A Cross-Cohort Changepoint Model for Customer-Base Analysis

Arun Gopalakrishnan

The Wharton School, University of Pennsylvania

Advised by:

Professor Eric T. Bradlow

Professor Peter S. Fader

Submitted as First Year Summer Paper

12 September 2011

Correspondence concerning this paper may be addressed to the above author at The Wharton School, 700 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, agop@wharton.upenn.edu. The author would like to thank Professors Shane Jensen and Raghu Iyengar for their feedback and suggestions.

Abstract

A Cross-Cohort Changepoint Model for Customer-Base Analysis

Many firms maintain a customer database with detailed transaction histories at the individual level. Despite this wealth of information, firms are yet to fully exploit the underlying structure in a customer database to forecast the behavior of brand new customers and detect whether new customers are markedly different from old customers. We propose to extract information from the underlying structure in a database by segmenting customers into cohorts indexed by period of acquisition, which results in a sequence of customer cohorts. We develop a Hierarchical Bayesian cross cohort changepoint model framework that (1) allows for cross-sectional heterogeneity within a cohort, (2) identifies shifts/changepoints in cohort behavior along the sequence of customer cohorts, and (3) enables predictions of recently acquired cohorts for whom the firm has little to no longitudinal transaction data. We apply our model to a discrete-time non-contractual setting using multi-cohort donation data from a non-profit and find that the changepoint model provides improved holdout prediction by reducing bias in the predictive distribution versus a static baseline model. The results demonstrate that detecting changepoints across a sequence of cohorts allows the firm to use data from relevant "peer" cohorts to make predictions about new cohorts and pick up discrete shifts in behavior across cohorts. The proposed framework can be applied to any setting with a sequence of cohorts, by selecting an appropriate cohort-level model.

Key words: Changepoint; Cross Cohort; Hierarchical Bayesian; Forecasting; Customer-base analysis.

I. Introduction

Customer-base analysis occupies a central role in many firms' marketing departments, especially in transaction-oriented industries such as telecommunications, internet retailing, and non-profits where customer acquisition and retention is critical to business success. Models of customer-base analysis have demonstrated accurate forecasting performance in a variety of settings (Schmittlein, Morrison and Colombo 1987; Netzer, Lattin and Srinivasan 2008; Fader, Hardie and Shang 2010). Many of these models account specifically for the heterogeneity in a customer base and some also model non-stationarity of behavior at the individual level (Allenby, Leone and Jen 1999; Fader, Hardie and Huang 2004). Implicitly, these models assume that customers in the data set come from the same population, such that the population-level model can be specified and estimated from all of the data. Suppose instead that the customer base is in fact comprised of multiple segments, each having its own unique model. One such segmentation is a sequence of cohorts, where each cohort consists of the set of customers acquired in the same time period. This broader model allows us to explore new structural patterns in the customer database, directly relevant to important managerial questions.

Consider a fictitious online retailer, Acme Apparel Inc. Acme has been selling clothing over the internet since 2001. Each year, they acquire a cohort of new customers. Anita, the sales manager wants to forecast future sales for newly acquired cohorts but has very little data about these customers. She turns to Bob, the data analyst for assistance. Bob is not sure if he should use all cohorts since 2001 to estimate the latest cohort's forecast, or to draw a line at a particular year because the behavior of the oldest cohorts may no longer reflect what can be expected today. He decides that more data will improve his forecast and therefore uses all the data at his disposal. In the meanwhile, Colleen the business intelligence manager has intuition that a competitor who entered the market in 2006 has been "skimming the best new customer prospects", such that the "quality" of the newer cohorts Acme has acquired might be significantly different compared to pre-2006 ones. She asks Bob for his insights. Bob knows there are good and bad customers in every cohort, and is not sure if he can state with confidence whether the data confirms Colleen's intuition.

The above example illustrates three key managerial questions: (1) when does data become outdated for forecasting behavior of new cohorts, (2) how to develop forecasts for new cohorts, and (3) is there evidence for a distinct shift in cohort behavior in a sequence of cohorts?

In this paper, we propose a Hierarchical Bayesian cross cohort changepoint model that provides guidance to these managerial questions. Our framework is applicable to a wide range of settings including counts of doctor visits, product interpurchase times, and even manufacturing and epidemiological processes as shown in Table 1. The only requirements are that a sequencing of cohorts exists, longitudinal data is available for each cohort at the individual level, and the cohort-level model is parametric. Our model yields probabilistic inferences about discrete shifts or changepoints in cohort behavior across the sequence of cohorts, and provides the ability to forecast new cohorts' behavior with little or no data from the new cohort itself.

[INSERT TABLE 1 HERE]

We calibrate the model in the context of a discrete-time non-contractual setting, using multi-cohort donation data from a U.S. non-profit organization, using the Beta-Geometric Beta-Bernoulli (BG/BB) model developed by Fader, Hardie and Shang (2010) to characterize each cohort's behavior. Comparing the changepoint model to a Hierarchical Bayesian static model that does not partition cohorts into different blocks, we find evidence for a changepoint and benefit from more accurate forecasting of holdout data versus the static model.

The remainder of this paper is organized as follows. In the next section, we review the relevant literature. We then present the model framework and the corresponding MCMC estimation. We discuss the empirical analysis based on both simulated data and a non-profit multiple donor cohort data set, and highlight the insights learned from applying this model. We conclude with suggestions for future work and overall implications for managers and researchers.

II. Literature Review

Models of cross-cohort relationships

Models that account for inter-relationships among units (whether cohorts or individuals) can be of two types. The first explicitly lays out a model with parameters that capture systematic differences/similarities across units. For instance, Schweidel, Fader and Bradlow (2008) introduce covariates relating to cohort age that affect a baseline hazard function in the context of modeling service retention. Related work capturing spatial relationships among units includes Yang and Allenby

(2003) who use an autoregressive term capturing preference dependency across units in a discrete choice model. Both of these allow for a range of possible patterns of cross-cohort effects but would tend to smoothen out sudden non-stationarity shifts in cohort characteristics rather than detect them. The second type introduces hierarchical priors which enable "information sharing" across units. Hofstede, Wedel and Steenkamp (2002) use spatial prior distributions that borrow information from neighboring units to help smooth posterior inference especially for units with limited data. Hui and Bradlow (2011) suggest a general framework of modeling relationships between units in a graphical configuration (particularly suited for spatial relationships), and use Bayesian methods to determine both the graphical configuration and associated parameters from the data. Since using MCMC sampling on the graphical configuration is infeasible, the posterior mode of the graphical configuration is found using simulated annealing. The model is appealing since it can identify general relationships among units.

Since our problem specifies a sequence of cohorts, we can provide more structure than a general model of inter-unit relationships, by constraining the search to the number and locations of changepoints along this sequence. We essentially seek to divide the sequence into blocks of similar cohorts, each characterized by a hierarchical block prior, which can determine the most relevant contiguous block to forecast the behavior of new cohorts. We assume independence in hierarchical prior distributions across blocks, conditional on the block partitioning pattern. This additional structure comes at the cost of ignoring possible cyclical patterns across cohorts, which may be a fruitful topic for future research.

Even more closely related to our problem is the large literature on detecting changepoints which falls into two basic categories. The first is the product partition approach outlined by Barry and Hartigan (1993), whereby the underlying sequence of parameters that characterizes a sequence of observations is divided into contiguous blocks. Within each block, the parameter value driving those observations is homogeneous. Across blocks, the parameter value is allowed to shift. Both the number of changepoints and their locations are treated as random variables with prior distributions, to allow their incorporation into a Bayesian framework. Though Barry and Hartigan (1993) provide an elegant framework for obtaining draws from the joint posterior of changepoint configuration and parameters, the product partition approach is limited to situations where conjugate priors are used, and it is possible to integrate out nuisance parameters such that the posterior parameters are known once the changepoint configuration is drawn using data augmentation. In other words, once the possible changepoint configuration is drawn, calculating the posterior parameters (of interest) is trivial. The product partition method is therefore unsuitable for more general scenarios where priors are not conjugate, and where

model parameters must also be drawn from a distribution conditional on the changepoint configuration. We do not pursue the product partition method for these reasons.

The second category deals with the complexity of scaling model dimensionality as the number of changepoints is allowed to vary. Suppose the number of changepoints were fixed up-front (at K) with unknown locations. It is straightforward to construct a model with (K + 1) distributions (each block of similar cohorts has its own prior distribution) and use an MCMC procedure. However, if we allow K to vary across MCMC iterations, each draw of K would result in a potential change in the dimensionality of the underlying model, posing difficulties for even evaluating the conditional posterior distribution of K. If a set of parameters is eliminated from the model, an absorbing state may result which violates a necessary condition for convergence (Carlin and Chib 1995). A number of approaches exist to circumvent the changing dimensionality problem.

Green (1995) presents an approach known as "reversible jump MCMC" which allows the number of components/dimensions of a model to shrink or increase, while appropriately handling the Bayesian machinery with an adaptation of the Metropolis-Hastings method. This sharply increases the complexity of the algorithm but provides an approach that obtains convergence with judicious tuning of the algorithm. Applications to changepoint models (Green 1995) as well as to modeling an unknown number of mixtures (Richardson and Green 1997) have been demonstrated.

Carlin and Chib (1995) provide an alternative approach to Bayesian model selection by using "linking densities". Adapting their approach to our context, we would define (K + 1) models to cover the possible cases of 0 to a maximum of K changepoints. Each MCMC iteration would involve taking draws for every model's parameters, even for models not drawn in the current iteration. They suggest running the algorithm separately for each model to obtain posterior draws which can be used as a pseudoprior or linking density in the next MCMC stage which includes model selection as a step. This avoids changing dimensionality at the cost of maintaining a set of models and the associated increase in memory storage. The parameters of the selected model are drawn using the usual Gibbs or Metropolis-Hastings sampler. In addition, parameters of models not selected are also drawn using the linking density. The limitation of this approach is that the number of models needs to be small to avoid computational challenges, but it provides a practicable solution when the number of changepoints is capped to a small value.

Chib (1998) formulates a multiple changepoint model in terms of a latent discrete state variable that indicates the regime from which an observation is drawn. This enables the change-point probability to vary depending on the regime, unlike the constant probability of change assumed in Barry and Hartigan (1993). This approach is analogous to using a hidden Markov model (HMM) to guide changepoint transitions, with restrictions in the state transition matrix. The number of changepoints needs to be fixed to be able to define a state transition matrix.

There is also a literature stream that 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, these models do not necessarily impose structure in the form of contiguous blocks of cohorts within regimes. Since our goal is to detect changepoints as well as model the distributions in different regimes, we do not utilize these approaches for our cross cohort changepoint model.

Either of Green (1995) or Carlin and Chib (1995) should be adequate for our context. At the timing of writing this first year summer paper, the general changepoint model is still under development, and our model framework fixes K at 1 (exactly one changepoint) with unknown location. A future version of this paper will implement the general changepoint model.

Cohort-level models for discrete-time data

For the non-profit data set that we analyze, we need to specify a cohort-level model. Fader, Hardie and Shang (2010) demonstrate the superior fit of the discrete-time non-contractual BG/BB model of customer behavior in contexts where the firm is interested in the customer response or purchase incidence with respect to a campaign, versus continuous-time models such as the Pareto/NBD model (Schmittlein, Morrison and Colombo 1987). The BG/BB model characterizes cohort heterogeneity in the probability of transaction incidence and the probability of dropout (which is a latent process not observed in the data). The rationale is that we can identify both dropout and transaction propensities at an individual level by observing the overall pattern of behavior. A very frequent donor who subsequently halts all activity, might be someone with a high transaction propensity, but also a high dropout propensity: while he is active, he consumes (i.e. donates) heavily; after a certain point, he ceases all activity. On the other hand, we can draw different conclusions about another donor who has the same aggregate amount of activity (in whatever measure) as the previous donor, but spreads the

activity in terms of occasional consumption. Such a person might be someone with a low transaction propensity, but also a low dropout propensity.

We describe the details of the non-profit dataset in the Empirical Analysis section, but note here that its characteristics are amenable to a discrete-time non-contractual model of donation incidences of individuals over time. We therefore adopt the BG/BB model for our cohort-level likelihood function. Fader, Hardie and Shang (2010) note that empirical identification becomes an issue for cohorts with few observations, and suggest pooling data from multiple cohorts to estimate one set of parameters in a Maximum Likelihood setting. This of course does not allow for possible changepoints across cohorts. With the Hierarchical Bayesian specification, we allow each cohort to have its own set of parameters, and use the hierarchical block prior distribution to allow for proper Bayesian shrinkage for those cohorts with insufficient data for identification.

In the next section, we describe a Hierarchical Bayesian cross cohort changepoint model where the number of changepoints and their locations are unknown, and are to be estimated along with the hierarchical block priors and cohort-level BG/BB model parameters.

III. Model Development

Changepoint configuration and hierarchical block distributions

There are three levels of hierarchy in our model: the changepoint configuration (random scalar K denoting number, and random vector Q denoting locations of changepoints), the (K+1) hierarchical prior distributions for each block of contiguous cohorts (Ω), and cohort-level parameters. With this hierarchy, each cohort *i* has its own set of parameters (η_i) that model the underlying phenomena of interest in the cohort-level data (Y_i). We denote the set of all cohorts' parameters as η and the set of all cohort data as Y. Based on these hierarchical relationships, the following properties ensue: (1) K and Q conditioned on Ω and η are independent of Y; (2) Ω conditioned on K, Q and η is independent of Y. This is easily seen in the graphical representation of the model shown in Figure 1, and is important to note as we describe the model and its conditional distributions.

[INSERT FIGURE 1 HERE]

The full conditional distributions for K, Q, Ω and η are needed for the MCMC estimation to converge to the joint posterior density $p(K=k, Q=\{q_1,...q_k\}, \Omega, \eta \mid Y)$.

The conditional distribution for the changepoint configuration is

$$
\begin{aligned}
& p(K=k, Q=\{q_1,...q_k\} \mid \Omega, \eta) \\
& \propto p(\Omega, \eta \mid K=k, Q=\{q_1,...q_k\}) \cdot p(K=k, Q=\{q_1,...q_k\}) \\
& = p(\Omega \mid K=k, Q=\{q_1,...q_k\}) \cdot p(\eta \mid \Omega, K=k, Q=\{q_1,...q_k\}) \cdot p(K=k, Q=\{q_1,...q_k\})
\end{aligned}
$$

For the remainder of the paper, we describe the implementation and results for a changepoint model with K fixed as 1 (i.e. exactly one changepoint) but with a random location Q (refer to Appendix section B for discussion on what a multiple changepoint would entail). This is implemented using a data augmentation approach (Tanner and Wong 1987) to draw the changepoint location, and then draw model parameters conditioned on the location. The prior on the changepoint configuration $p(K=k, Q=\{q_1,...q_k\})$ can now be defined as $p(Q=q)$ where Q is the index of the cohort where the changepoint occurs.

We adopt the convention that a changepoint at index Q=*q* in a sequence of cohorts indexed from 1 to n_c (where n_c is the total number of cohorts in the sequence) divides the sequence into two blocks {1,…,q-1} and {q,…, n_c}. The value of Q can range from 2 to n_c. Each block has its own hierarchical prior ω_b. Let Ω be the set of prior parameters for the two blocks $\{\omega_1, \omega_2\}$. Let η, as previously defined, be the set of cohort parameter vectors for the sequence of n_c cohorts (each cohort having its own parameter vector η_i).

The probability of *Q = q* conditional on Ω and η is given by Bayes rule:

$$
p(Q=q \mid \Omega, \eta) = \frac{p(Q=q) \cdot p(\Omega, \eta \mid Q=q)}{\sum_{j=2}^{n_c} p(Q=j) \cdot p(\Omega, \eta \mid Q=j)} = \frac{p(Q=q) \cdot p(\eta \mid Q=q, \Omega) \cdot p(\Omega \mid Q=q)}{\sum_{j=2}^{n_c} p(Q=j) \cdot p(\eta \mid Q=j, \Omega) \cdot p(\Omega \mid Q=j)}
$$

We can simplify the above conditional probability by noting that $p(\Omega \mid Q=q)$ is equal to the prior probability $p(\Omega)$ since Q conveys no information to update $p(\Omega)$ in the absence of η (see Figure 1).

$$p(Q=q\,|\,\Omega,\eta)=\frac{p(Q=q)\cdot p(\eta\,|\,Q=q,\Omega)}{\sum_{j=2}^{n_c} p(Q=j)\cdot p(\eta\,|\,Q=j,\Omega)} \tag{1}$$

By defining prior $p(Q=q)$, Equation (1) now allows us to evaluate $p(Q=q\,|\,\Omega,\eta)$ at each possible value of Q, and draw from a multinomial distribution as a Gibbs sampling step. The uniform distribution would be the non-informative choice of prior for $p(Q=q)$. If prior knowledge is available suggesting that a certain location is more likely, this knowledge can be embedded in the appropriate probabilities for each location *q*.

Since our data set involves a discrete-time non-contractual setting, we use the BG/BB model at the cohort level. The BG/BB model does not have a conjugate prior and we therefore use a multivariate normal block prior for a specific parameterization of the BG/BB model that we will discuss below.

Ω is thus defined by two sets of mean and covariance parameters corresponding to each block's multivariate normal distribution.

$$\Omega=\{\omega_1,\omega_2\}=\{\mu_{g_1},\Sigma_{g_1},\mu_{g_2},\Sigma_{g_2}\}$$

The probability of cohort *i*'s parameter vector η_i given Ω and Q is $p(\eta_i\,|\,\Omega,Q=q)=MVN(\eta_i;\mu_{g_b},\Sigma_{g_b})$ where the block index *b* is 1 if i < q, and 2 otherwise. Every cohort must be assigned to one of the two blocks by the changepoint configuration.

$$\Sigma_{g_b}\sim IW(\nu_0,\Sigma_0) \tag{2}$$

$$\mu_{g_b}\,|\,\Sigma\sim MVN(\mu_0,\Sigma/\kappa_0) \tag{3}$$

$$\kappa_0=0.01;\ \mu_0=\begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix}^T;\ \nu_0=7;\Sigma_0=I_4$$

The hyperpriors for Σ_{g_b} and $\mu_{g_b}\,|\,\Sigma$ are the Inverse Wishart (Equation 2) and Multivariate Normal (Equation 3) distributions. This Normal-Inverse-Wishart parameterization is used by both Murphy (2007) and Gelman et al (2004). As is typical in changepoint models (Barry and Hartigan 1993), block parameters are assumed to be conditionally independent across blocks, given η and Q. We set the hyperprior constants to be non-informative resulting in a reasonably diffuse prior (Rossi and Allenby 2003).

The parameters of the Normal-Inverse-Wishart model for each block *b* can be updated using η and Q . The conditional posterior $p(\Omega \mid Q = q, \eta) \propto p(\eta \mid Q = q, \Omega) \cdot p(\Omega)$ is now defined by the product of two updated Normal-Inverse-Wishart distributions. See Appendix section A for details of Bayesian updating.

Given Equations (2) and (3), and the assumption of a non-informative uniform prior $p(Q = q)$, we can define equation (1) as a product of two multivariate normal distributions (from the two blocks). The log posterior density for Q is proportional to $\log p(\eta \mid Q = q, \Omega)$ and can be expressed as follows.

$$
\begin{aligned}
&\log p(Q = q \mid \Omega, \eta) \\
&\propto \log p(\eta \mid Q = q, \Omega) \\
&= \sum_{i=1}^{q-1} \log p(\eta_i \mid \mu_{g_1}, \Sigma_{g_1}) + \sum_{i=q}^{n_c} \log p(\eta_i \mid \mu_{g_2}, \Sigma_{g_2}) \\
&\propto -\frac{1}{2} \left[\begin{array}{l} \sum_{i=1}^{q-1} (\eta_i - \mu_{g_1})^T \cdot \Sigma_{g_1}^{-1} \cdot (\eta_i - \mu_{g_1}) + \sum_{i=q}^{n_c} (\eta_i - \mu_{g_2})^T \cdot \Sigma_{g_2}^{-1} \cdot (\eta_i - \mu_{g_2}) \\ + (q-1) \log \left| \Sigma_{g_1} \right| + (n_c - q + 1) \log \left| \Sigma_{g_2} \right| \end{array} \right]
\end{aligned}
\tag{4}
$$

Equation (4) highlights that $\log p(\eta \mid Q = q, \Omega)$ is proportional to sum of the squared Mahalanobis distances (Bishop 2006) from each block mean for cohorts assigned to that block (given *q*), along with two terms relating to the amount of variance in each block.

We contrast equation (4) to a constrained k-means clustering approach (Bradley et al 2000; Wagstaff et al 2001), where block membership is required to be contiguous. Firstly, the block means and covariances are not drawn from posterior distributions but obtained directly from the data. The choice of *q*, rather than being drawn from a multinomial distribution would be the following optimization.

$$
q^* \in \arg\min_{q,\Omega} \left[\sum_{i=1}^{q-1} (\eta_i - \mu_{g_1})^T \cdot \Sigma_{g_1}^{-1} \cdot (\eta_i - \mu_{g_1}) + \sum_{i=q}^{n_c} (\eta_i - \mu_{g_2})^T \cdot \Sigma_{g_2}^{-1} \cdot (\eta_i - \mu_{g_2}) \right]
$$

The Hierarchical Bayesian machinery enables us to understand how uncertainty in one parameter can impact uncertainty in other parameters. We discuss the impact of uncertainty when analyzing the posterior predictive distributions from the model. This would not be possible using a constrained k-means approach.

We now present an extension of the changepoint model that can provide more detailed diagnostics about cross-cohort shifts, with the addition of new assumptions about the correlation among parameters in η. Suppose that some or all of the cohort-level parameters represented in η are mutually uncorrelated[1]. We can then model separate block priors for subgroups of parameters in η and also allow for a different changepoint configuration for each subgroup. This leads to a vector-changepoint model as opposed to the scalar-changepoint model we defined earlier in which all parameters of η contribute to the identification of a "global" discrete shift. In contrast, the vector-changepoint model identifies discrete shifts along each subgroup's parameter space, partitioning data for each subgroup in a data-driven fashion[2]. If every element of η is mutually uncorrelated with other elements, all the results from equations (1) through (4) hold for each element, resulting in a vector of Q's and block prior parameters, corresponding to each element. Note that the cohort-level prior on each parameter would then be univariate normal, with the block priors being normal inverse chi-squared distributed. We discuss the relative merits of the vector-changepoint and scalar-changepoint approaches in the Simulation Analysis section.

We also define an alternate static Hierarchical Bayesian model with no changepoints. The drawing of changepoint location step is no longer needed, and $\Omega = \{\mu_g, \Sigma_g\}$, consists of just one multivariate Normal-Inverse-Wishart distribution. We use the static model as a benchmark to determine if the changepoint model demonstrates superior prediction.

Cohort-level parameters

We now motivate our chosen cohort-level parameterization of the BG/BB model (represented by dimensions of η).

The BG/BB model provides a parsimonious framework to capture transaction incidence behavior in a cohort consisting of multiple individuals. We chose transaction incidence for a cohort over time as the primary feature of cohort behavior that we model. Future research could include monetary amount of donations as an additional feature[3] (Fader, Hardie and Lee 2005a). In the context of our data set (further

[1] Since the cohort-level parameters have a normal prior distribution, uncorrelated parameters implies independence.
[2] An advantage of dividing the parameter space into subgroups is that a changepoint in one subgroup does not hamper the other subgroup from sharing data across all cohorts, which is not possible in a scalar-changepoint model.
[3] Fader, Hardie and Lee (2005) discuss empirical evidence for modeling the distribution of monetary amounts as independent of the model of transaction incidence. If we assume independence in these two distributions, the posterior cohort-level parameter distributions would be unaffected by the addition of a monetary amount feature.

details in the Empirical Analysis section) changes in transaction incidence across cohorts provides diagnostic insights on behavioral patterns over the cohort sequence.

Hence individual behavior within a cohort is represented by a vector of binary transaction incidence choices across discrete time periods. The first observation corresponds to the first repeat transaction opportunity, and thus excludes the initial transaction which identifies the cohort an individual belongs to. Recency and frequency statistics can be computed from an individual's vector of choices. Recency indicates the most recent time period of a transaction. Frequency indicates the total number of transactions over the duration of observation.

Due to the specification of the BG/BB model (Fader, Hardie and Shang 2010), recency and frequency together serve as a sufficient statistic for that individual's behavioral pattern. Since each individual is assumed to have a Bernoulli probability *p* of making a transaction at any given time period, and a geometric distribution with parameter θ of dropping out or "dying" at the beginning of each period, the recency statistic identifies the last known period the individual was "alive". By definition, all transactions for an individual take place before or at the time period indicated by the recency statistic. Since an individual's transactions are assumed i.i.d and therefore exchangeable, the probability of any sequence of transactions with the same recency and frequency is identical and given by

$$p(x, t_x, n \mid p, \theta) = E_Z[p(x, t_x, n \mid p, \theta, Z)] = p^x(1-p)^{n-x}(1-\theta)^n + \sum_{i=0}^{n-t_x-1} p^x(1-p)^{t_x-x+i}\theta(1-\theta)^{t_x+i}$$

where x is frequency, t_x is recency, n is the number of periods and Z is the period where the individual dropped out (which is unobserved). Taking the expectation over Z results in the probability of observing $\{x, t_x, n\}$ given $\{p, \theta\}$.

Heterogeneity across individuals in cohort *i* is modeled using independent Beta distributions such that $p \sim Beta(\alpha_i, \beta_i)$ and $\theta \sim Beta(\gamma_i, \delta_i)$. Fader, Hardie and Shang (2010) also define the SBB/GB model which allows p and θ to be arbitrarily correlated. A possible future extension is to replace the BG/BB cohort-level model with the SBB/GB model.

For cohort *i* observed over n_i periods, we categorize all individuals into $J = 0.5n_i(n_i+1)+1$ possible recency-frequency combinations. Equation (5) shows the likelihood function for any Y_{ij} that is

categorized as the j^{th} recency (t_{x_j}) and frequency (x_j) pattern, given the $\alpha_i, \beta_i, \gamma_i, \delta_i$ parameters of the BG/BB model after integrating out *p* and θ. Note that $B(\cdot,\cdot)$ in equation (5) refers to the Beta function.

$$p(Y_{ij} \mid \alpha_i, \beta_i, \gamma_i, \delta_i, x_j, t_{x_j}, n_i) = \frac{B(\alpha_i + x_j, \beta_i + n_i - x_j)}{B(\alpha_i, \beta_i)} \cdot \frac{B(\gamma_i, \delta_i + n_i)}{B(\gamma_i, \delta_i)} + \sum_{i=0}^{n_i - t_{x_j} - 1} \frac{B(\alpha_i + x_j, \beta_i + t_{x_j} - x_j + i)}{B(\alpha_i, \beta_i)} \cdot \frac{B(\gamma_i + 1, \delta_i + t_{x_j} + i)}{B(\gamma_i, \delta_i)} \quad (5)$$

By counting the incidence of recency-frequency patterns within a cohort, the entire cohort's transactions can be represented by the sufficient statistic of a recency-frequency table of counts for the BG/BB model. Let Y_i represents the data set (i.e. recency-frequency table) associated with cohort *i*.

$$p(Y_i \mid \alpha_i, \beta_i, \gamma_i, \delta_i) = \prod_{j=1}^{J_i} \left(p(Y_i \mid \alpha_i, \beta_i, \gamma_i, \delta_i, x_j, t_{x_j}, n_i) \right)^{f_j} \quad (6)$$

Equation (6) describes the BG/BB model likelihood function for a cohort's data, given the four cohort-level model parameters. J_i represents the total number of recency-frequency combinations for cohort *i*. f_j represents the total consumer count for pattern *j* in the recency-frequency table. The appendix includes an example of a recency-frequency table to illustrate the data needed to compute equation (6).

To improve convergence properties for estimation we reparameterize the BG/BB model as follows. First we parameterize the two Beta distributions in terms of their means and polarizations as shown in equation (7). This has the advantage of compartmentalizing changes in parameters across cohorts, substantially reducing correlations between parameters, which may enable consideration of the vector-changepoint model. It would be more valuable to know that the mean of the transaction or dropout process has shifted across cohorts, than simply to know that the original Beta distribution parameters have changed. Second, we take the logit transform for each of $\mu_{p_i}, \phi_{p_i}, \mu_{\theta_i}, \phi_{\theta_i}$, since each of these parameters is bounded between 0 and 1.

$$\mu_{p_i} = \frac{\alpha_i}{\alpha_i + \beta_i}; \phi_{p_i} = \frac{1}{1 + \alpha_i + \beta_i}; \mu_{\theta_i} = \frac{\gamma_i}{\gamma_i + \delta_i}; \phi_{\theta_i} = \frac{1}{1 + \gamma_i + \delta_i} \quad (7)$$

$$\text{logit}(\mu_{p_i}) = \log(\frac{\alpha_i}{\beta_i}); \text{logit}(\phi_{p_i}) = -\log(\alpha_i + \beta_i); \text{logit}(\mu_{\theta_i}) = \log(\frac{\gamma_i}{\delta_i}); \text{logit}(\phi_{\theta_i}) = -\log(\gamma_i + \delta_i)$$

The parameter vector $\eta_i : \{ \text{logit}(\mu_{p_i}), \text{logit}(\phi_{p_i}), \text{logit}(\mu_{\theta_i}), \text{logit}(\phi_{\theta_i}) \}$ captures the transformed parameterization. Thus the prior for η_i can be represented by a multivariate normal distribution from the block it is assigned by the changepoint location Q.

$$p(\eta_i \mid Y_i, Q = q, \Omega) \propto \prod_{j-1}^{J} \left(p(Y_i \mid \eta_i, x_j, t_{x_j}, n_i) \right)^{f_j} \cdot p(\eta_i \mid Q = q, \Omega) \tag{8}$$

The posterior density of η_i defined up to a proportionality constant is given in equation (8). This completes the model specification of the cross-cohort changepoint model for a discrete-time setting.

IV. Model Estimation and Identification

The MCMC pseudo-algorithm for the scalar-changepoint model is described in Figure 2. Note that we can also obtain the static Hierarchical Bayesian model, by simply assigning all the data to one of the block prior parameters and essentially ignoring the other block parameters. We would also remove the changepoint location draw step, replacing it with a constant location (e.g. q = 1) that essentially contains all cohorts in one block. We compare the performance of the changepoint model to that of the static model in the empirical analysis section.

[INSERT FIGURE 2 HERE]

Empirical identification of the BG/BB model is a known issue when very few observations (1 or 2 periods of data) are available. Because the BG/BB model looks to capture both overt transaction behavior and latent dropout behavior, multiple explanations can exist for limited data. A cohort with 1 or 2 observations tends to have an informative but multi-modal likelihood function. We cannot therefore identify the parameters of the latent dropout process which are critical for predicting future behavior of this cohort[4] by accounting for churn in the customer base. If such cohorts are included in the MCMC estimation, their parameters may not converge, affecting the hierarchical block parameters as well. Hence, a cohort's data is only included in the model estimation if it has at least three periods of observation, which in simulation testing has been shown to have good convergence properties.

[4] With the emphasis on prediction of holdout data, it is not of interest to include a cohort with 1 or 2 periods, just to improve in-sample fit for the cohort.

In the case of cohorts acquired sequentially, we can choose to analyze the data by assuming a cut-off year, and examining data from cohorts observed up to that point. Due to our "3 observation minimum" requirement, the most recent two cohorts will necessarily be excluded from the model. For example, if we examine data up till and including 2001, the latest cohort included would be the 1998 cohort, and we would exclude the 1999 and 2000 cohorts which would have 2 and 1 repeat observations respectively.

The question then arises of how we predict behavior for cohorts that fall below the "3 observation minimum" requirement. We are able to take advantage of the hierarchical block parameters to draw cohort-level parameters for those cohorts with limited or even zero data. Since excluded cohorts will always follow to the right of included cohorts (as a sequence), we use the mean and covariance parameters of the second block to generate cohort parameters η_i from the multivariate normal prior distribution.

To provide a specific example, suppose that 1990 is the oldest cohort, 2000 is the data cut-off year, and that the model detected a changepoint at 1996 with high posterior probability. The cohorts exerting the greatest influence on the second block's parameters would be the 1996 and 1997 cohorts. The first block would have six cohorts (1990 – 1995) driving its parameters, and would be expected to have lower uncertainty in both the mean and covariance distributions. However, if a changepoint is indeed accurately detected at the 1996 cohort, we should expect a bias-variance tradeoff between the posterior predictive distributions of the changepoint versus the static model for excluded cohorts (1998 and 1999 cohorts). Specifically, the changepoint model should have lower bias than the static model, in terms of the centering the predictive distribution closer to the actual values. However, the predictive distributions of excluded cohorts should exhibit larger variance when generated from the changepoint model since by definition the block parameters are based on fewer cohorts than the static model. For prediction, a lower bias is advantageous and we use holdout data to compare the two models in the Empirical Analysis.

We note that cohorts included in model estimation have very informative likelihood functions, due to the large number of individuals in each cohort. These cohorts will exhibit little to no shrinkage as a result, and therefore relatively unaffected by the choice of prior (whether coming from a changepoint or static model). Hence, we expect the predictive distributions for these cohorts to be very similar across the two models. Since our emphasis is on predicting behavior of new cohorts with limited observations, this is not an aspect of the data that is relevant for model comparison.

We used the following configuration to run the MCMC algorithm on the actual data set. The first 10,000 iterations are utilized as burn-in. Most parameters converged within 2,000 iterations but this buffer helps ensure that all model parameters converge (according to the Gelman-Rubin statistic by running 3 chains with over-dispersed initial parameter values). Autocorrelation is significant for cohort level parameters (especially ϕ_{θ_i} which represents the polarization for the dropout Beta distribution) and requires considerable thinning to result in approximately i.i.d draws. We ran 550,000 iterations, and thinned aggressively after burn-in, by taking every 500th draw. These draws exhibited very low autocorrelation and can be considered a reasonable approximation to i.i.d samples for the purposes of estimating posterior intervals. The per-iteration run time was about 0.052 seconds on a 4-core 2.70GHz processor in the R programming environment, with a total run time of approximately 8 hours. The appendix contains computational details for the MCMC algorithm.

In the next two sections, we describe the results from running the described MCMC algorithm on simulated and actual data sets.

V. Simulation Analysis

We performed a simulation analysis to answer the following questions: (1) can the scalar-changepoint model detect discrete shifts in a cohort sequence? (2) how does the scalar-changepoint model behave when the data does not admit a clear changepoint? (3) how does the scalar-changepoint model perform versus the vector-changepoint model if shifts occur in different parameters, at different cohorts?

All of our simulations were based on a sequence of 11 cohorts to mirror the total number of cohorts available in the non-profit data set. Each cohort was assigned a four-dimensional parameter vector $\{\mu_{p_i}, \phi_{p_i}, \mu_{\theta_i}, \phi_{\theta_i}\}$, and data was simulated for a cohort using the assigned parameters, assuming 6 periods of observations for each individual and 10,000 individuals per cohort. We standardize the number of periods and cohort size as our objective in running simulations is to test the changepoint detection capability of the algorithm, and not to generate predictive distributions for new cohorts.

The results are obtained by running 50,000 MCMC iterations, allowing for 20,000 iterations as burn-in and drawing 1,000 thinned samples from the post-burn-in period. All cohort parameters and block priors converge within the burn-in period (based on Gelman-Rubin statistic, running 3 chains with over-dispersed initial values). The posterior distribution of the changepoint location is compared to the actual location to assess convergence and accuracy.

Figure 3a shows that a changepoint is clearly detected when there is a fairly large discrete jump in one of the parameters (we pick μ_{p_i}) in three different scenarios (shift at cohort 2, 6 and 10 respectively). The remaining parameters $\{\phi_{p_i}, \mu_{\theta_i}, \phi_{\theta_i}\}$ are similar across cohorts. Figure 3b shows the posterior interquartile range for μ_{p_i} across cohorts, showing that the cohort-level parameter has been identified with a tight distribution, and we observe the jump at cohort 6 corresponding to the scenario where we coded the jump at this cohort. Figure 3c presents the posterior interquartile range of each hierarchical block mean for the cohort 6 jump coding. We observe that the first block's mean reflects the higher value of $\mu_{p_i} = 0.6$ set for the first five cohorts. The second block's mean also reflects the lower value of $\mu_{p_i} = 0.3$ set for the last six cohorts. Appropriate posterior block mean distributions are obtained for the other jump codings. Hence, our first objective of detecting discrete shifts using the scalar-changepoint model has been satisfied.

[INSERT FIGURE 3a,b,c HERE]

Figure 4a shows the result from having no changepoint in the sequence of 11 cohorts – only small fluctuations in parameters from cohort to cohort are encoded, with no large discrete jumps (as seen in the posterior interquartile range for μ_{p_i} across cohorts in Figure 4b). We observe the apparently counter-intuitive assignment of roughly symmetric high posterior probability to cohort 2 and cohort 11. In fact, this assignment is consistent with the estimation of the joint posterior density of the scalar-changepoint model. In particular, the draws of the changepoint location and block distributions need to be viewed holistically. Because our model has the number of changepoints fixed at one, which is a mismatch to the data which does not admit a discrete shift, the changepoint location is pushed to either end of the sequence to maximize the number of cohorts in a block[5]. We observe from Figure 4c that the posterior interquartile ranges of the two block means are fairly similar. When we move to a general changepoint model that would nest the possibility of K = 0, this mismatch should no longer occur. Hence, our model is able to detect the likely absence of a changepoint via this assignment of probabilities.

[INSERT FIGURE 4a,b,c HERE]

[5] The assignment of changepoint posterior probability to the ends of the cohort locations where a changepoint is permissible, is observed whenever the data does not suggest a sizeable discrete jump. For the K = 1 model we currently have, the observation of symmetric probability to q = 2 and q = 11 is an indication of the lack of evidence for a changepoint. Should there be a discrete jump at one of the end locations such as q = 2 (Figure 3a), this is clearly identified by the changepoint model.

Our final simulation compares the scalar-changepoint model whose performance we have characterized thus far, with the vector-changepoint model. Figure 5 shows the result for the scalar-changepoint model for a cohort sequence with two different discrete jumps, occurring in two different parameters of η. Since the scalar-changepoint model is constrained to fit one changepoint, its performance is sub-optimal when presented with a data set with multiple changepoints. From Figure 5a, we observe that a high posterior probability is assigned at cohort 6. The other jump at cohort 3 is picked up by the model by assigning the second highest posterior probability, but is overwhelmed by the jump at cohort 6. This is entirely consistent with how our model with one changepoint is expected to behave. However, we now examine the performance of the vector-changepoint model in being able to detect multiple shifts if they occur in different parameters, even without migrating to a multiple-changepoint framework.

[INSERT FIGURE 5a,b,c HERE]

We can make the assumption that the parameters of η are mutually independent since this is a simulated data set. When we use the vector-changepoint model based on this assumption, any correlation between pairs of parameters in the simulated data is ignored. A vector of posterior changepoint location distributions is obtained corresponding to $\{\mu_{p_i}, \phi_{p_i}, \mu_{\theta_i}, \phi_{\theta_i}\}$. Figure 6a shows the results of posterior changepoint probability for each parameter. Changepoints are accurately detected at cohort 3 for μ_{p_i} (transaction process mean) and cohort 6 for μ_{θ_i} (dropout process mean). The changepoint probability assignments for ϕ_{p_i} and ϕ_{θ_i} are reflective of no changepoints in these parameters (see previous discussion relating to Figure 4). In Figure 6b, discrete jumps can be observed at cohort 3 and cohort 6 in the μ_{p_i} and μ_{θ_i} parameters respectively (very similar to Figure 5b). In Figure 6c, note that each parameter of the block mean has its interquartile range deriving from its own changepoint configuration. The two blocks have markedly different distributions for parameters 1 and 3, while parameters 2 and 4 are similar (since the data does not shift in these parameters). The results show the advantage of the vector-changepoint model[6] in detecting multiple shifts in a cohort sequence, if they occur in different parameters of the cohort-level model. When considering a multiple-changepoint model, the vector-changepoint approach can still be applicable to detect *multiple* shifts in different parameters.

[6] The vector-changepoint model also improves speed of computation since a k-dimensional multivariate distribution is replaced with k univariate normal distributions. However, this alone is not a reason to use the vector-changepoint model over the scalar-changepoint model.

[INSERT FIGURE 6a,b,c HERE]

A cautionary note here is to verify the assumption of mutual independence among cohort-level parameters. Running the scalar-changepoint model and examining the posterior distribution of correlation coefficients between pairs of parameters would provide some indication of whether the assumption of zero correlation can be justified. The researcher can still impose the assumption of independence if this can be inferred from prior knowledge or previous data sets, at the cost of distorted results should the assumption be false.

Judicious selection of parameter sub-groups can provide managerially relevant changepoint signals. For instance, a researcher may decide that the transaction process parameters $\{\mu_{p_i}, \phi_{p_i}\}$ form one subgroup while the dropout process parameters $\{\mu_{\theta_i}, \phi_{\theta_i}\}$ form another subgroup, and that these subgroups are mutually independent. The covariance matrix for the cohort-level prior then has a block-diagonal structure, and is a middle path to avoid assuming that *all* parameter pairs are mutually independent. As the number of features modeled at the cohort level increases, the vector-changepoint model can bring additional insights and computational savings when the assumption of independence is justified.

VI. Empirical Analysis

To test the model's capabilities with actual data, we obtained a multi-cohort donor data set from a regional U.S. non-profit public television station. The 11 cohorts in the data set were acquired from 1990 through 2000. The data for each cohort runs from the year of acquisition through to the end of 2009. For each cohort, the raw data contains transaction dates and donation amounts for each donor. We convert this data into a binary sequence of annual transaction incidences for each donor. The range of a cohort's size is from 8,900 to 23,153 donors. In this context, focusing on yearly giving behavior as the time scale is appropriate since the non-profit organization plans a set of campaigns across the year (involving TV appeals and direct mail). Choosing a time scale shorter than a year can result in variations across periods that are due to seasonality or other temporal factors rather than stable donor behaviors[7], which would distort cohort parameter estimates. In other contexts, it may be that the time scale for a

[7] Fader, Hardie and Shang 2010 found that using a time scale shorter than annual incidence results in undesirable data distortions in a non-profit setting.

period is quarterly, monthly or even weekly. Our changepoint model can be applied to any of these settings, and the only differences would be in the number of observations.

The management team of the television station was interested to know if the characteristics of their donor base had changed over time, whether the change was gradual or involved a distinct changepoint, and how to best forecast the behavior of new cohorts. We estimate the changepoint and static models to address these questions. The changepoint model provides the posterior probability of a shift at each possible cohort. The cohort-level parameter distributions tracked across the cohort sequence provides guidance on the nature of the shift (if one exists). Posterior predictive distributions for new cohorts obtained from the changepoint and static models are compared with actual data to determine which model is a better fit.

With the plethora of quantities available to assess model fit, we focused on two measures: posterior predictive distributions of *holdout data* for the newest cohort in our data set (acquired in 2000), and log marginal likelihood. Per the earlier discussion on empirical identification of the BG/BB model and the highly informative likelihood function for cohorts with at least three periods of repeat observations (we use the terms 'observation' and 'repeat observation' interchangeably), parameters of cohorts with sufficient data undergo little to no shrinkage. Therefore, in-sample features of the data are fit equally well by the changepoint and static models, which have very similar cohort-level posterior distributions. We return to this point when presenting the log marginal likelihood comparison. As a result, we expect using a five year holdout period for the 2000 cohort's data to provide a solid measure of each model's predictive accuracy and face validity in capturing features of the data set not used in model estimation.

To provide a true test of the difference between the changepoint and static models, we deliberately picked "cut-off" years where the 2000 cohort would have very few or no observations[8]. The cut-off years we use are 2000, 2001, 2002, and 2003. In Figure 7, we show how cohorts included for model estimation at each cut-off year have differing periods of repeat observations, with the 1990 cohort always having the largest number of observations. Cohorts with fewer than three observations are excluded from model estimation but we can still generate parameter draws and posterior predictive distributions of interest using the estimated model's MCMC draws. For example, the 2000 cohort has no repeat observations in the year 2000, and could not be estimated at all using Maximum Likelihood methods.

[8] Note that our data set contains transactions up to 2009 for every cohort. If we used all this data, each cohort from 1990 to 2000 would have more than the minimum three observations required for identification, and there would be practically no difference in prediction between the changepoint and static models. Hence, we use a more realistic test putting ourselves back in time such that the 2000 cohort has very few observations.

Our hierarchical Bayesian models enable predictions to be made by borrowing information either from all cohorts (static model) or *relevant cohorts* if a changepoint is detected (changepoint model). In 2003, the 2000 cohort would have 3 observations and is included in model estimation and we will observe that the predictive distributions of the changepoint and static models become very similar. We compute mean squared error (MSE) across predictive distribution draws and decompose MSE into the bias and variance components to highlight the difference between the changepoint and static models.

[INSERT FIGURE 7 HERE]

We also considered whether the vector-changepoint or scalar-changepoint model would be appropriate for analyzing this data set. Based on the discussion in the Simulation Analysis section, we first checked the correlation between pairs of parameters using the scalar-changepoint model, shown in Table 2. We observe a large 95% posterior interval width for all possible correlation coefficients for pairs of parameters in each block. We therefore defer the analysis using vector-changepoint model until potential distortions it may introduce are investigated. Presently, we use the scalar-changepoint model which allows for correlations between parameters.

[INSERT TABLE 2 HERE]

We start by looking at posterior probability of changepoints as we move from the 2000 cut-off to the 2003 cut-off. This provides a view of what management could have taken note of in "real time".

Using the progression of the four cut-off years from 2000 to 2003 shown in Figure 8, we can infer the following shifts in the cohort sequence. There appears to have been a moderate shift detected at the 1997 cohort in the year 2000. However, this has been superseded by a larger shift at the 1998 cohort (which was only possible to detect once the 1998 cohort had enough data to be included in model estimation in 2001). In 2002 and 2003, the 1998 cohort continues to have a large posterior probability of a change. This pattern therefore suggests that cohorts acquired in 1998 or later have different behavioral characteristics compared to pre-1998 cohorts.

[INSERT FIGURE 8 HERE]

In which parameters might shifts have occurred which lead to the above changepoint probabilities? We examine the 2002 cut-off as an illustrative example. In Figure 9, we plot the posterior interquartile ranges across cohorts for each parameter in $\{\mu_{p_i}, \phi_{p_i}, \mu_{\theta_i}, \phi_{\theta_i}\}$. We observe that the transaction process

parameters appear to be relatively similar across cohorts. In the dropout process parameters however, we see a substantial shift in both the mean and polarization at the 1998 cohort, and a further (but smaller) shift at the 1999 cohort. This corresponds to the high posterior probability of a changepoint at the 1998 cohort, and the next-highest probability at the 1999 cohort for the 2002 cut-off depicted in Figure 8. The managerial implication is that the 1998 and 1999 cohorts appear to have a much higher dropout rate on average than their predecessors, and with greater heterogeneity as well in the dropout rate. The television station would need to examine any environmental or policy changes that may have resulted in these poorer-performing cohorts.

[INSERT FIGURE 9 HERE]

Based on the evidence pointing towards a changepoint at the 1998 cohort, as well as the parameter shifts, we should expect to find corroborating evidence in the predictive accuracy of the changepoint model versus the static model. Specifically, we can hypothesize that the static model would overpredict the number of transactions of the 2000 cohort over the holdout period, compared to the changepoint model. This follows from the increased dropout mean evident in the 1998 and 1999 cohorts, which would get "smoothed over" by the static model.

We now examine the predictive performance of the two models with respect to five year holdout data of the 2000 cohort. For the 2000 cut-off, the holdout period is 2001-2005; for the 2001 cut-off, the holdout period is 2002-2006, and so forth. Apart from the 2003 cut-off, the 2000 cohort is not included in model estimation for the other cut-off years. Hence, how closely the posterior predictive distribution characterizes the actual holdout data is an indication of the resemblance of the 2000 cohort to its predecessors. Figure 10 shows the actual holdout transactions versus the mean of the posterior predictive distribution from the changepoint and static models. From Figure 10, we observe that the mean of the posterior predictive distribution for the changepoint model is centered closer to the actual holdout values for every holdout period and every cut-off year except 2003. In 2003, the 2000 cohort has enough observations to be included in model estimation and the posterior predictive distributions are very similar for both models.

[INSERT FIGURE 10 HERE]

At the 2000 cut-off, data from the 1998 cohort is not included in the model estimation. Since we observe a changepoint at the 1998 cohort at a later cut-off, we can see that the bias of the changepoint model's predictive distribution, while lower than that of the static model, is still off from the actual

values. At the 2001 cut-off, we observe a much improved fit using the changepoint model, while the static model has a large positive bias since it is influenced by older cohorts which seem to have a much lower dropout propensity than the 2000 cohort. In the 2002 cut-off, the changepoint model fit is good for the first two holdout years, and then deviates with an upward bias. The explanation for this can be found in Figure 9 which shows the 1999 cohort with a dip in the transaction process polarization relative to other cohorts. Because the scalar-changepoint model we are using heavily weights the 1998 and 1999 cohorts in the estimation of the second block's parameters, all parameters from these two cohorts (including of the transaction process) influence the 2000 cohort predictive distribution. Thus, the 1999 transaction process polarization causes the prediction for the 2000 cohort to be upwardly biased. The vector-changepoint model may be useful in such instances to avoid the tendency of the scalar-changepoint model to assign all parameters to the same block structure due to a large shift in a subset of the parameters. Finally, Figure 10 shows that the holdout predictions for the changepoint and static models become practically identical in the 2003 cut-off, since the 2000 cohort's own data are primarily driving its parameter draws, with the information "borrowed" from the 1998 and 1999 cohorts having a weak influence. That both models produce biased predictions for the 2003 cut-off is a reflection that the trajectory of the 2000 cohort's predictive distribution is not entirely given by information in the three observations of this cohort[9] (from 2001-2003).

While we show the holdout performance over five years to illustrate overall fit, the most important prediction is for the following year because the model can be re-estimated each year as new data arrives. Therefore, predictions made for the longer-term can be periodically revised. The changepoint model provides predictions with much lower bias than the static model, which helps forecast behavior of the 2000 cohort even when there is very little to draw from that cohort's data itself. The lower bias of the changepoint model does not come without a cost. In Table 3, we decompose the mean squared error across posterior predictive draws into bias and variance components. Per the discussion in the Model Estimation section, we expect a higher variance for the changepoint model. We show the results for the first holdout year in each cut-off, but the general pattern of a bias-variance tradeoff persists across each of the holdout years.

[INSERT TABLE 3 HERE]

[9] While a cohort's BG/BB likelihood function is identified with three observations, the shape of the likelihood function can still fluctuate as new observations are added, and only "settle" after more than three observations. This can vary from cohort to cohort.

From Table 3, the changepoint model has an overall higher MSE which is driven primarily by variance in the posterior predictive distribution, since block parameters are driven by fewer cohorts and exhibit more uncertainty. The static model has a lower MSE but has a large bias component. The posterior predictive p-values also indicate that the changepoint model has better face validity in explaining the holdout data versus the static model whose p-values are on the higher end of the spectrum. These results are consistent with Figure 10, indicating that the 2000 cohort is significantly underperforming in terms of transactions compared to earlier cohorts, and the changepoint model is able to pick up on some of the change and shield the predictions for the 2000 cohort from being influenced by the older cohorts. Notably, the 2003 cut-off uses the 2000 cohort's data for its predictions, resulting in a much lower overall MSE and similar statistics for the changepoint and static models.

We examine "global model fit" by computing log marginal likelihood for each model and cut-off year using the harmonic mean method (Newton and Raftery 1994), as shown in Table 4. We observe both models to have very similar log marginal likelihoods, because cohorts included in model estimation have at least three observations and have very informative likelihood functions due to the large number of donors in a cohort. Thus, little to no shrinkage towards the prior is observed for cohorts included in model estimation for either model, and cohort level parameters are tightly distributed. Cohort-level parameters converge to similar posterior distributions in both models, leading to similar predictive distributions.

[INSERT TABLE 4 HERE]

We note that the changepoint model's primary purpose is to detect discrete shifts across cohorts, and enable improved prediction of *new* cohorts with limited observations. Therefore, the log marginal likelihood comparison is of limited value compared to the posterior predictive checks we discussed earlier.

VII. Discussion and Managerial Implications

The comparison of the changepoint and static models shows that utilizing the changepoint framework provides new information about the structure in a cohort sequence, and that this new information plays a significant role in enhancing predictive accuracy of holdout data by reducing bias.

We return to the managerial questions posed at the beginning: (1) when does data become outdated for forecasting behavior of new cohorts, (2) how to develop forecasts for new cohorts, and (3) is there evidence for a distinct shift in cohort characteristics in a sequence of cohorts?

Bob, the analyst from Acme Apparel, has drawn parallels from the non-profit data set to his own context. Using the changepoint model, he saw that a line could be drawn at the cohort where the posterior probability of a changepoint is highest. He is now able to determine which cohorts would be relevant predecessors for the new cohorts that Anita, the sales manager, is interested in. In the non-profit example, cohorts prior to 1998 are identified by the model as being from a different "regime". Without this output from a changepoint model, Bob would understandably be reluctant to disregard 6 to 8 cohorts' worth of data to forecast new cohorts' behavior and management may be skeptical of analysis which uses less data than what is available. Thus, decision making based on heuristics of what is relevant can be augmented with data-driven evidence of regime change.

Bob has also learned that he can examine the parameters which have shifted at the changepoint location to better understand the intuition of what cohort-level behavior is responsible for the shift. He may be able to confirm Colleen's intuition by analyzing whether the transaction or dropout propensity of newer cohorts have undergone a significant shift since the new competitor's entry. He can now provide Colleen with data-driven corroboration of intuition regarding changes in the marketplace that impact customer behavior, which he was previously unable to quantify.

While the changepoint model is not dynamic in the sense of "sequential Monte Carlo" methods, it can be re-run after each period in which new data is received, thus providing a moving snapshot of cross-cohort changes. To run the MCMC algorithm for 550,000 iterations (needed for aggressive thinning) took about 8 hours on a 4-core 2.70GHz processor in the R programming environment. An analyst like Bob can therefore re-run the model on a periodic basis to determine relevant cross-cohort patterns.

Bob is keenly interested in the bias-variance tradeoff between the changepoint and static models. He understands that using fewer cohorts to estimate block parameters results in greater uncertainty in those parameters. The mean squared error of the changepoint model appears larger than the static model, but is centered closer to the actual performance. He knows that Anita, who has read Lodish (1986), would prefer to be "vaguely right" than "precisely wrong". For making predictions, it is more valuable to have low bias than low variance, if these had to be traded off.

We also note that the MSE of the changepoint prediction will reduce as more cohorts are added over time; however, the bias of the static model improves very slowly with additional data, as the older cohorts will dominate if they are greater in number than the cohorts in the second regime. Thus, while there appears to be a "tradeoff" between bias and variance when comparing the changepoint and static models, this is not a real tradeoff that a manager has to make. The changepoint prediction is more relevant, and its variance provides information on uncertainty which is also important for the manager to know. Having a false sense of certainty can lead to decisions that are sub-optimal for the organization. For instance, Anita may overcommit her sales organization to over-ambitious goals by relying on forecasts that do not take changepoints into account. Management of startup companies whose valuations rely upon customer acquisition and retention may portray a less accurate view of future growth if they project new cohorts to behave like the "average" of all previous cohorts.

For researchers, these findings suggest that investigating cross-cohort shifts can be a powerful tool when dealing with a dataset that has an inherent cohort sequence structure. New insights can be obtained with the general changepoint model framework, regardless of the type of cohort-level behavior being modeled. The absence of evidence of a changepoint can also provide confirmatory feedback to conduct analysis which uses a single population model for the entire dataset.

VIII. Extensions/Future Work

Ongoing Extensions

As described in the Model Development section, we are currently developing a multiple-changepoint model to supersede the fixed changepoint model implemented in this paper. Approaches such as reversible jump MCMC (Green 1995) and model selection using linking densities (Carlin and Chib 1995) are under consideration, and we will include implementation details for the multiple-changepoint model in a future version of this paper. The multiple-changepoint model will be valuable to allow for the detection of an arbitrary number of shifts, which may well occur in a long cohort sequence. We will extend the comparison of the scalar-changepoint and vector-changepoint approaches to the multiple-changepoint model.

We also plan to characterize the nature of distortions introduced by using a vector-changepoint model when cohort parameters are in fact correlated. If the vector-changepoint model is reasonably robust

when parameters are correlated, we can uncover more detailed insights about shifts by parameter dimension compared to the scalar-changepoint model.

Future Work

While the changepoint model provides useful insights on cross cohort shifts, it has a number of limitations, which can be addressed in future research.

First, the model does not currently have covariates which can explain potential causes of cross cohort shifts. For instance, the state of the environment and firm at the time of acquiring a cohort may explain part of the cohort's characteristics. A variety of macroeconomic and firm-internal variables may be hypothesized as reasons for the shifting behavior of customer cohorts. It would be useful to build a model with covariates that can test for a link between changepoint locations and explanatory variables. Endogeneity needs to be accounted for, especially if the covariates involve firm-internal strategies which may be influenced by the behavior of previous cohorts.

Second, it is possible that cohorts are not necessarily related to each other in contiguous blocks. Returning to the idea of a more general model of relationships amongs cohorts (e.g. Hui and Bradlow 2011), 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 will not have a cohort sequence structure, but there can still be underlying patterns which can be uncovered.

Third, our model does not allow for correlations between the transaction and dropout process parameters within a cohort. This limitation can be addressed by replacing the BG/BB cohort-level model with a more general SBB/GB model that allows for correlations.

Fourth, our model focused on transaction incidence, giving the same weight to a transaction regardless of monetary amount. Modeling transaction monetary amounts within a cohort and tracking changes in this dimension across cohorts can provide additional insights on behavioral changes at the cohort level. This would enable the firm to compute the overall cohort lifetime value and distinguish between frequent, low monetary spend customers and infrequent, high spend customers. Empirical evidence of independence between the distributions of monetary amount and transaction incidence (Fader, Hardie

and Lee 2005a) may facilitate the use of the vector-changepoint approach to separate the changes related to spend and incidence.

Finally, our changepoint model is not dynamic in the sense of a sequential Monte Carlo approach. The model has to be re-run with the entire new data set, if new time periods of observations become available. This may be reasonable if computation does not have to be in real-time. However, there can be scenarios (especially in rapidly growing online social networks), where cross cohort patterns need to be analyzed on a weekly or even daily basis. In this case, sequential Monte Carlo methods such as particle filtering could enable real-time updating of model parameters as new data comes in (Doucet, de Freitas and Gordon, 2001).

IX. Conclusions

We have demonstrated the value and importance of considering change points in multi-cohort sequences. Since many businesses acquire cohorts of customers over time, monitoring the evolution of cohort behaviors may provide a telling indication of change, either confirming managers' intuitions or providing a fresh impetus to search for external or internal shifts. Managers may discover that new cohorts' behaviors have been shifting significantly due to recent promotions by a competitor, which they had initially assumed to be ineffective. Managers may discover signs of saturation in the market they are targeting – successive new cohorts may appear to have less attractive characteristics than older cohorts, perhaps because the best prospects were already converted in the older cohorts, and the ones remaining are the "customers nobody wants to have". Lumping all customers, new and old, into one grouping will not enable the detection of such changes.

We have developed a Hierarchical Bayesian cross cohort changepoint model that provides guidance to these managerial issues. Our framework is applicable to any scenario where a sequencing of cohorts exists, longitudinal data is available for each cohort, and the cohort-level model is parametric. Our model yields probabilistic inferences about discrete shifts or changepoints in cohort characteristics across the sequence of cohorts, and provides the ability to forecast new cohorts' behavior with little or no data from the new cohort itself. We hope this work spawns further research in uncovering new structural patterns in customer databases that help managers make the most effective use of their data.

References

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

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

Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer

Bradley, P.S., Bennett, K.P., and Demiriz, A. (2000). Constrained K-Means Clustering. Microsoft Research Technical Note, MSR-TR-2000-65

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

Chib, S. (1998). Estimation and Comparison of Multiple Change-Point Models. Journal of Econometrics, 86, 221-241

Doucet, A., de Freitas, N., and Gordon, N. (2001). Sequential Monte Carlo Methods in Practice. Springer

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. (2005a). RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. Journal of Marketing Research, 42, 4, 415-430

Fader, P.S., Hardie, B.G.S, and Lee, K.L. (2005b). "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

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

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. (2011). Multi-Resolution Analysis with Applications in Marketing. Working Paper

Lodish, L.M. (1986). The Advertising and Promotion Challenge: Vaguely Right or Precisely Wrong? Oxford University Press

Madouasse, A. (2009). An Evaluation of Milk Recording, Somatic Cell Counts and Reproductive Performance in a Large Cohort of Dairy Herds in England and Wales. Doctoral Thesis – University of Nottingham

Murphy, K.P. (2007). Conjugate Bayesian Analysis of the Gaussian Distribution. Technical Note

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

Newton, M.A., and Raftery, A.E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society B, 56, 1, 3-48

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

Rosenblatt, M.J., and Lee, H.L. (1986). Economic Production Cycles with Imperfect Production Processes. IIE Transactions, 18, 1, 48-55

Rossi, P.E., and Allenby, G.M. (2003). Bayesian Statistics and Marketing. Marketing Science, 22, 3, 304-328

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 Cohrots Using Limited Information. Journal of Marketing, 72, 1, 82-94

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

Wagstaff, K., Cardie, C., Rogers, S., and Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. Proc. Eighteenth International Conference on Machine Learning, 577-584

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

Tables and Figures

Table 1: Examples of multi-cohort data sets to which the cross-cohort changepoint model can be applied

Cohort-level data	Cohort-level model	Hierarchical prior for cohort-level parameters	References
Number of doctor visits in the last three months for an age-demographic cohort of patients	Negative Binomial Distribution	Log normal	Winkelmann (2004)
Interpurchase time for a financial firm's customer cohort	Generalized Gamma	Log normal	Allenby, Leone and Jen (1999)
Purchase of CDs for a cohort of customers over time	Beta-Geometric/NBD	Logit normal & Log normal	Fader, Hardie and Lee (2005b)
Proportion of defective products per batch in manufacturing process	Bernoulli	Beta	Rosenblatt and Lee (1986)
Milk yield in dairy herd: one of N categories of yield	Multinomial	Dirichlet	Madouasse (2009)

Note:

The papers referenced utilize the cohort-level model for these applications, but do not use a cohort sequence approach to analyze cross-cohort differences. Our suggestion is that the cross-cohort changepoint model can be applied to such settings.

Table 2: Correlations between cohort parameters in a scalar changepoint-model using 2001 as cut-off year

Correlation coefficient/Quantile	2.50%	25%	50%	75%	97.50%
p1_12 (block 1, par 1 and par 2)	-0.58	-0.21	-0.01	0.18	0.52
p1_13 (block 1, par 1 and par 3)	-0.49	-0.13	0.08	0.27	0.59
p1_14 (block 1, par 1 and par 4)	-0.41	0.01	0.21	0.40	0.67
p1_23 (block 1, par 2 and par 3)	-0.54	-0.16	0.04	0.23	0.56
p1_24 (block 1, par 2 and par 4)	-0.62	-0.31	-0.11	0.09	0.50
p1_34 (block 1, par 3 and par 4)	-0.47	-0.14	0.10	0.31	0.64
p2_12 (block 2, par 1 and par 2)	-0.74	-0.28	-0.01	0.29	0.72
p2_13 (block 2, par 1 and par 3)	-0.73	-0.30	0.02	0.31	0.77
p2_14 (block 2, par 1 and par 4)	-0.71	-0.26	0.05	0.34	0.78
p2_23 (block 2, par 2 and par 3)	-0.76	-0.30	-0.01	0.27	0.77
p2_24 (block 2, par 2 and par 4)	-0.73	-0.32	-0.03	0.26	0.73
p2_34 (block 2, par 3 and par 4)	-0.72	-0.24	0.09	0.41	0.79

Note:

We observe that the distributions of Pearson correlation coefficients (obtained from draws of block covariance matrices) are centered near zero, but exhibit a fair amount of uncertainty. The 95% posterior interval covers a wide range of values.

Table 3: Posterior predictive statistics for 2000 Cohort, when using data cut-offs at 2000, 2001, 2002 and 2003

Cut-off Year		Actual Holdout Year 1 Value	Mean (T(y_rep))	MSE	MSE - Bias %	MSE - Var %	RMSE	Pr(T(y_rep) > T(y))
2000	Changepoint Model	2,625	3,197	1,841,345	17.7%	82.2%	1,357	0.675
	Static Model	2,625	3,563	1,397,720	63.0%	37.0%	1,182	0.910
2001	Changepoint Model	2,201	2,218	1,400,150	0.0%	100.0%	1,183	0.463
	Static Model	2,201	2,808	851,641	43.2%	56.8%	923	0.821
2002	Changepoint Model	1,822	1,728	1,035,545	0.9%	99.1%	1,018	0.395
	Static Model	1,822	2,247	636,159	28.3%	71.7%	798	0.735
2003	Changepoint Model	1,551	1,590	5,933	25.6%	74.4%	77	0.728
	Static Model	1,551	1,594	5,950	30.4%	69.6%	77	0.747

Note:

Mean(T(y_rep)) refers to the mean of the posterior predictive distribution for the statistic, which in this case is holdout transactions.

MSE is the mean squared error for the posterior predictive draws.

MSE – Bias % is the proportion of MSE due to bias.

MSE – Var % is the proportion of MSE due to variance in the predictive draws.

RMSE is the root mean squared error, presented to provide a measure of error in the same units as the number of transactions.

Pr(T(y_rep)>T(y)) is the posterior predictive p-value. A p-value closer to 0.5 would indicate that the actual value is located close to the center of the predictive distribution. A p-value at the extremes (close to 0 or 1) would indicate a poorer model fit with the data.

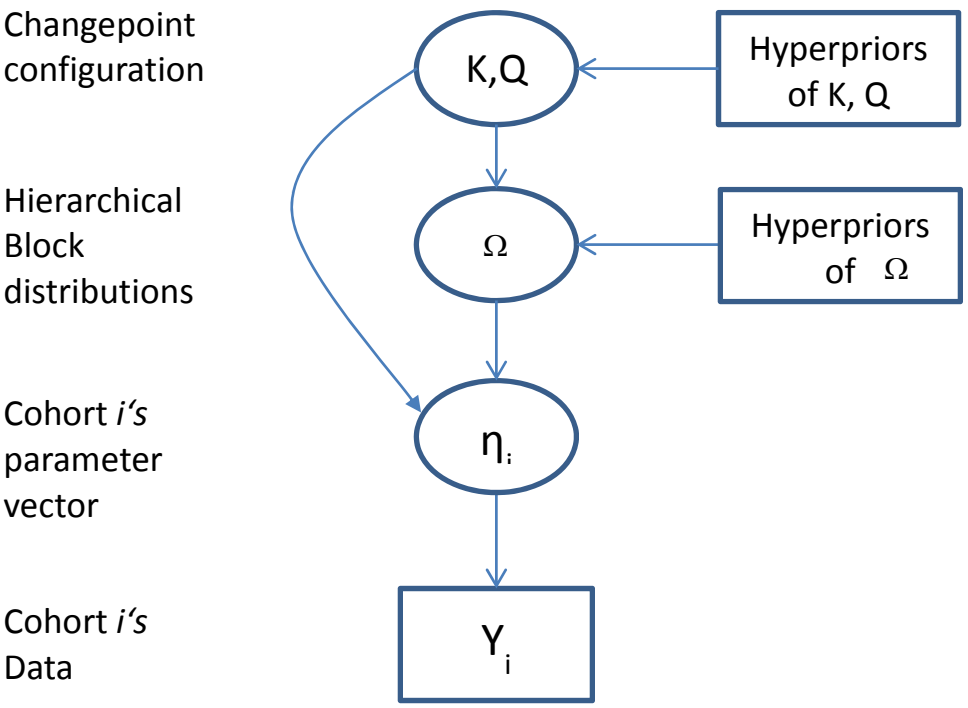
Table 4: Log Marginal Likelihood for Changepoint and Static models under each cut-off year

Cut-off Year / Model	Changepoint Model	Static Model
2000	-366,528	-366,529
2001	-410,579	-410,579
2002	-464,984	-464,984
2003	-517,352	-517,352

Note:

The log marginal likelihoods were calculated using the harmonic mean method (Newton and Raftery 1994) under the Changepoint and Static models. We observe both models to have very similar log marginal likelihoods, because cohorts included in model estimation have at least three observations and have very informative likelihood functions due to the large number of donors in a cohort. Thus, little to no shrinkage towards the prior is observed for cohorts included in model estimation for either model, and cohort level parameters are tightly distributed.

Figure 1: Directed acyclic graph of hierarchical model



Note:

K and Q are the number and locations of changepoints respectively. In this paper, K is set to 1 and only Q is random.

Ω contains the hierarchical block distributions. For K = 1, Ω contains exactly two distributions.

Figure 2: MCMC Pseudo-algorithm for scalar-changepoint model

Generate random starting values for $\eta_i \forall i$

Generate random starting value for changepoint location q

Loop through iterations from 1:nIterations

  Draw new values for $\mu_{g_1}, \Sigma_{g_1}, \mu_{g_2}, \Sigma_{g_2}$ based on q and $\eta_i \forall i$, by Gibbs sampling from conditional posterior distributions.

  Loop through each cohort

    Draw new η_i using Metropolis-Hastings step

    Accept or reject candidate using posterior density ratio test

  End Loop

  Draw new changepoint location q based on $\mu_{g_1}, \Sigma_{g_1}, \mu_{g_2}, \Sigma_{g_2}$ and $\eta_i \forall i$ using Gibbs sampling from multinomial density function

End Loop

Figure 3a: Large jump from $\mu_{p_i} = 0.6$ to $\mu_{p_i} = 0.3$ and minor fluctuations within-block with changepoint at cohort 2 (left), cohort 6 (middle) and cohort 10 (right) of a 11 cohort simulated sequence

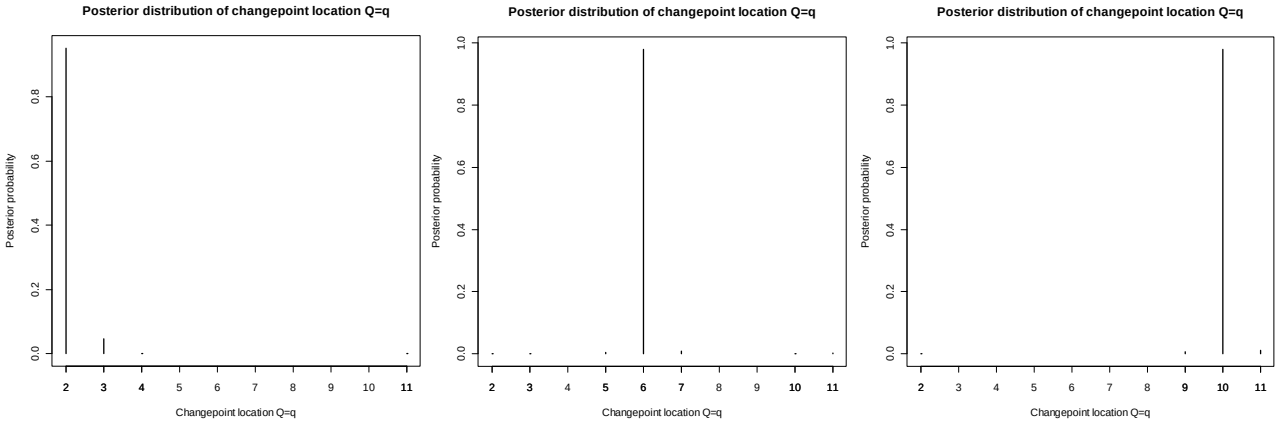
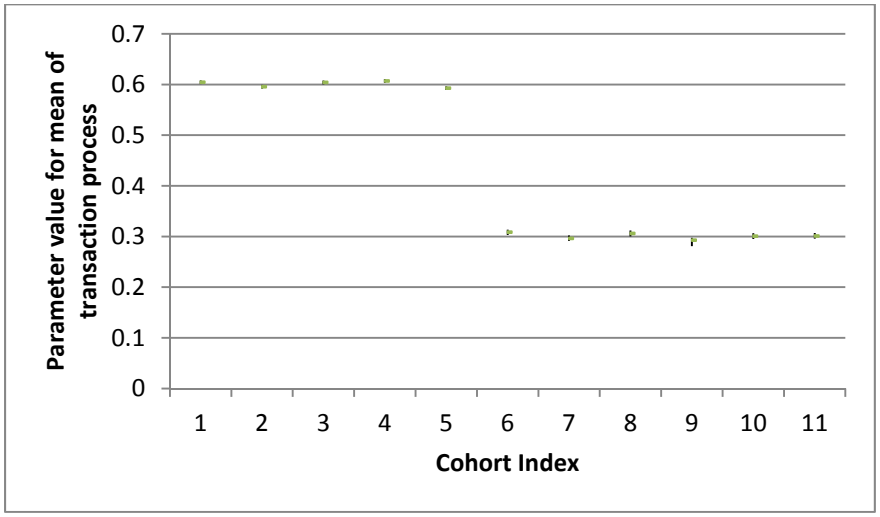
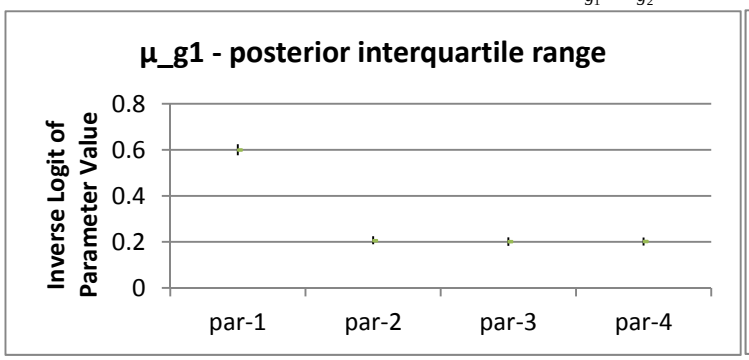


Figure 3b: Posterior interquartile range for each cohort's μ_{p_i} when discrete jump coded at cohort 6



Note that interquartile range is fairly tight for each cohort, and may not be easy to view from this graph. It suffices to observe the discrete shift present at cohort 6, which is picked up in the middle graphic of Figure 3a.

Figure 3c: Posterior interquartile range of μ_{g_1}, μ_{g_2}



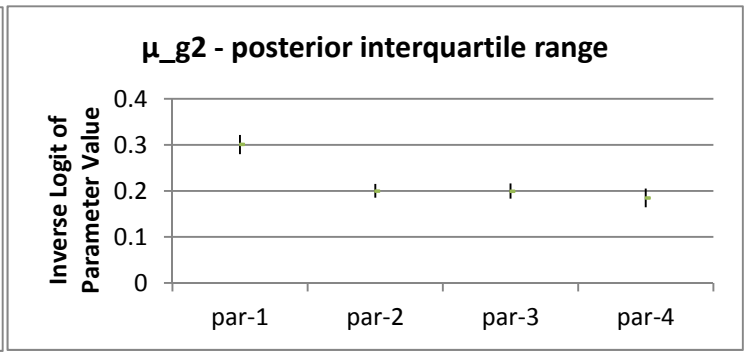


Figure 4a: No jumps, only minor fluctuations across 11 cohort simulated sequence

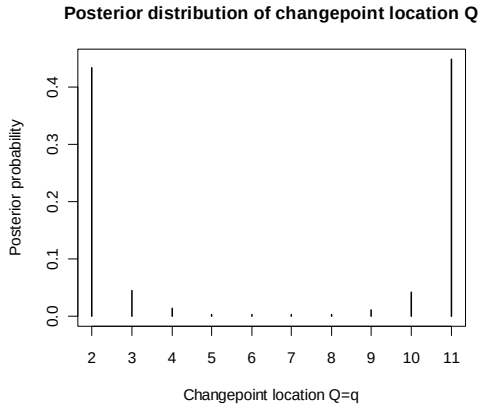


Figure 4b: Posterior interquartile range for each cohort's μ_{p_i}

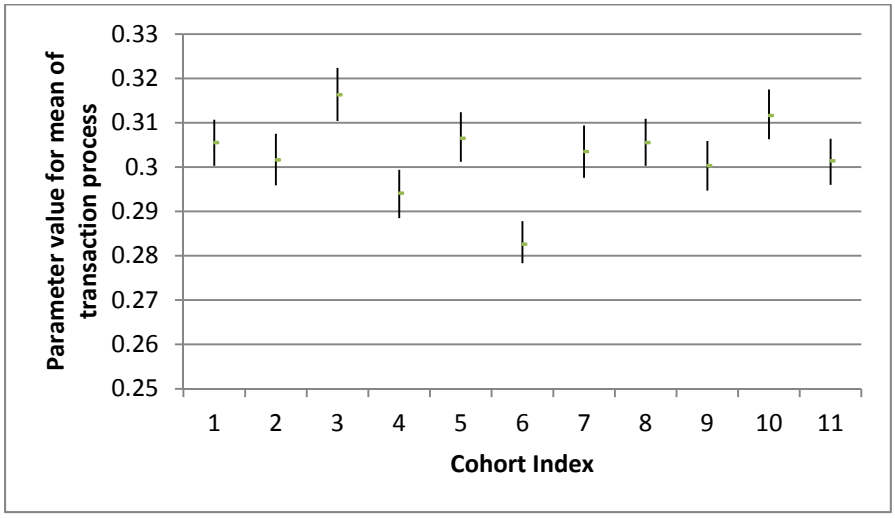
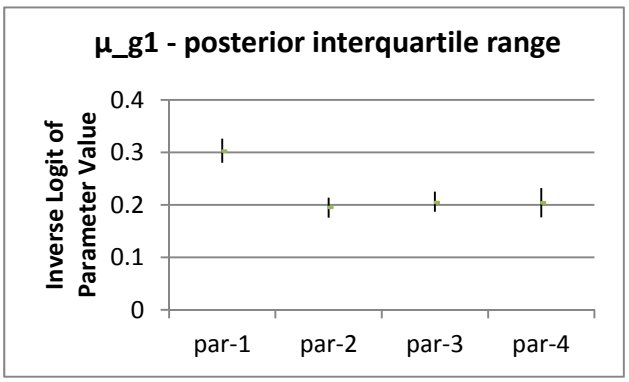


Figure 4c: Posterior interquartile range of μ_{g_1}, μ_{g_2}



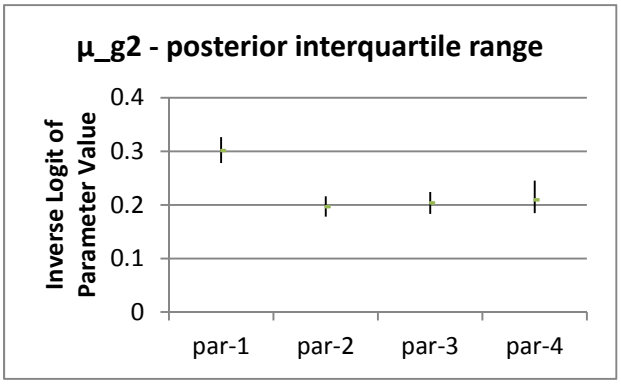


Figure 5a: Scalar-Changepoint: Large jump from $\mu_{p_i} = 0.3$ to $\mu_{p_i} = 0.6$ at cohort 3 and large jump from $\mu_{\theta_i} = 0.2$ to $\mu_{\theta_i} = 0.5$ at cohort 6; other parameters similar across the cohort sequence

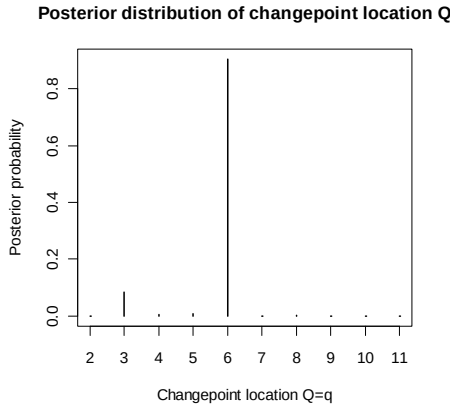


Figure 5b: Posterior interquartile range for each cohort's μ_{p_i} and μ_{θ_i}

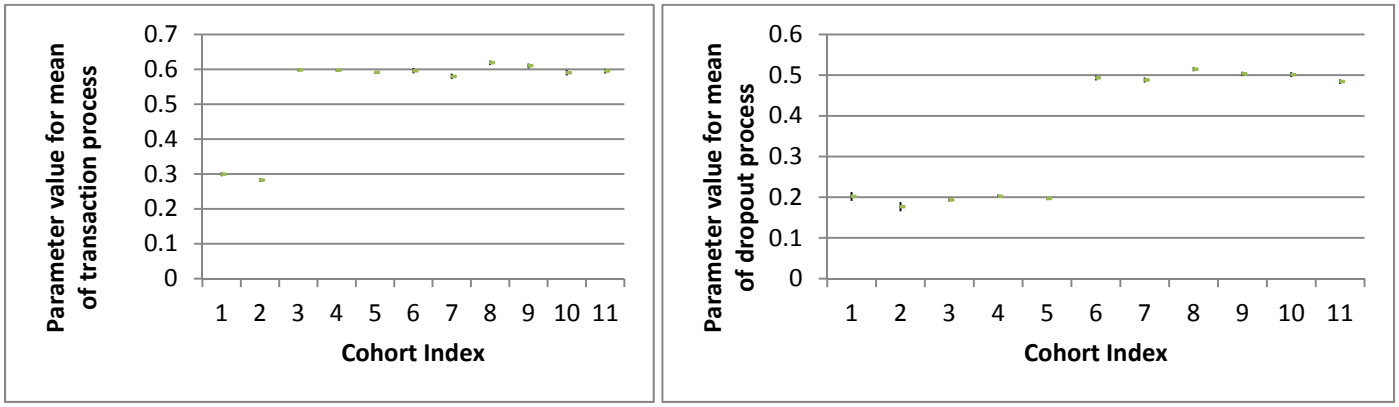


Figure 5c: Posterior interquartile range of μ_{g_1}, μ_{g_2}

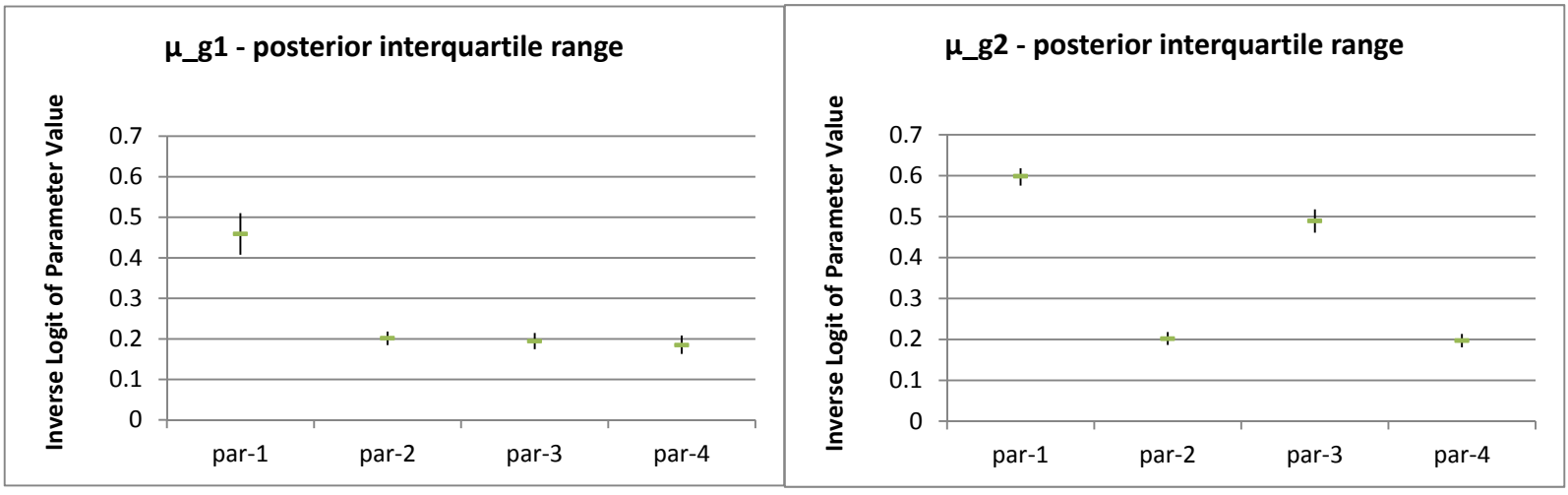


Figure 6a: Vector-Changepoint: Large jump from $\mu_{p_i} = 0.3$ to $\mu_{p_i} = 0.6$ at cohort 3 and large jump from $\mu_{\theta_i} = 0.2$ to $\mu_{\theta_i} = 0.5$ at cohort 6; other parameters similar across the cohort sequence

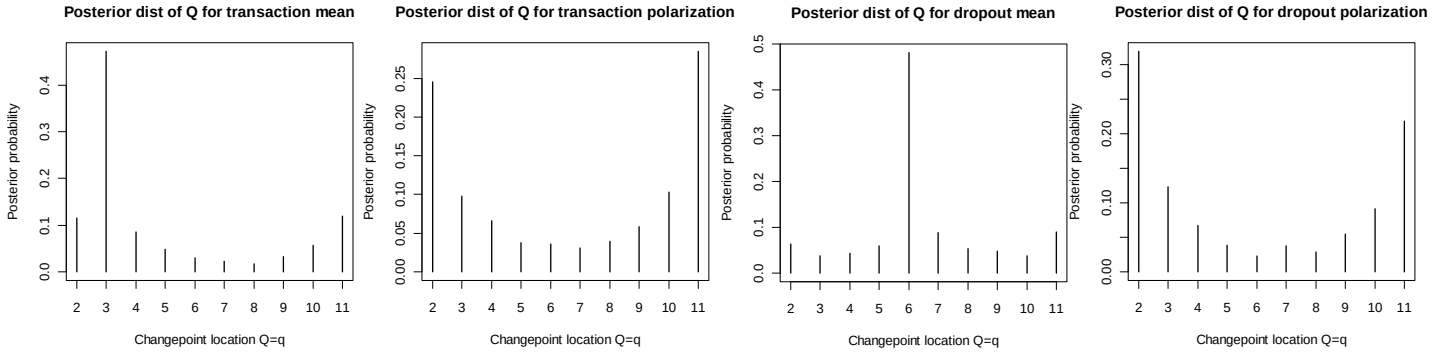


Figure 6b: Posterior interquartile range for each cohort's μ_{p_i} and μ_{θ_i}

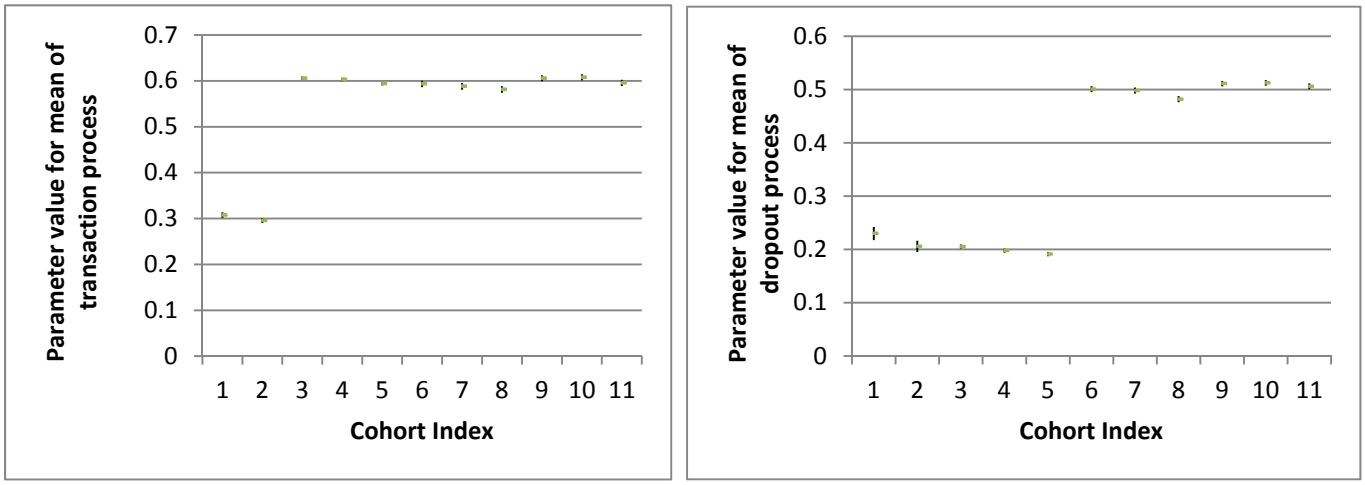


Figure 6c: Posterior interquartile range of μ_{g_1}, μ_{g_2} (each parameter driven by own changepoint configuration)

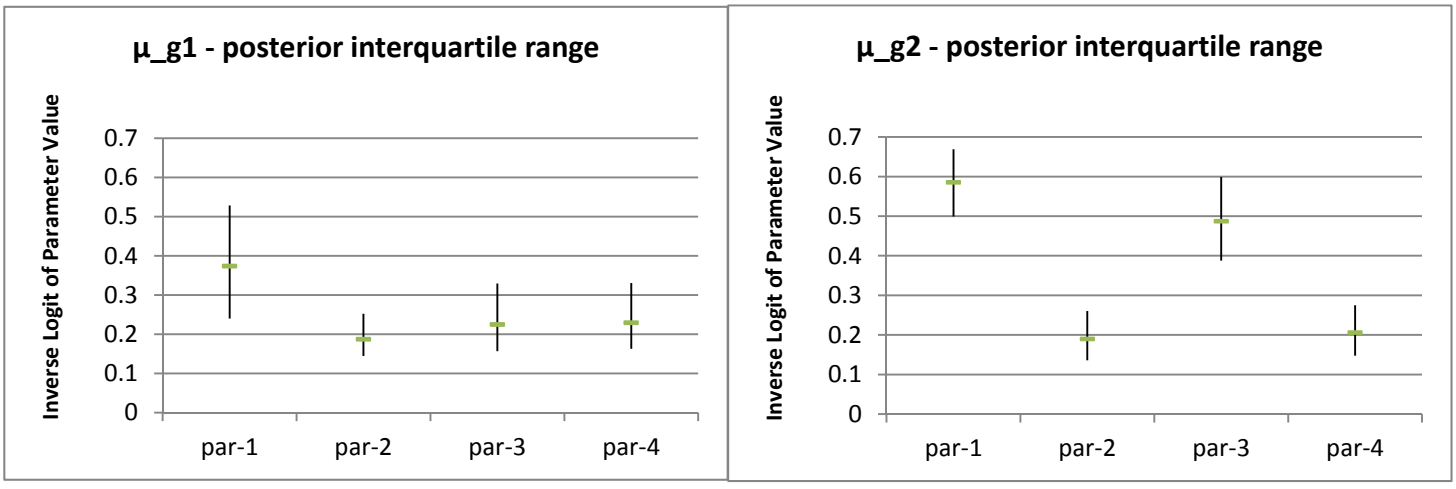
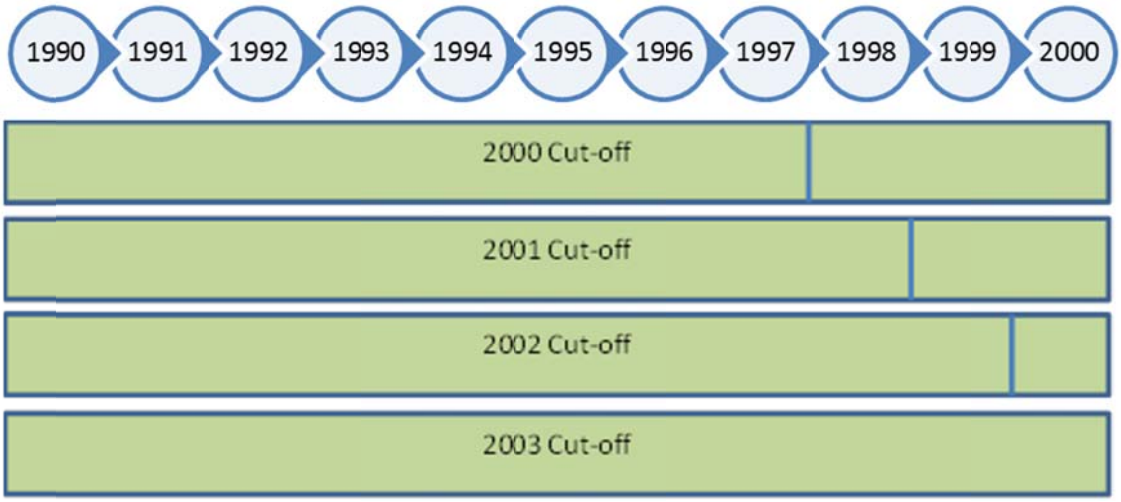


Figure 7: Cut-off years and cohorts included for model estimation



Note:

There are three aspects of cutting off data at the various "cut-off" years that are noteworthy. First, only cohorts with at least three repeat observations are included for model estimation. For instance, at the 2000 cut-off, the latest cohort included in model estimation is the 1997 cohort, with repeat observations in 1998, 1999 and 2000. By the 2003 cut-off, every cohort in our data set is included in model estimation as even the 2000 cohort has 3 repeat observations.

Second, each cohort (whether included or excluded for model estimation) has a different number of observations. The 1990 cohort always has the largest number of observations. For instance, at the 2001 cut-off, the 1990 cohort has 11 observations and the 2000 cohort has 1 observation.

Third, any cohort that is excluded from model estimation can still be predicted using MCMC draws from the various model parameters to generate posterior predictive distributions of interest.

Figure 8: Posterior probability of changepoint using 2000 cut-off (top left), 2001 cut-off (top right), 2002 cut-off (bottom left), 2003 cut-off (bottom right).

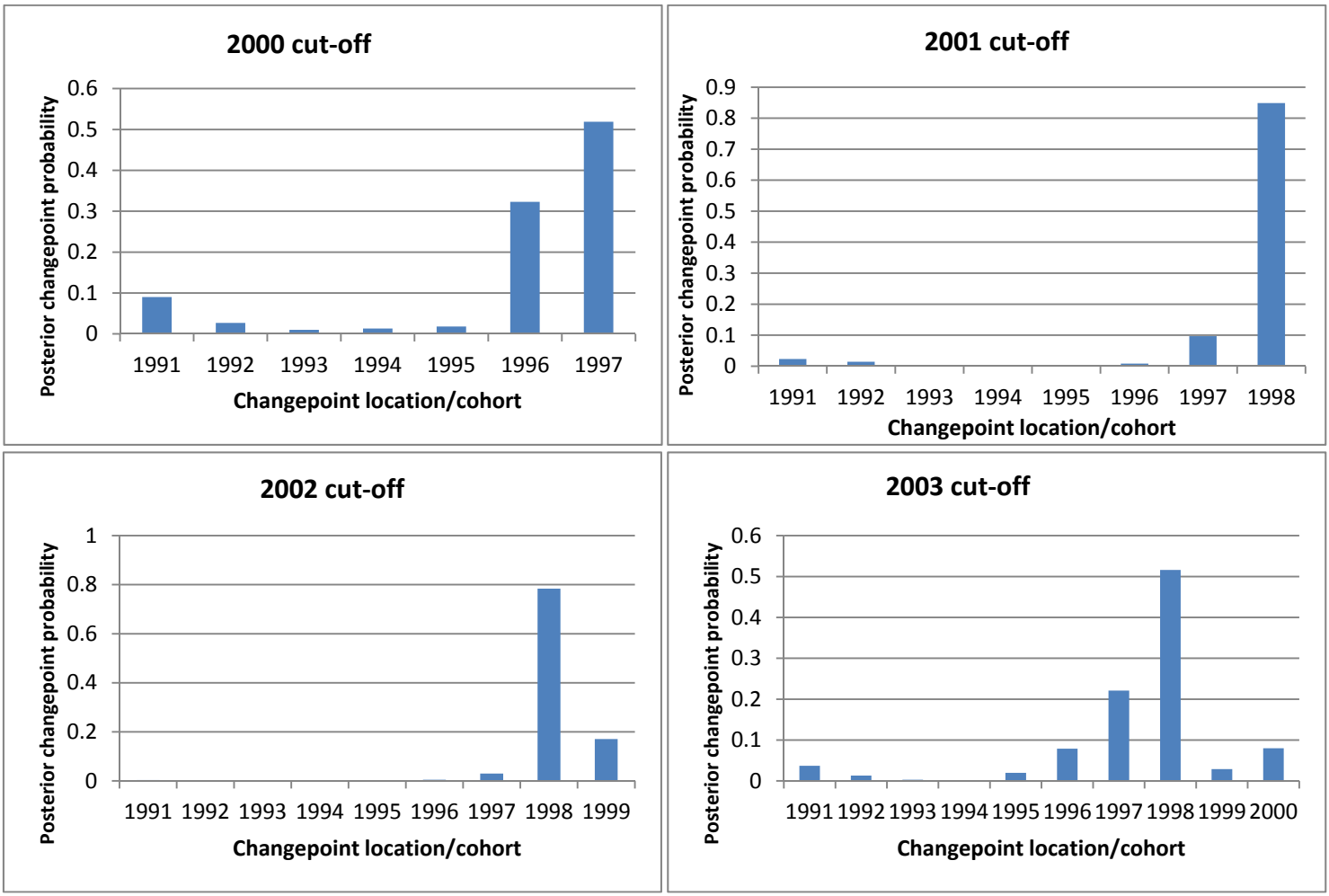


Figure 9: Posterior interquartile range for parameters across cohort sequence (2002 cut-off)

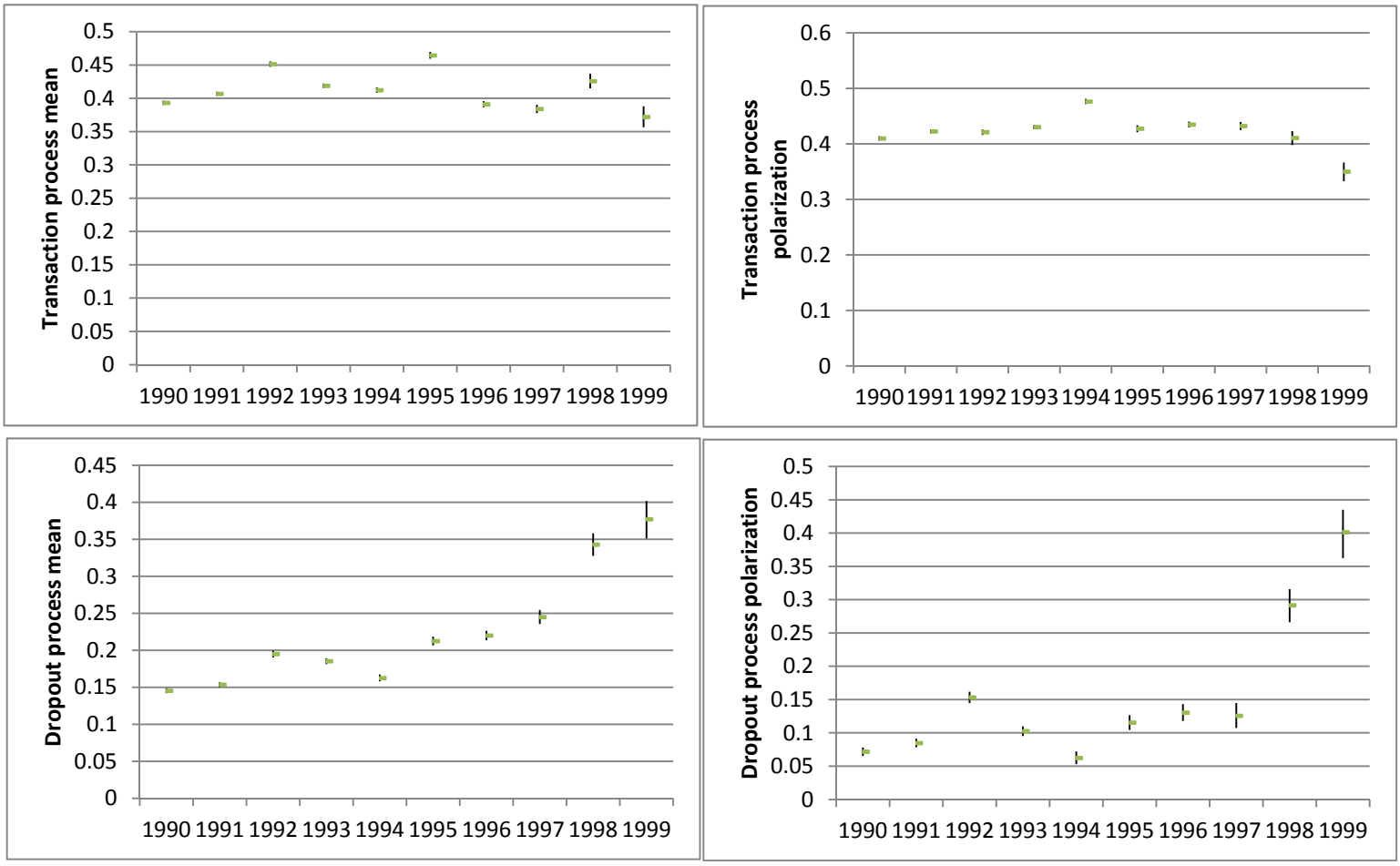
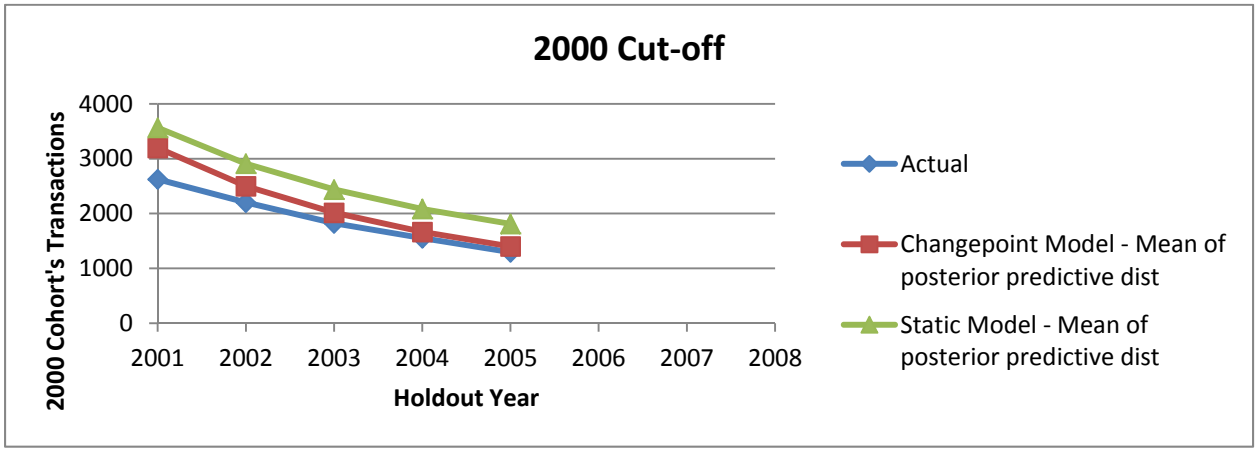
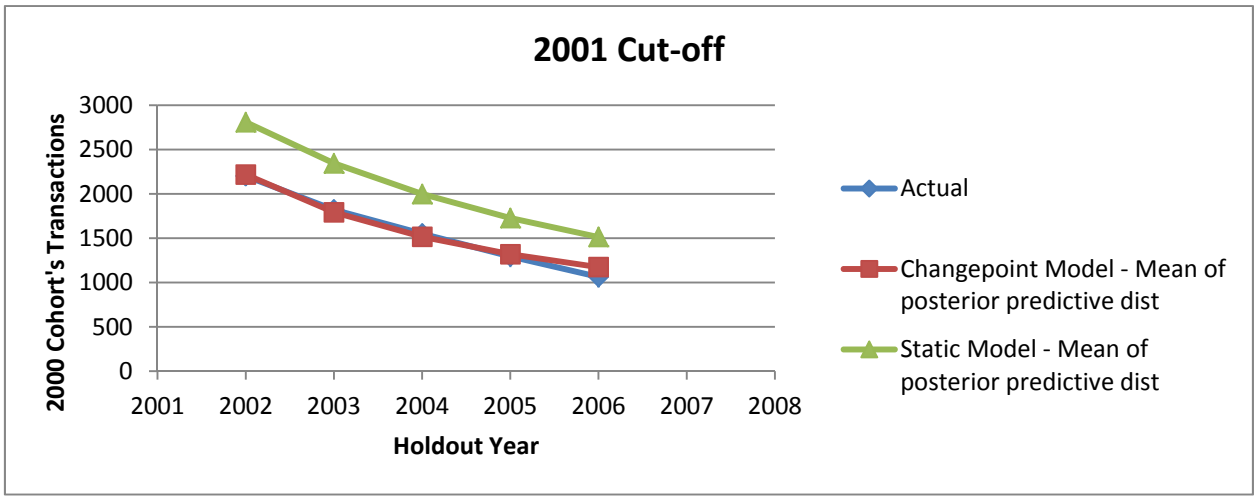


Figure 10: Actual holdout transactions versus posterior predictive distribution means for 2000 cohort at cut-off years ranging from 2000 to 2003





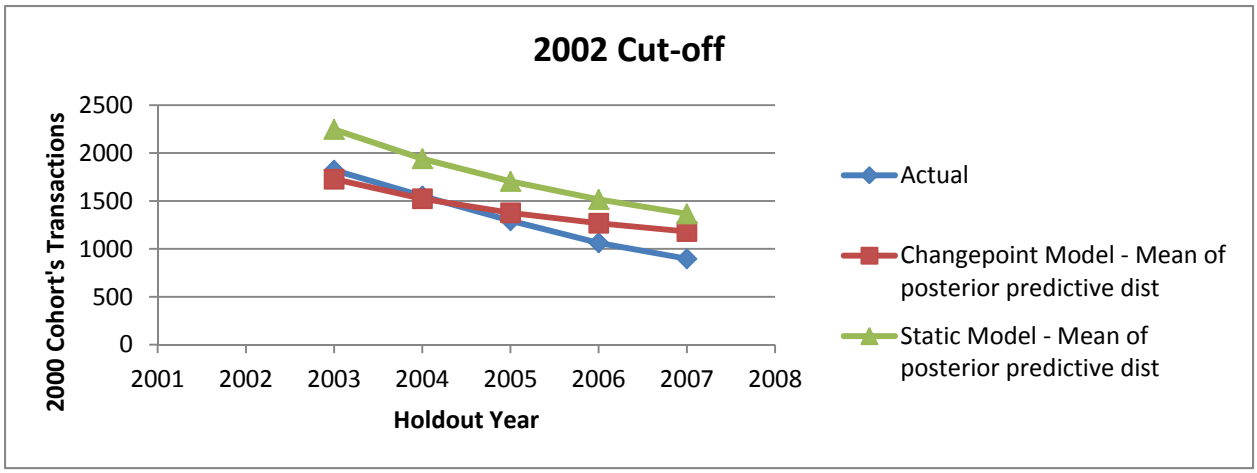
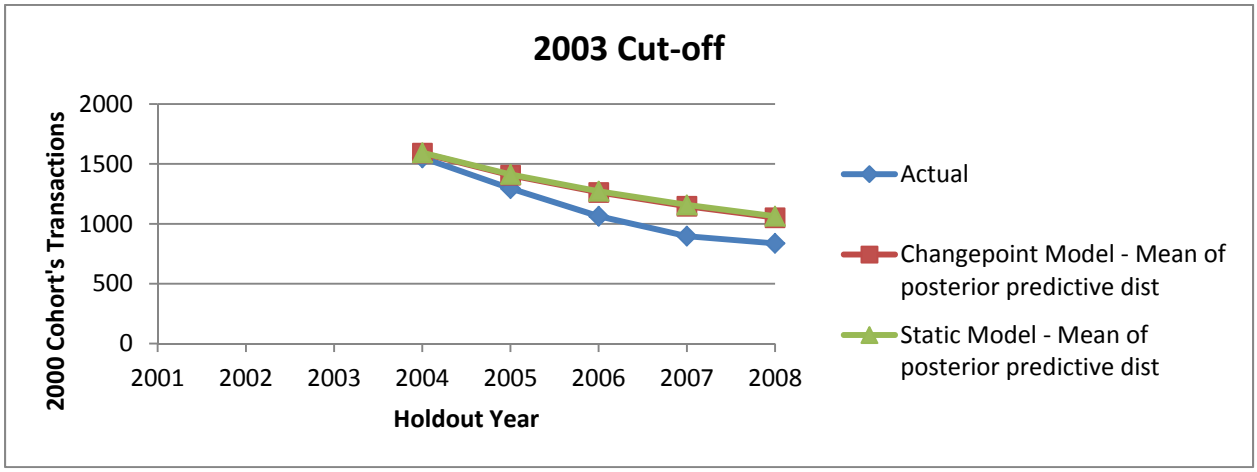


Figure 10 (continued):



Appendix

A. MCMC Implementation Details

To implement the MCMC algorithm, we need full conditionals for Q, Ω and η.

From equation (1), we have

$$p(Q = q \mid \Omega, \eta) = \frac{p(Q = q) \cdot p(\eta \mid Q = q, \Omega)}{\sum_{j=2}^{n_c} p(Q = j) \cdot p(\eta \mid Q = j, \Omega)}$$

Q can be drawn as a Gibbs sampling step from a multinomial distribution by calculating equation (1) for every value of Q.

Drawing the parameters of Ω conditional on *q* and η is straightforward if we use a conjugate prior. We choose the Normal-Inverse-Wishart conjugate prior for the multivariate normal $p(\eta \mid Q = q, \Omega)$ so that it becomes a Gibbs sampling step. We draw block parameters ω_1 and ω_2 separately as they are conditionally independent given Q and η.

$$p(\Omega \mid Q = q, \eta) \propto p(\eta \mid Q = q, \Omega) \cdot p(\Omega)$$

The hyperpriors for each block's Normal-Inverse-Wishart distribution are set to be non-informative:

$\kappa_0 = 0.01$; $\mu_0 = [0 \quad 0 \quad 0 \quad 0]^T$; $\nu_0 = 7$; $\Sigma_0 = I_4$

κ_0 is a measure of the amount of "prior data" upon which μ_0 is based. By setting it to a small positive constant, it scales variance Σ such that the distribution of $\mu_g \mid \Sigma$ is diffuse. $\nu_0 = 7$ is set so that the degrees of freedom equals $rank(\Sigma_0) + 3$. $\Sigma_0 = I_4$ is the identity matrix. Both these settings result in a reasonably diffuse prior (Rossi and Allenby 2003). While the prior Σ_0 does not allow for correlations in the parameter space of η, the posterior distribution of Σ_{g_b} can be influenced by data-driven correlations in the cohort-level parameter vectors.

The parameters of the Normal-Inverse-Wishart model for each block *b* can be updated using η and Q before drawing posterior block parameters.

$$\Sigma_{g_b} \sim IW(\nu_{p_b}, \Sigma_{p_b})$$
$$\mu_{g_b} | \Sigma \sim MVN(\mu_{p_b}, \Sigma / \kappa_{p_b})$$

Let $n_b = \sum_{i=1}^{n_c} I_{i \in b}$ be the number of cohorts assigned to a block and $\overline{m}_b = \frac{1}{n_b} \sum_{i=1}^{n_c} I_{i \in b} \cdot \eta_i$ be the mean of the cohort vectors assigned to block *b*.

The Normal-Inverse-Wishart parameter updates are given by the following equations:

$$\kappa_{p_b} = \kappa_0 + n_b$$
$$\mu_{p_b} = (\kappa_0 \cdot \mu_0 + n_b \cdot \overline{m}_b) / \kappa_{p_b}$$
$$\nu_{p_b} = \nu_0 + n_b$$
$$\Sigma_{p_b} = \Sigma_0 + \sum_{i=1}^{n_c} I_{i \in b} \cdot (\eta_i - m_b) \cdot (\eta_i - m_b)^T + \frac{\kappa_0 \cdot n_b}{\kappa_{p_b}} \cdot (m_b - \mu_0) \cdot (m_b - \mu_0)^T$$

The final step is to draw a parameter vector for each cohort, and this requires a Metropolis-Hastings step since the multivariate normal prior is not conjugate to the BG/BB likelihood function in Equation (6). The unnormalized posterior density used to calculate the conditional probability of the proposed candidate is $p(\eta_i^{cand} | Y_i, Q = q, \Omega) \propto \prod_{j-1}^{J} \left(p(Y_i | \eta_i^{cand}, x_j, t_{x_j}, n_i) \right)^{f_j} \cdot p(\eta_i^{cand} | Q = q, \Omega)$.

Note that our current implementation requires the model to fit one changepoint to the data. In the case that there are multi-modalities (multiple changepoints) in the data, this may hinder the accurate detection of the most sizeable discrete jump.

We used the following configuration to run the MCMC algorithm on the actual data set. The first 10,000 iterations are considered the burn-in period. Most parameters converged within 2,000 iterations but this buffer helps ensure that all model parameters converge. It is difficult to examine the convergence of the parameters in each model with graphical checks. We use the Gelman-Rubin statistic for each estimand (using draws from three different chains with overdispersed starting parameter values), and determine that every parameter of each model converges (using $\hat{R}$ threshold of 1.1 per Gelman et al 2004).

Autocorrelation is significant for cohort level parameters (especially ϕ_{θ_i} which represents the polarization for the dropout Beta distribution) and requires considerable thinning to result in

approximately i.i.d draws. We ran 550,000 iterations, and thinned aggressively after burn-in, by taking every 500th draw. These draws exhibited very low autocorrelation and can be considered a reasonable approximation to i.i.d samples for the purposes of estimating posterior intervals.

We used a constant jump variance or step of 0.0025 (for each parameter in η_i) for the Metropolis-Hastings step in drawing each cohort's parameter vector. This was tuned to provide a reasonable acceptance rate (around 10% to 30% depending on the cohort), while keeping autocorrelation from being too high. A lower jump variance boosts the acceptance rate but at the expense of more significant autocorrelation. A higher jump variance reduces the acceptance rate. We did not implement an adaptive jump variance as we found good convergence and mixing properties with the fixed jump variance.

B. Random changepoint model discussion

The terms $p(\eta \mid \Omega, K = k, Q = \{q_1, ...q_k\})$ and $p(\Omega \mid K = k, Q = \{q_1, ...q_k\})$ pose the challenge in a model with a random number of changepoints. The number of blocks in the cohort sequence is (K + 1) and therefore dependent on K. The number of hierarchical priors is equivalent to the number of blocks, and the MCMC algorithm will have to handle potentially changing dimensionality of Ω to even draw a new changepoint configuration, and to enable convergence and good mixing. The reversible jump MCMC method (Green 1995) would entail a modified Metropolis-Hastings step which can draw K, and evaluate whether to accept or reject this draw. A process of creating new hierarchical priors or removing a prior will need to be defined per Green (1995). Carlin and Chib (1995) can also be applied to this context by indexing a set of models based on the number of changepoints from 0 to $k_{\max}$, and maintaining a set of $\Theta = \{\Omega_0, ...\Omega_{k_{\max}}\}$ by taking draws for all values of K, even those that are not drawn in that MCMC iteration.

C. Recency-Frequency Sufficient Statistic - Example

For each cohort, we construct a recency-frequency table that serves as sufficient statistics for the BG/BB model. As an illustration, we provide the recency-frequency table for the 1996 cohort after four periods of observation (from 1997 – 2000) in Table C1. Of the 18,527 donors in this cohort, over half have no transaction activity since their initial donation, after four observations (1997 – 2000). The BG/BB likelihood function is described in equation (6) as a product of likelihoods corresponding to each row of the recency-frequency table.

Frequency (x)	Recency (t_x)	# donors
4	4	1,343
3	4	826
2	4	494
1	4	239
3	3	661
2	3	640
1	3	553
2	2	786
1	2	707
1	1	1,994
0	0	10,284

Table C1: Recency-Frequency Table for 1996 Cohort after four observations (data as of year 2000)