components.
HB vector changepoint (HB-VCT)	Each parameter has its own regime configuration and each cohort i has its own parameter vector η_i, plus time effects using linear, quadratic and log components.

Table 2: Number of initial donors by cohort (after sampling 33% of cohort)

Cohort	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Donors	5,085	6,467	5,851	7,494	5,223	5,379	5,989	4,364	2,876	4,357	3,338

Table 3: In-sample MAPE by cohort for each model

Model	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
B-PE	4.3%	4.1%	4.7%	7.3%	3.3%	7.0%	6.4%	6.7%	6.8%	4.6%
HB-0	4.5%	3.3%	4.2%	3.5%	3.8%	4.7%	5.6%	5.6%	6.3%	4.4%
HB-CC	4.5%	3.4%	4.2%	3.4%	3.8%	4.7%	5.7%	5.5%	6.4%	4.4%
B-VC	7.6%	9.3%	7.8%	6.9%	13.2%	12.1%	6.8%	18.0%	24.9%	33.9%
HB-VC	4.5%	3.4%	4.3%	3.5%	3.7%	4.7%	5.7%	5.6%	6.5%	4.5%
B-PET	4.5%	4.5%	4.8%	7.3%	3.4%	7.0%	6.4%	6.5%	6.6%	4.4%
B-VCT	7.4%	3.6%	4.9%	5.4%	4.3%	7.7%	6.3%	5.4%	6.6%	14.9%
HB-VCT	4.6%	3.8%	4.3%	3.4%	3.7%	4.6%	5.4%	5.5%	6.5%	4.4%

Table 4: Out-of-sample MAPE by cohort for the 2003-2006 time period for each model. The 2000 cohort (last column) is shaded to denote that its data are entirely held out from model estimation.

Model	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000 (Heldout)**
B-PE	7.0%	11.0%	9.8%	5.1%	10.4%	14.5%	9.8%	10.2%	23.5%	21.2%	39.8%
HB-0	9.2%	9.5%	7.5%	6.8%	10.4%	11.6%	8.7%	12.2%	19.1%	15.4%	38.1%
HB-CC	9.2%	9.5%	7.6%	6.8%	10.5%	11.5%	9.1%	11.9%	18.1%	14.5%	39.2%
B-VC	11.6%	15.6%	10.4%	9.9%	20.5%	15.9%	6.8%	16.7%	15.0%	22.5%	25.4%
HB-VC	8.7%	9.1%	7.6%	7.0%	10.6%	12.2%	9.0%	12.7%	19.7%	15.4%	22.6%
B-PET	11.0%	14.4%	11.5%	4.6%	11.4%	15.6%	10.4%	10.5%	23.3%	20.5%	39.2%
B-VCT	8.5%	17.3%	15.6%	5.4%	12.8%	13.7%	17.7%	11.7%	17.0%	7.1%	17.2%
HB-VCT	9.3%	9.3%	8.2%	7.3%	10.2%	12.4%	8.8%	11.3%	16.5%	12.9%	15.2%

Table 5: HB-VCT four-period-ahead out-of-sample MAPE using a rolling validation time window. Grayed cells refer to cohorts that were used for model estimation and cells in yellow refer to the heldout cohort. Note that the results for last time period of 2002 exactly corresponds to the HB-VCT results in Table 4 and is replicated for ease of comparison.

			Out-of-sample MAPE by cohort										
Last time period	Last cohort used	Holdout periods	**90**	**91**	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**00**
2002	1999	2003-2006	9.3%	9.3%	8.2%	7.3%	10.2%	12.4%	8.8%	11.3%	16.5%	12.9%	15.2%
2001	1998	2002-2005	33.6%	34.7%	35.0%	34.9%	32.5%	35.6%	37.3%	39.5%	42.3%	31.1%	
2000	1997	2001-2004	17.8%	18.8%	16.6%	19.1%	17.7%	22.0%	23.8%	24.4%	26.2%		
1999	1996	2000-2003	10.3%	10.2%	10.4%	10.3%	8.9%	10.6%	12.2%	24.7%			
1998	1995	1999-2002	12.6%	12.3%	13.9%	16.8%	15.4%	13.4%	21.8%				
1997	1994	1998-2001	13.6%	11.2%	11.6%	12.1%	13.6%	22.3%					
1996	1993	1997-2000	13.2%	14.1%	17.4%	18.2%	22.5%						
1995	1992	1996-1999	13.2%	12.8%	13.1%	22.6%							

Table 6: HB-VCT model slope parameter estimates

Parameter	**Posterior mean (standard error)**
Attrition: linear time	-0.34 (0.71)
Attrition: quadratic time	0.03 (0.03)
Attrition: log time	4.71 (2.06)
Trans: linear time	-1.58 (0.27)
Trans: quadratic time	0.07 (0.01)
Trans: log time	4.45 (0.69)
Attrition: ln(distance+1)	0.09 (0.01)
Attrition: HH Size	0.01 (0.01)
Attrition: ln(HHIncome+1)	-0.22 (0.02)
Trans: ln(distance+1)	0.06 (0.01)
Trans: HH Size	-0.04 (0.01)
Trans: ln(HHIncome+1)	0.18 (0.02)

Figure 1: Example of priors from the vector changepoint model. No two cohorts need have the same multivariate prior unless their parameter-specific block memberships are identical. Cohorts 1 through 5 have identical priors but the prior for Cohort 6 differs from the first five cohorts due to the changepoint in parameter 1.

<table>
<tr><td>Coh. 1</td><td>Coh. 2</td><td>Coh. 3</td><td>Coh. 4</td><td>Coh. 5</td><td>Coh. 6</td><td>Coh. 7</td><td>Coh. 8</td><td>Coh. 9</td><td>Coh. 10</td></tr>
<tr><td colspan="5">Par 1: Regime 1 - $N(\mu_{g11}, \sigma^2_{g11})$</td><td colspan="5">Par 1: Regime 2 - $N(\mu_{g12}, \sigma^2_{g12})$</td></tr>
<tr><td colspan="8">Par 2: Regime 1 - $N(\mu_{g21}, \sigma^2_{g21})$</td><td colspan="2">Par 2: Regime 2 - $N(\mu_{g22}, \sigma^2_{g22})$</td></tr>
<tr><td colspan="10">Par 3: Regime 1 - $N(\mu_{g31}, \sigma^2_{g31})$</td></tr>
<tr><td colspan="10">Par 4: Regime 1 - $N(\mu_{g41}, \sigma^2_{g41})$</td></tr>
<tr><td colspan="10">Par 5: Regime 1 - $N(\mu_{g51}, \sigma^2_{g51})$</td></tr>
</table>

$K_1 = 1, Q_1 = 6; K_2 = 1, Q_2 = 9, K_3 = 0, K_4 = 0, K_5 = 0$. Let R be a valid correlation matrix.

$$\eta_1, \eta_2, \eta_3, \eta_4, \eta_5 \sim MVN\left(\begin{bmatrix}\mu_{g11}\\ \mu_{g21}\\ \mu_{g31}\\ \mu_{g41}\\ \mu_{g51}\end{bmatrix}, \begin{bmatrix}\sigma_{g11} & 0 & 0 & 0 & 0\\ 0 & \sigma_{g21} & 0 & 0 & 0\\ 0 & 0 & \sigma_{g31} & 0 & 0\\ 0 & 0 & 0 & \sigma_{g41} & 0\\ 0 & 0 & 0 & 0 & \sigma_{g51}\end{bmatrix} R \begin{bmatrix}\sigma_{g11} & 0 & 0 & 0 & 0\\ 0 & \sigma_{g21} & 0 & 0 & 0\\ 0 & 0 & \sigma_{g31} & 0 & 0\\ 0 & 0 & 0 & \sigma_{g41} & 0\\ 0 & 0 & 0 & 0 & \sigma_{g51}\end{bmatrix}\right)$$

$$\eta_6, \eta_7, \eta_8, \sim MVN\left(\begin{bmatrix}\mu_{g12}\\ \mu_{g21}\\ \mu_{g31}\\ \mu_{g41}\\ \mu_{g51}\end{bmatrix}, \begin{bmatrix}\sigma_{g12} & 0 & 0 & 0 & 0\\ 0 & \sigma_{g21} & 0 & 0 & 0\\ 0 & 0 & \sigma_{g31} & 0 & 0\\ 0 & 0 & 0 & \sigma_{g41} & 0\\ 0 & 0 & 0 & 0 & \sigma_{g51}\end{bmatrix} R \begin{bmatrix}\sigma_{g12} & 0 & 0 & 0 & 0\\ 0 & \sigma_{g21} & 0 & 0 & 0\\ 0 & 0 & \sigma_{g31} & 0 & 0\\ 0 & 0 & 0 & \sigma_{g41} & 0\\ 0 & 0 & 0 & 0 & \sigma_{g51}\end{bmatrix}\right)$$

$$\eta_9, \eta_{10} \sim MVN\left(\begin{bmatrix}\mu_{g12}\\ \mu_{g22}\\ \mu_{g31}\\ \mu_{g41}\\ \mu_{g51}\end{bmatrix}, \begin{bmatrix}\sigma_{g12} & 0 & 0 & 0 & 0\\ 0 & \sigma_{g22} & 0 & 0 & 0\\ 0 & 0 & \sigma_{g31} & 0 & 0\\ 0 & 0 & 0 & \sigma_{g41} & 0\\ 0 & 0 & 0 & 0 & \sigma_{g51}\end{bmatrix} R \begin{bmatrix}\sigma_{g12} & 0 & 0 & 0 & 0\\ 0 & \sigma_{g22} & 0 & 0 & 0\\ 0 & 0 & \sigma_{g31} & 0 & 0\\ 0 & 0 & 0 & \sigma_{g41} & 0\\ 0 & 0 & 0 & 0 & \sigma_{g51}\end{bmatrix}\right)$$

Figure 2: Identification of HB-VC model using a simulated sequence of 10 cohorts (SNR = ratio of across-regime mean shift to within-regime standard deviation). A true changepoint is detected with high probability when SNR is large enough as shown in Fig 2a. The algorithm is conservative and prefers a false negative at low SNRs (Fig 2b) to ensure a low probability of false positives (Fig 2d). This also holds for changepoint location, and for various lengths of a new regime (Fig 2e to 2h). The magnitudes of low and high mean shifts are 0.1 and 0.75 respectively.

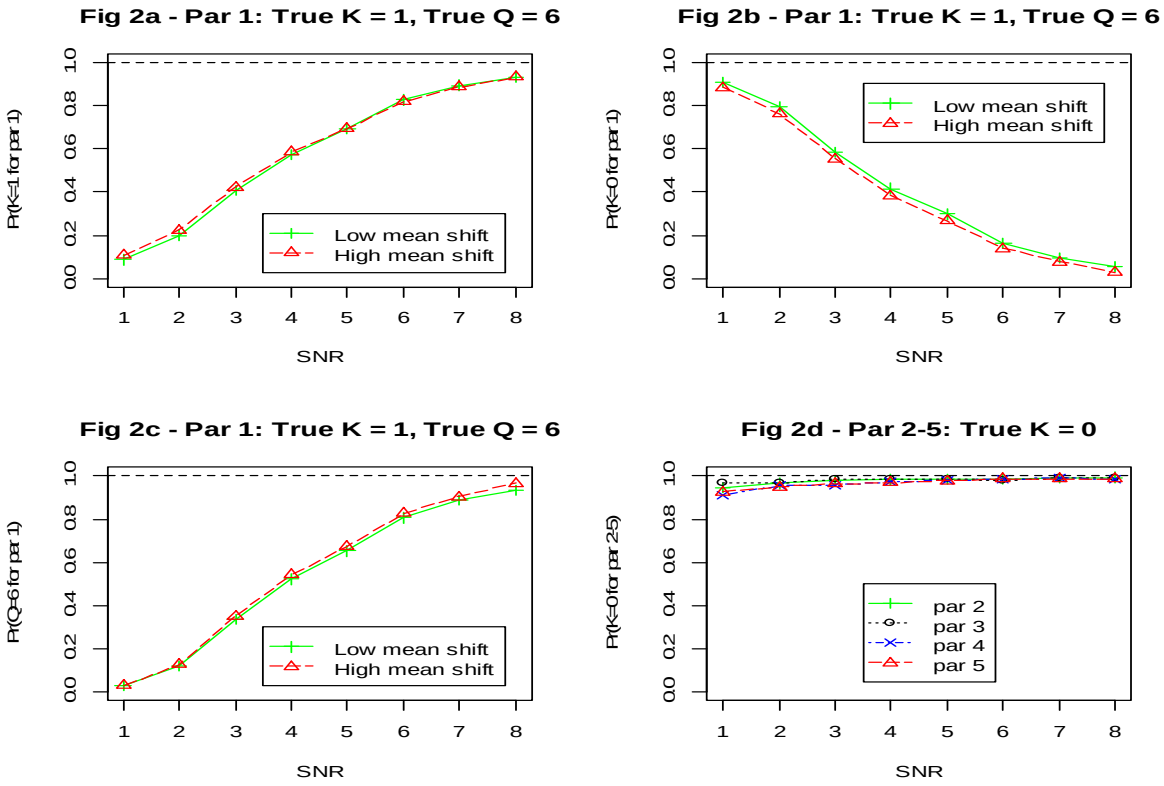


Fig 2e - Par 1: True K = 1, True Q = 9

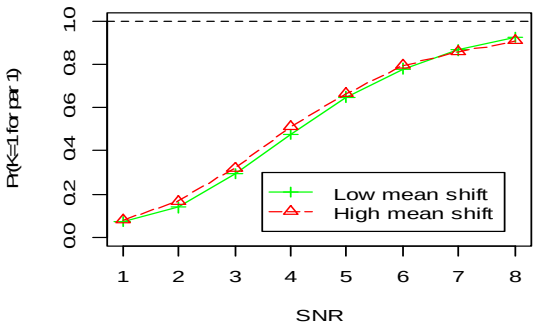


Fig 2f - Par 1: True K = 1, True Q = 9

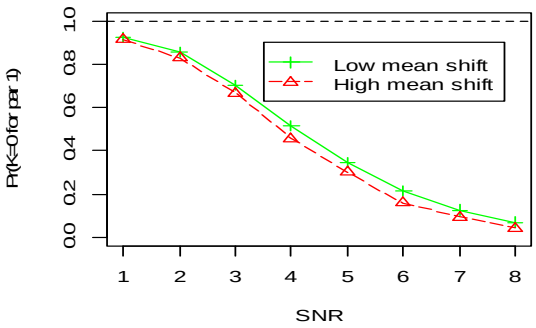


Fig 2g - Par 1: True K = 1, True Q = 9

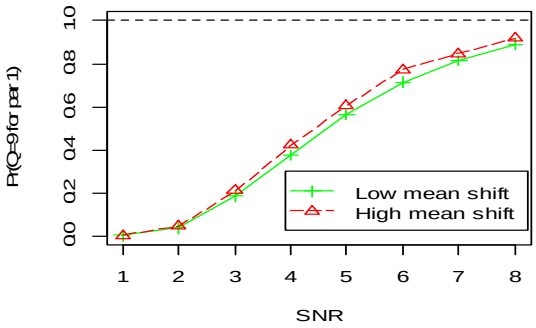


Fig 2h - Par 2-5: True K = 0

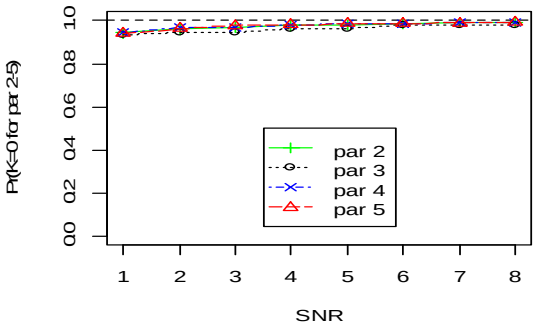


Figure 3: Percentage of active donors for each cohort in calendar time. Since cohorts are acquired at different points in time, this helps identify any calendar-time effects.

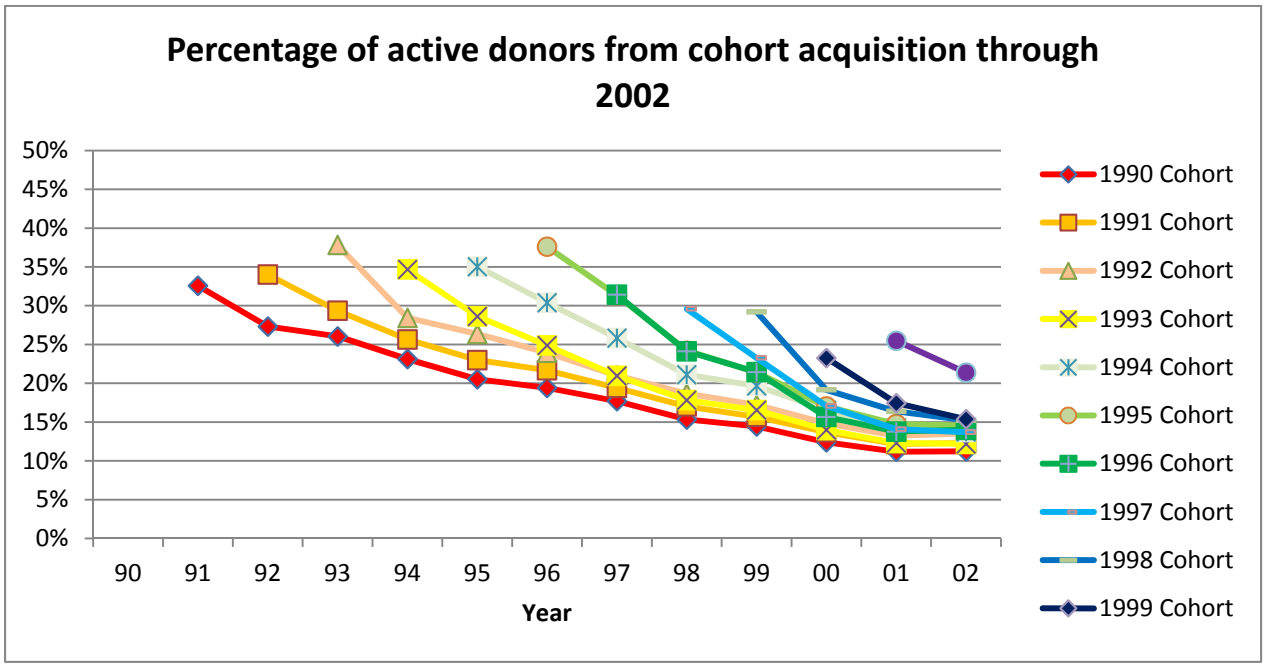


Figure 4: Percentage of active donors in years since cohort acquisition. Note that the 2000 cohort has only two repeat observations.

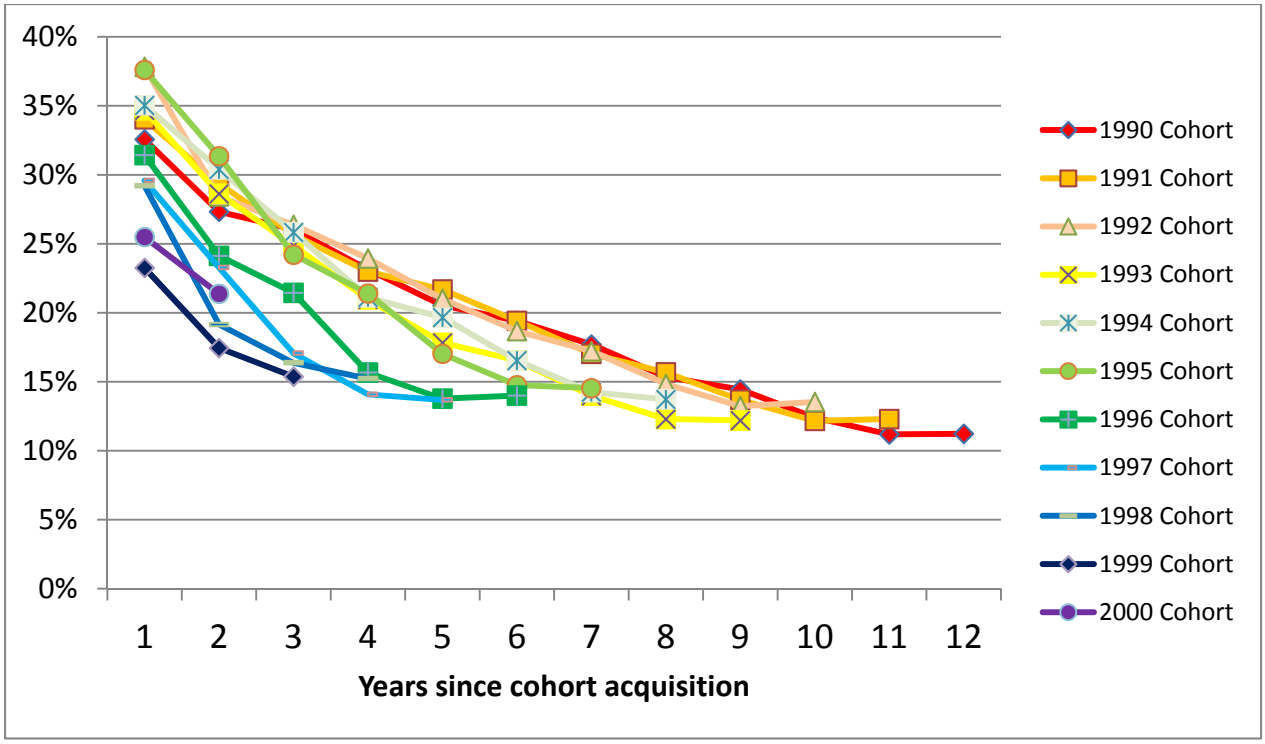


Figure 5: HB-VCT model estimates: (a) There is a 73% probability of a changepoint in the attrition-transaction correlation parameter (with most of this mass at K = 1 changepoint); (b) The most probable location of this changepoint is the 1996 cohort; (c) The HB-VCT prior mean for each cohort is shown in dotted lines to demonstrate the mutual consistency of K and Q with priors. The posterior cohort parameters are also shown in Figure 5c (solid lines).

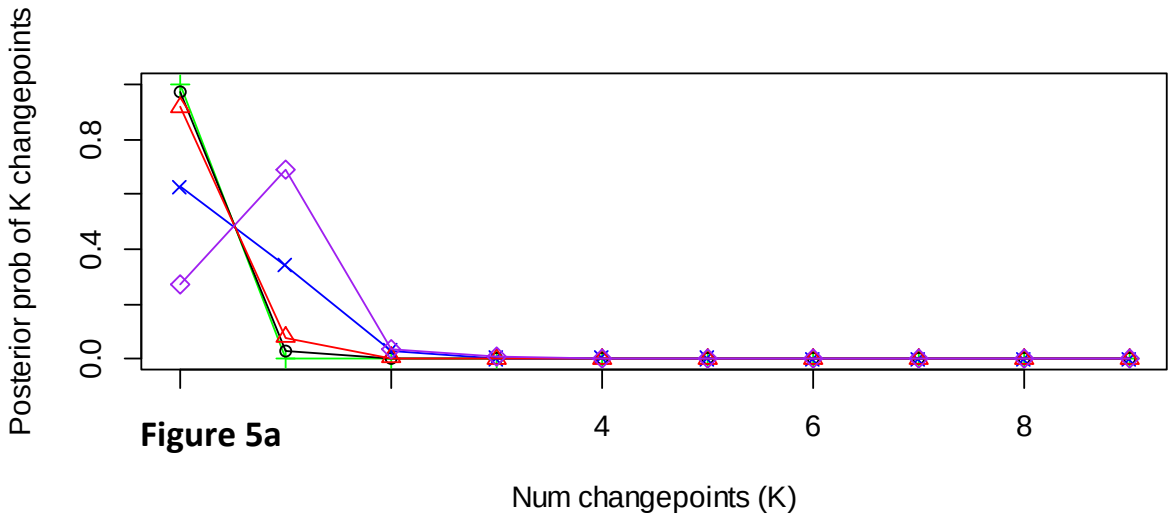


Figure 5a

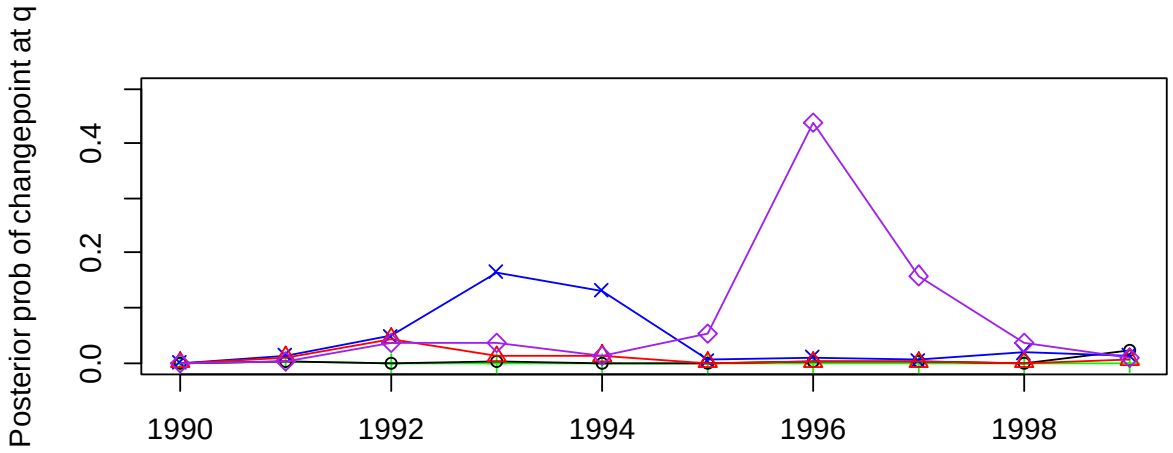
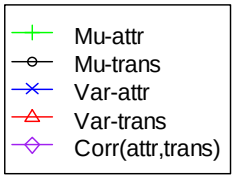


Figure 5b



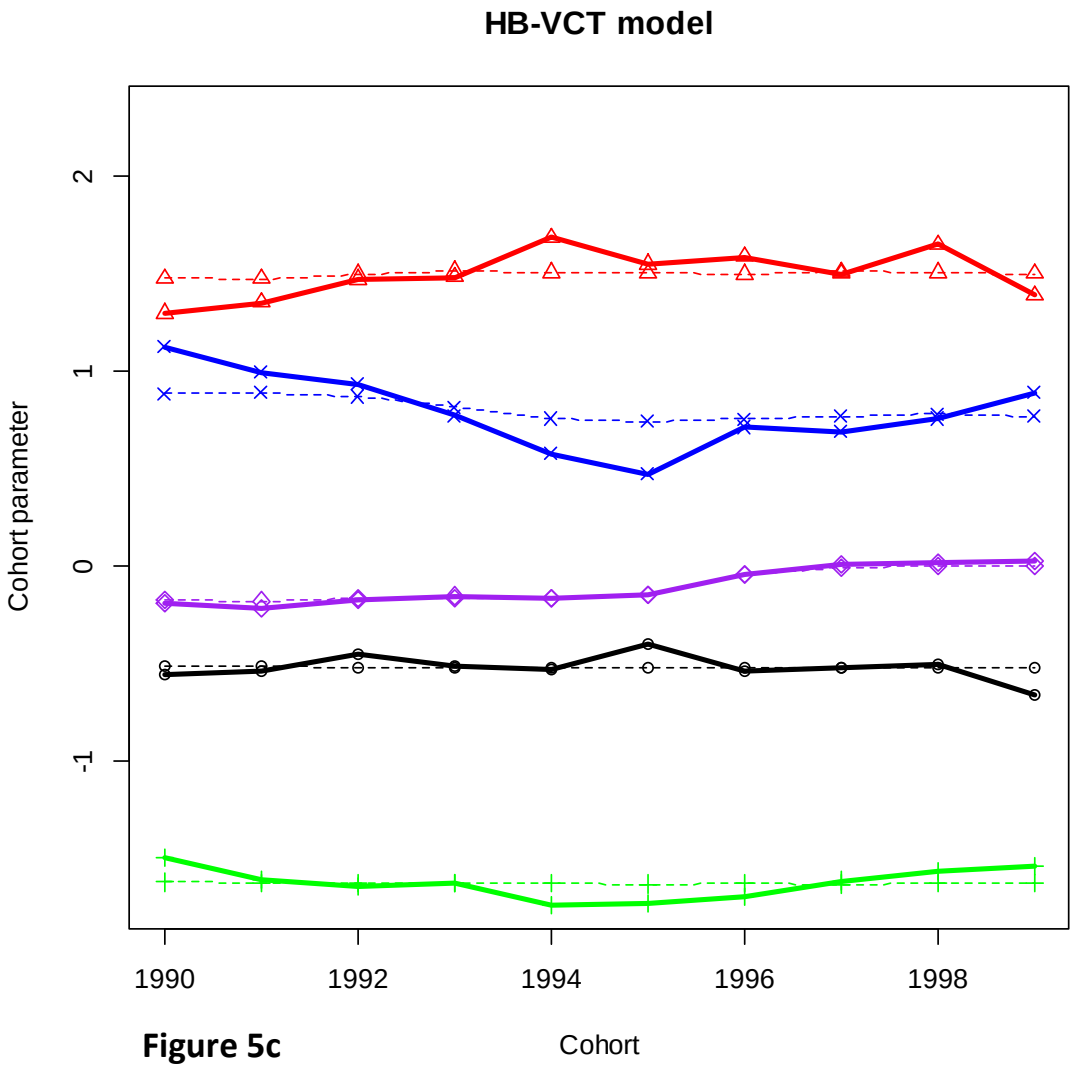
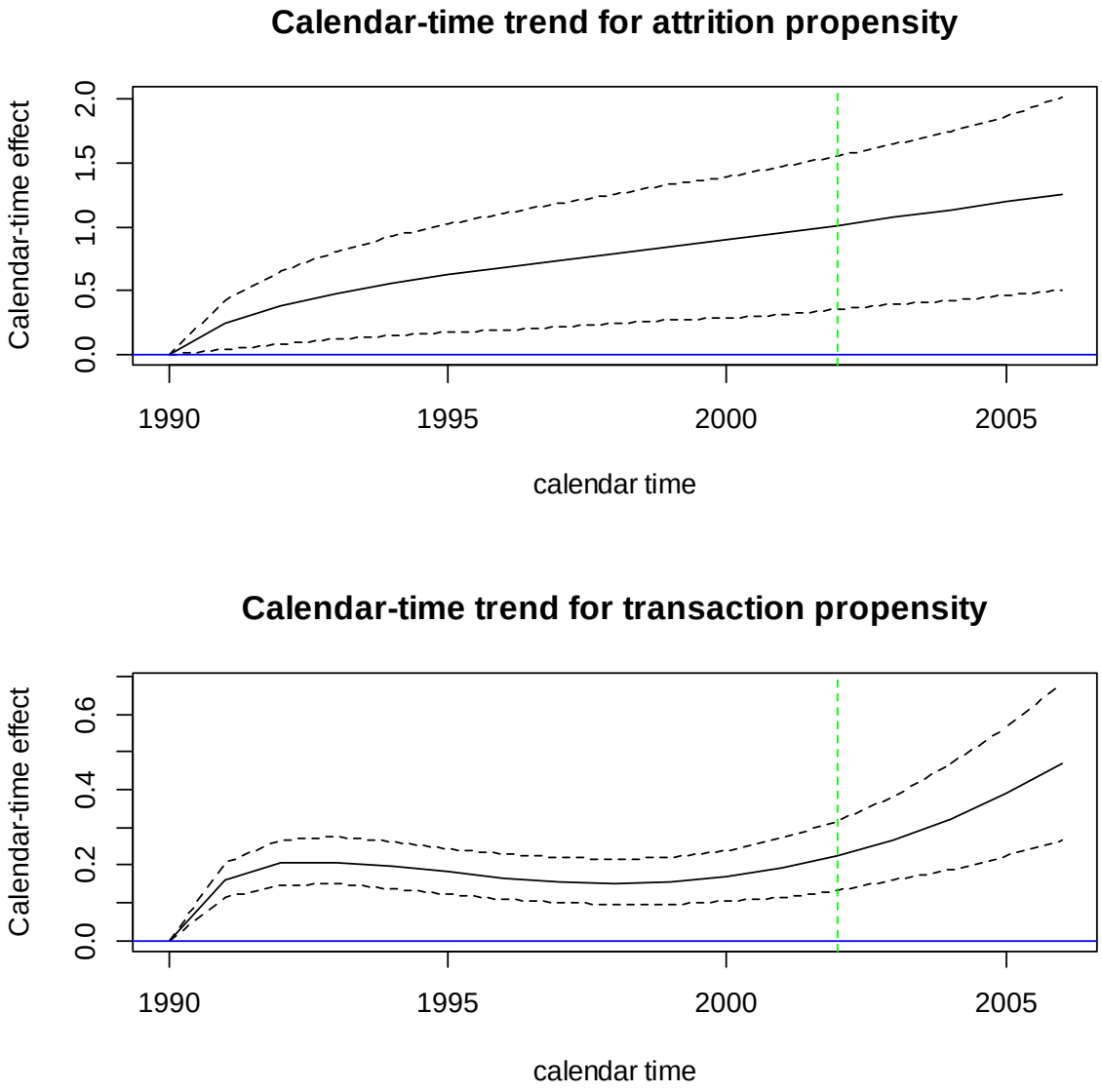

HB-VCT model
Cohort parameter
2
1
0
-1
1990
1992
1994
1996
1998
Cohort
Mu-attr
Mu-trans
Var-attr
Var-trans
Corr(attr,trans)

Figure 5c

Figure 6: Estimated calendar-time trends from HB-VCT model.



Online Technical Appendix

accompanying

A Cross-Cohort Changepoint Model for Customer-Base Analysis

Technical Appendix A – MCMC Implementation

Sampling cohort-specific parameters

We use a partially collapsed Gibbs sampler to efficiently obtain draws of cohort-specific parameters. As an illustration, suppose we look to sample from the following joint posterior distribution for a single cohort $p(\eta_c, \{\alpha_{icz}\}, \{\alpha_{icy}\}, \{L_{ic}\}, \{Y^*_{ic}\}, \{Z^*_{ic}\} | Y_c, X_c, \beta_z, \beta_y, M_{VC})$, where Y^*_{ic} is a vector of augmented random variables $\{Y^*_{ict}\}$ such that $y_{ict} = I(Y^*_{ict} \geq 0)$ and Z^*_{ic} is a vector of augmented random variables $\{Z^*_{ict}\}$ such that $z_{ict} = I(Z^*_{ict} \geq 0)$. Y^*_{ic} and Z^*_{ic} enable Gibbs sampling by simplifying the full conditional distributions. Y_c is a matrix of transaction incidences for customers in cohort *c* over T_c periods and X_c contains (possibly) time-varying covariates for each customer and time period. Our sampler is partially collapsed (Park and van Dyk 2009) in that we marginalize and trim step 1 to speed up convergence and reduce computation.

1. Draw $L_{ic} \sim p(L_{ic} | \alpha_{icz}, \alpha_{icy}, \beta_z, \beta_y, Y_i, X_{ic})$, from a multinomial distribution with support $\{rec_{ic}, \ldots . T_c\}$ whose probabilities are computed for each latent lifetime included in the support – note the importance of conditioning on recency which acts as a lower bound on lifetime.
2. Draw each $Z^*_{ict} \sim p(Z^*_{ict} | L_{ic}, \alpha_{icz}, \beta_z, X_{ict}) \propto \exp\left(-\frac{1}{2}(Z^*_{ict} - \alpha_{icz} - X_{ict}\beta_z)^2\right) [I(Z^*_{ict} \leq 0, t \leq L_{ic}) + I(Z^*_{ict} \geq 0, t > L_{ic})] \; \forall \; t \; \in \{1, \ldots L_{ic} + 1\}$
3. Draw each $Y^*_{ict} \sim p\left(Y^*_{ict} \middle| L_{ic}, \alpha_{icy}, Y_i, \beta_y, X_{ict}\right) \propto \exp(-\frac{1}{2}\left(Y^*_{ict} - \alpha_{icy} - X_{ict}\beta_y\right)^2)[I(Y^*_{ict} \leq 0, Y_{ict} = 0) + I(Y^*_{ict} \geq 0, Y_{ict} = 1)] \; \forall \; t \; \in \{1, \ldots L_{ic}\}$
4. Draw each $\alpha_{icz} \sim p(\alpha_{icz} | Z^*_{ic}, L_{ic}, \alpha_{icy}, \beta_z, X_{ic}, \eta_c) \propto \prod_{t=1}^{\min(L_{ic}+1, T_c)} \exp\left(-\frac{1}{2}(Z^*_{ict} - \alpha_{icz} - X_{ict}\beta_z)^2\right) \cdot N(\alpha_{icz} | \alpha_{icy}, \eta_c)$
5. Draw each $\alpha_{icy} \sim p(\alpha_{icy} | Y^*_{ic}, L_{ic}, \alpha_{icz}, \beta_y, X_{ic}, \eta_c)) \propto [\prod_{t=\min(1, L_{ic})}^{\mathrm{L_{ic}}} \exp\left(-\frac{1}{2}\left(Y^*_{ict} - \alpha_{icy} - X_{ict}\beta_y\right)^2\right)]^{I(L_{ic}>0)} \cdot N(\alpha_{icy} | \alpha_{icz}, \eta_c)$
6. Draw $\eta_c \sim p(\eta_c | \{\alpha_{icz}\}, \{\alpha_{icy}\}, M_{VC})$, which is a random-walk Metropolis-Hastings step, whose jump size we tune during burn-in to optimize the acceptance rate

Steps 1 through 6 are repeated for each of the 10 cohorts used for model estimation, and are parallelized for efficiency. Parallelization requires a multi-threaded environment for which we used Amazon Web Services (AWS) elastic cloud computing resources. Steps 2 and 3 are especially computationally intensive and benefit from a 10-fold speed increase due to parallelization. We implemented this using the R library "parallel", which allows the dynamic creation and allocation of tasks to worker processes spawned as new threads. In this case, the task is to compute steps 1 through 6 for any given iteration. When all cohorts' parameters are drawn, the algorithm continues on to sample the hierarchical model and common slope parameters.

Sampling common slope parameters

Given that we store all augmented variables from each cohort's draws in steps 1 through 6, it is a straightforward Bayesian regression step to obtain the slopes. We use a diffuse Multivariate Normal prior centered at zero for all coefficients.

7. Draw $\beta_z \sim p(\beta_z | \{\alpha_{icz}\}, \{L_{ic}\}, \{Z_{ic}^*\}, X_c)$, which is a Bayesian regression update step.
8. Draw $\beta_y \sim p(\beta_y | \{\alpha_{icy}\}, \{L_{ic}\}, \{Y_{ic}^*\}, X_c)$, which is a Bayesian regression update step.

Sampling hierarchical changepoint model parameters

We use a Reversible-Jump sampler for each parameter dimension (see Technical Appendix B for a brief overview). $\tau_0 = 0.5, \tau_1 = 0.5$ are set to allow for reasonably efficient R-J proposals.

9. For each parameter dimension *d*,

   Draw new model (K',Q',Ω') based on current model (K, Q, Ω) and $\{\eta_{cd} \forall c\}$

   a. If $0 < K < N - 1$, K' = K with prob τ_0 and K' = K+1, or K' = K-1 with prob $\frac{1-\tau_0}{2}$ each
      Else if $K = 0$, K' = K with prob τ_0 and K' = K+1 with prob $1 - \tau_0$
      Else K' = K with prob τ_0 and K' = K-1 with prob $1 - \tau_0$
   b. If K' = K, Q' = Q with prob τ_1 and Q' new draws with prob $1 - \tau_1$
   c. If K'=K+1, Q'=c(s,Q), where s is a randomly selected new changepoint from the feasible set of changepoint locations.
   d. If K'=K-1, Q'=Q[-s], where s is a randomly selected changepoint to be removed.
   e. Update means and variances of blocks which have *changed* from Q to Q', by drawing from posterior distributions. Other block parameters are the same.

f. Calculate acceptance probability ratio, and either accept or reject the proposed model

To sample the hierarchical correlation matrix R (if not set to the identity matrix), we use a Metropolis-Hastings step with a Barnard prior $p_k(R) \propto |R|^{-(k+1)}\left(\prod_i r^{ii}\right)^{-\frac{(k+1)}{2}}$, where $k = D$ is the number of diagonal elements in R and r^{ii} is the *i-th* element on the main diagonal of R^{-1} (Barnard et al 2000). One challenge in step 10 is to propose valid correlation matrices (that satisfy the positive definiteness constraint). While a number of algorithms have been proposed to draw matrices that satisfy the positive definiteness constraint (Liechty et al 2004; Joe 2006; Budden et al 2008; Hürlimann 2012), we implement a random-walk Metropolis-Hastings version of Budden et al (2008) which exactly draws from the space of allowed correlation matrices. The M-H approach uses the current draw of correlation coefficients and iteratively proposes a new set of coefficients, while handling the scenarios where the coefficients in the previous iteration are infeasible given some of the coefficients in the new iteration.

10. Draw $R \sim p(R|M_{VC}, \{\eta_c\}) \propto \prod_{c=1}^{N} p(\eta_c|M_{VC}, R) \cdot p(R)$

Iterating over Steps 1 through 10 (with steps 1 to 6 executed in parallel for each cohort) leads to draws from the joint posterior distribution in equation (3).

Modifications for the B-VC model

The B-VC model imposes that parameters of cohorts in the same regime are identical, as opposed to being i.i.d draws from the same regime-specific distribution (HB-VC). Steps 1 through 5 in the HB-VC algorithm remain the same, but Step 6 is redundant since cohort-level parameters are deterministically obtained from the changepoint model.

Step 9e changes since a single scalar parameter needs to be proposed for each regime (in each dimension), which we draw from a known closed-form distribution (further detailed in Technical Appendix B). Step 10 is redundant since cohort parameters are not i.i.d draws from a regime-specific distribution. Conditional on the changepoint model, cohort parameters are known values.

Technical Appendix B – Reversible-Jump Sampler for Vector Changepoint Model

HB-VC Model

No closed-form density exists for the joint posterior distribution in equation (3) so we are unable to sample directly from it. We therefore need to use Markov Chain Monte Carlo (MCMC) methods such that the stationary distribution of the chain is the joint posterior density in equation (3). Traditional samplers in the MCMC family include the Gibbs sampler (based on the full conditional distribution for a parameter) and Metropolis-Hastings sampler.

However, the issue of changing parameter dimensionality arises in devising a sampler for $p(K_d, Q_d, \Omega_{\mathrm{d}} | \{\eta_{cd} \forall c\}, R)$. Suppose that K_1 = 0 in the current MCMC iteration so that there is only one block and associated normal distribution. This corresponds to the standard HB model without changepoints where pooling across all cohorts is enabled. However, in our model, K_1 can take on any value between 0 and N-1. Suppose that the sampler proposes a move to $K_1 = 1$. This would then generate two blocks, and therefore an additional normal distribution (whose mean and variance parameters were previously non-existent). Standard MCMC algorithms such as M-H assume that the number of parameters do not vary across proposals. We therefore utilize the Reversible-Jump (R-J) sampler introduced by Green (1995) that generalizes the M-H sampler to allow for proposals that change the dimensionality of the parameter space. We then use D such R-J steps to update the vector changepoint model for each of the D cohort-level parameters (which collapses to just one step for the classical changepoint model).

Since the R-J sampler is not as commonly used as the M-H and Gibbs samplers, we provide a brief description of how we apply it in this context. Readers can refer to Green (1995) and Hastie (2005) for additional details. For notational ease, let $\{\mathrm{K}, \mathrm{Q}, \Omega\}$ be the current draw and $\{\mathrm{K}', \mathrm{Q}', \Omega'\}$ be the proposal, which is obtained according to the researcher's specification of allowed moves. To respect the detailed balancing condition required for convergence, we assume that a vector of random variables u and u' are obtained such that $[\Omega', u'] = T([\Omega, u])$, where T is a deterministic transformation and $\dim(\Omega') + \dim(u') = \dim(\Omega) + \dim(u)$. Let

$A = \frac{p(K', Q', \Omega' | \{\eta_{cd} \forall c\}, R) p(K, Q | K', Q') f'(u')}{p(K, Q, \Omega | \{\eta_{cd} \forall c\}, R) p(K', Q' | K, Q) f(u)} |J|$, where J is the Jacobian of the transformation T with respect to (Ω, u). The acceptance probability for the proposal is given by $\min(1, A)$.

The intuition for the expression in A is similar to that of the M-H sampler. The ratio of the posterior densities of the proposal and incumbent states is exactly the same as what would appear in a standard M-H acceptance probability. The expression $\frac{p(K, Q | K', Q')}{p(K', Q' | K, Q)}$ corrects for the asymmetry in making this type of move. A move which is highly probable to be proposed will be penalized if the reverse direction is not as probable. The expression $\frac{f'(u')}{f(u)}$ accounts for the asymmetry in drawing the random variables for the detailed balancing condition. Finally, the determinant of the Jacobian matrix, surprisingly does not arise from changing dimensionality (Hastie 2005) but from the indirect transformation T.

We allow for three types of K moves: split, combine and keep. A split move increments K by one (K' = K+1), a combine move decrements K by one (K' = K-1), and a keep move (as the name implies) leaves K'=K at the same value. K changes with probability τ_0, and so long as both an increase and decrease are possible, will shift up or down with equal probability. If K happens to be at 0 or N-1, then it will move in the only possible direction with probability τ_0. In alignment with the generation of K', Q' involves either adding a new changepoint location (split move), removing an existing changepoint location (combine move), reshuffling all changepoint locations (keep move), or no change (keep move). Given a keep move, the reshuffle of Q occurs with probability τ_1. A graphical depiction of the allowed moves is shown in Figure B-1 for the case where K is neither 0 nor (N-1).

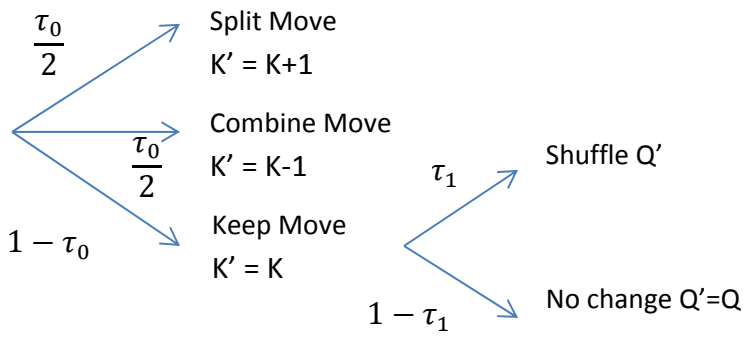


Figure B-1: Tree generating changepoint model moves

Ω' is generated conditional on $\{K', Q', \{\eta_{cd} \forall c\}\}$. If a keep move has been chosen, then Ω' is drawn from the posterior NIX distributions for each block since a closed form exists for $p(\Omega'|K', Q', \{\eta_{cd} \forall c\})$. For a keep move, the acceptance probability reduces to an M-H step. For a split or combine move, only the blocks affected by the change will have new parameters drawn from the posterior NIX distributions for those blocks. Essentially, $f(u) = p(\Omega_c'|K', Q', \{\eta_{cd} \forall c\})$, where Ω_c' is the set of parameters from blocks that require updating. For detailed balancing, we set $u' = \Omega_c$ so that $f'(u') = p(\Omega_c|K, Q, \{\eta_{cd} \forall c\})$. The remaining parameters are then identical across Ω and Ω'. The transformation described above results in T being the identity matrix so that the determinant of the Jacobian is 1, which disappears from the acceptance probability expression. The NIX hyperparameters $\{\mu_0 = 0, \kappa_0 = 0.01, \nu_0 = 0.01, \Sigma_0 = 0.01\}$ are Bayesian updated using as data the parameters from cohorts assigned to a block.

Split/combine locations are equally likely to be selected from the feasible set of locations. Thus, $\log\left(\frac{p(K, Q|K', Q')}{p(K', Q'|K, Q)}\right) = (K' - K)\log(\frac{N-1-\min(K,K')}{\min(K,K')+1})$.

Modifications for the B-VC model

Individual-level attrition and transaction propensities $\alpha \equiv \{\alpha_{icz}, \alpha_{icy}\}\ \forall i, c$ serve as the data for the B-VC model, instead of cohort-level parameters. The acceptance probability expression becomes:

$$A = \frac{p(K', Q', \Omega'|\alpha)p(K, Q|K', Q')f'(u')}{p(K, Q, \Omega|\alpha)p(K', Q'|K, Q)f(u)}|J|.$$

K' and Q' moves are unchanged, while Ω' is drawn from a closed form distribution conditioned on α for efficient mixing.

The priors for K and Q need to be informative to avoid fitting boundary solutions such as fitting each cohort to its own regime. Given that parameters are constant in a regime, the likelihood of the data is maximized when each cohort is in its own regime. Without an informative prior, the B-VC model would therefore fit “spurious changepoints” to maximize model fit. We therefore used a prior that placed a floor of 4 cohorts for the minimum regime size and at most one regime change. The issue does not arise in the HB-VC model since each cohort is considered an i.i.d draw from a regime-specific distribution.

Technical Appendix C – Empirical Recovery of CPN/PN model with common slopes

We briefly discuss the empirical recovery of a standalone cohort with time-invariant covariates before focusing on model recovery that includes time-varying covariates. Similar to other BTYD models, true parameters that do not result in "too sparse a dataset" will be robustly recovered as long as the number of repeat observations is sufficiently large. As an extreme example of sparsity, imagine a cohort *c* with parameters $\mu_{cz} = -5, \mu_{cy} = -5, \sigma_{cz}^2 = 0.25, \sigma_{cy}^2 = 0.25$. A dataset simulated with these parameters will consist of almost all individuals having zero repeat transactions, because the transaction probability is extremely low for any given individual. However, the algorithm will find it difficult to pin down whether the zeros are due to *high attrition propensity* or *low transaction propensity*. A cohort with moderate attrition and transaction means, and a reasonable amount of heterogeneity will be recovered, albeit estimator variance is more sensitive to the length of the temporal window (number of repeat observations) than the number of individuals in the cohort.

Of greater interest is the ability to tease apart cross-cohort and calendar-time effects in a cohort sequence. Can these be recovered empirically for relevant parameter values in our empirical setting? We use the B-PE model to highlight that even in a sequence of 10 cohorts, recovery is robust. The difference between B-PE and the HB models is that the cross-cohort effect, instead of being defined as a function of parametric curves, is introduced through the structure in the hierarchical prior. In the simulations below, we turn on and off linear cross-cohort and calendar-time effects on both the attrition and transaction processes, run the algorithm and obtain the posterior means of all parameters to compare against the true ones. The number of individuals in each cohort matches Table 2 to mirror the amount of data in each actual cohort. The covariates are coded as follows: $X_{ict,cohort} = \frac{c-1}{10}, \quad X_{ict,caltime} = \frac{c-1+t-1}{10}$. Note that $t \in \{1, \dots T_c\}$ and indexes cohort repeat purchase opportunities whereas $X_{ict,caltime}$ is a scaled version of calendar time. The cross-cohort slopes shift the attrition and transaction mean parameters by cohort, and the calendar-time slopes shift the attrition and transaction mean parameters at each calendar time period the same way for all cohorts.

Simul #	μ_{cz}	μ_{cy}	σ^2_{cz}	σ^2_{cy}	σ_{czy}	$\beta_{z,cohort}$	$\beta_{z,caltime}$	$\beta_{y,cohort}$	$\beta_{y,caltime}$
1: True	-0.89	-0.38	0.31	1.46	-0.18	0.00	0.00	0.00	0.00
1: Estim	-0.87	-0.40	0.23	1.43	-0.18	0.17	-0.14	0.02	0.02
2: True	-1.35	-0.36	0.93	1.26	-0.31	0.00	1.14	0.00	0.17
2: Estim	-1.40	-0.37	0.97	1.30	-0.29	-0.06	**1.23***	-0.04	**0.16***
3: True	-1.26	-0.37	0.55	1.29	-0.30	0.31	0.70	0.00	0.17
4: Estim	-1.28	-0.38	0.55	1.29	-0.28	**0.32***	**0.72***	-0.04	**0.21***
4: True	-1.26	-0.37	0.55	1.29	-0.30	-0.31	0.70	0.10	-0.20
4: Estim	-1.26	-0.36	0.66	1.32	-0.30	**-0.36***	**0.70***	**0.15***	**-0.26***

Table C-1: True and estimated parameters (for slope parameters, * indicates that 95% posterior interval excludes zero)

Model 1 is the null model in which all cross-cohort and calendar-time effects are switched off, and none of the slope estimates are significantly different from zero. We observe that the estimated CPN/PN parameters are close to the true ones. In Model 2, we turn on calendar time effects and find they are robustly recovered. In Models 3 and 4, we enable both cross-cohort and calendar-time effects and find that these are both recovered for the attrition and transaction processes. Critical to identification of these slopes is having a multi-cohort structure, as there would be no variation in calendar-time for individuals of the same cohort.

Technical Appendix D: Empirical recovery of HB-VCT model using simulated dataset.

We simulate a sequence of 10 cohorts with a changepoint at the seventh cohort for the correlation (between attrition and transaction propensities) parameter (similar results are obtained for changepoints in other parameters). The first cohort has 12 time observations (during the in-sample period) and every successive cohort features one fewer time observation. We simulate an eleventh cohort and hold it out from estimation. In addition, the last four time observations for each cohort are held out for model validation. The regime-specific mean parameters were as follows:

Parameter	**Regime 1 (Cohorts 1 to 6)**	**Regime 2 (Cohorts 7 onwards)**
Attrition mean	-1.75	-1.75
Transaction mean	-0.50	-0.50
Attrition variance	0.75	0.75
Transaction variance	1.50	1.50
Correlation between attrition and transaction propensities	-0.25	0.00

Table D-1: Simulated changepoint parameters.

The data generating process also included calendar-time effects for both the attrition and transaction processes, using a combination of linear, quadratic and logarithmic time terms, generating time effects as shown in Figure D-1. The parameters were chosen to resemble the estimates from the HB-VCT model based on the actual data set used for empirical analysis.

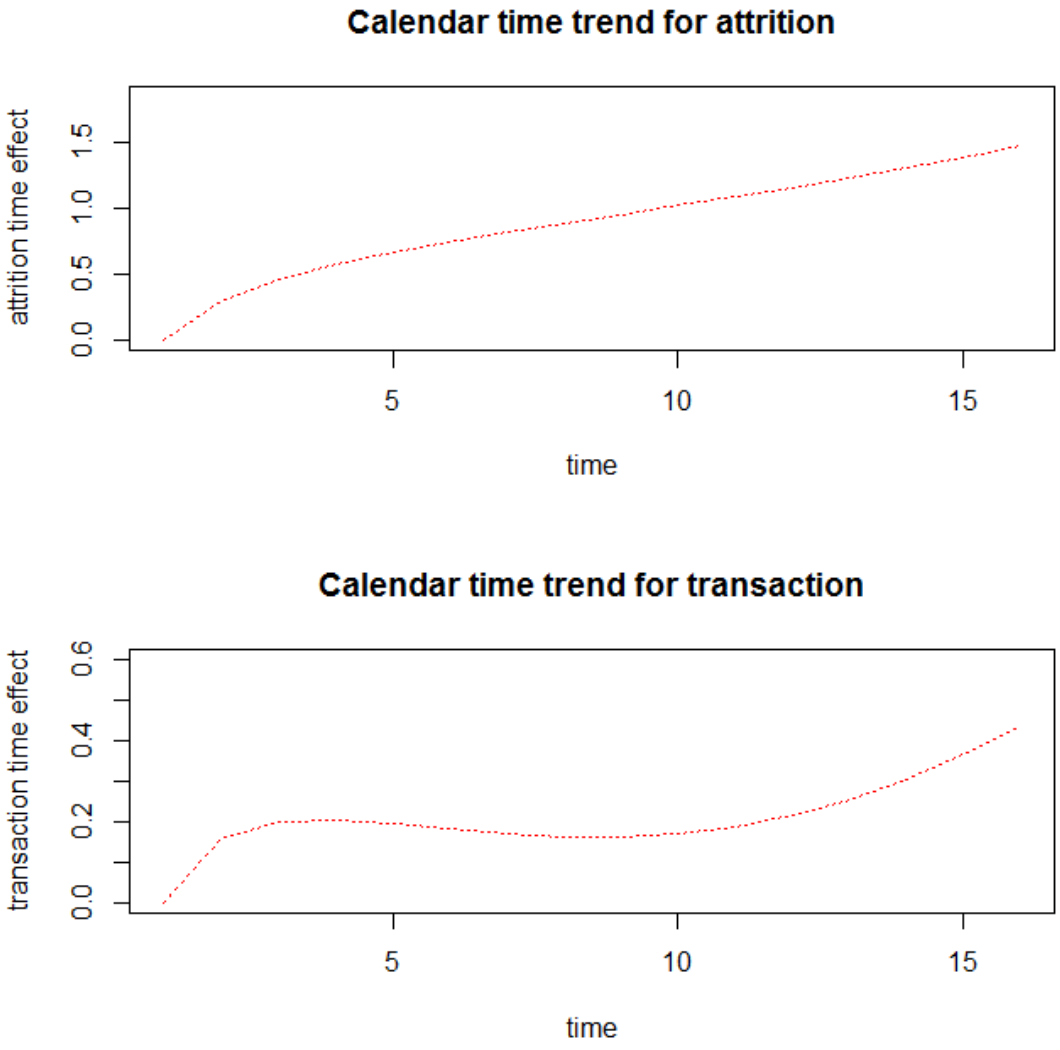


Figure D-1: Simulated calendar-time effects for attrition and transaction processes. Only the first 12 time periods are covered in the in-sample data.

The model estimates a probability of exactly one changepoint as 93.0% for the correlation parameter, and the probability of zero changepoints for other parameters are 98.8% or higher. The probability of a changepoint in the correlation parameter located at the seventh cohort is estimated as 95.4%. The model therefore recovers the changepoint structure with high accuracy for all parameters. The cohort-level parameters (Table D-2) reflect the changepoint at the seventh cohort for the correlation parameter.

Parameter/Cohort	1	2	3	4	5	6	7	8	9	10
Attrition mean	-1.82	-1.82	-1.78	-1.80	-1.82	-1.80	-1.76	-1.75	-1.83	-1.77
Transaction mean	-0.51	-0.51	-0.51	-0.53	-0.50	-0.50	-0.51	-0.50	-0.51	-0.52
Attrition variance	0.71	0.72	0.71	0.72	0.71	0.72	0.72	0.70	0.71	0.72
Transaction variance	1.50	1.49	1.48	1.49	1.48	1.48	1.50	1.50	1.49	1.51
Correlation between attrition & transaction	-0.22	-0.21	-0.20	-0.22	-0.22	-0.26	0.02	0.02	0.02	0.02

Table D-2: Recovered cohort-level parameters

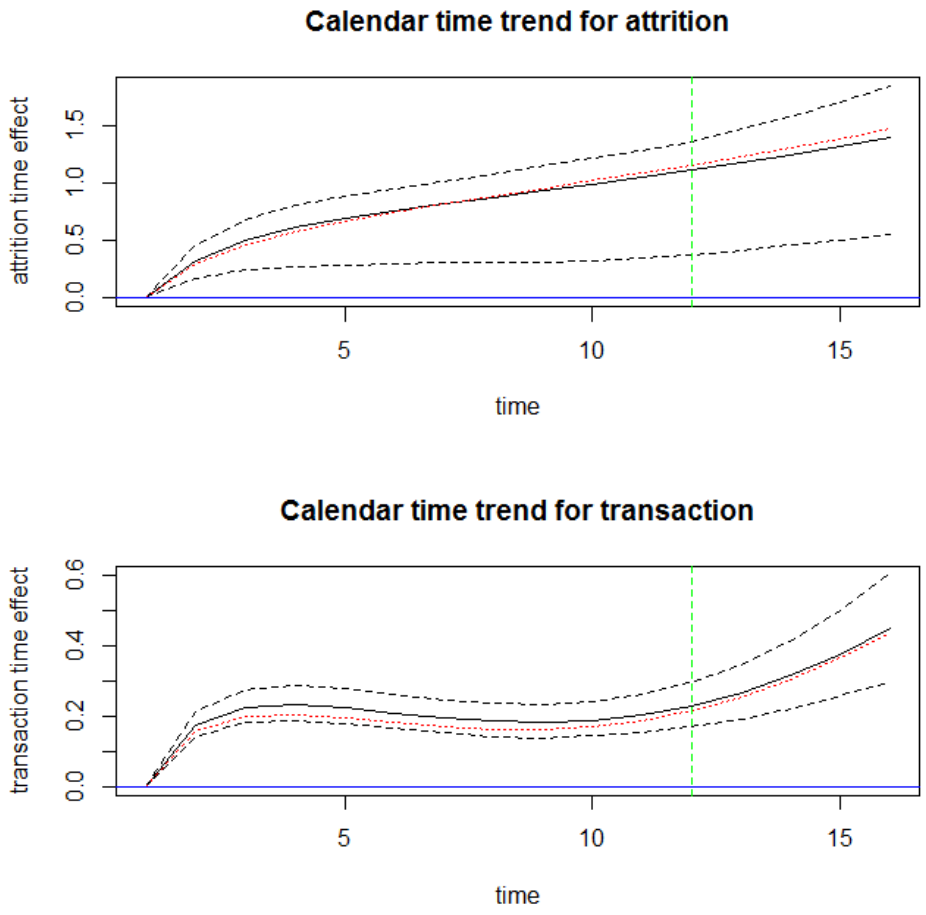


Figure D-2: Recovered calendar-time effects. Solid line is the posterior median time effect. Large-dash dotted lines represent the 95% posterior interval for the time effect. Small-dash dotted lines represent the true calendar-time effect. Vertical line shows in-sample cutoff.

As shown in Figure D-2, the calendar-time effects for the attrition and transaction propensities are recovered well, including out-of-sample (periods 13 through 16). Overall, this analysis shows that the HB-VCT model is able to accurately recover changepoint structure and cohort-level parameters, while separating calendar-time effects from simulated data.

Finally, we present the in-sample and out-of-sample MAPEs by cohort in Table D-3. Out-of-sample MAPEs show that the model accurately generalizes beyond the in-sample time periods, and also accurately forecasts an entirely held out eleventh cohort based on the estimated changepoint structure and calendar-time effects.

Metric/Cohort	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11 (Heldout)**
In-sample MAPE	3.3%	2.8%	3.2%	3.1%	4.7%	3.1%	3.5%	3.5%	4.6%	4.0%	n/a
Out-of-sample MAPE	5.1%	4.9%	5.1%	4.6%	5.5%	5.1%	7.3%	6.4%	6.6%	7.0%	9.0%

Table D-3: In-sample and out-of-sample MAPEs for first ten cohorts used for model estimation, and heldout eleventh cohort