Submitted to the Annals of Applied Statistics

A SCALABLE PATH ALGORITHM FOR CLUSTERWISE MESSAGING COMPONENT ANALYSIS IN DIGITAL COMMUNICATIONS

BY RASHMI RANJAN BHUYAN[1,a], WREETABRATA KAR[2,c] AND GOURAB MUKHERJEE[1,b]

[1]*Marshall School of Business, University of Southern California,* [a]*bhuyanr@usc.edu;* [b]*gourab@usc.edu*

[2]*Mitch Daniels School of Business, Purdue University,* [c]*wkar@purdue.edu*

We develop a novel statistical procedure for analyzing the effects of various messaging components in digital communications. A key feature of the proposed method is that it segments consumers based on their historical characteristics and provides fine-grained estimates of the heterogeneous effects that messaging components have on the different user segments. The proposed method CURM, provides *clusterwise* analysis of *users' responses* to digital *messages*. While CURM uses historical characteristics to segment users via generalized clusterwise linear regression, and through dynamic indices, it incorporates (a) short-term changes in user engagement, and (b) the impact of recent messages on users' responses.

We conduct Bayesian estimation of the model parameters by using a novel Gibbs algorithm which is highly scalable due to the usage of Polya-Gamma distributions based data-augmentation strategy in handling Binomial likelihoods of users' responses to messages. Finally, through a path-algorithm (PA), we provide an integrated framework for providing fine-grained analysis of the messaging component effects at various levels of heterogeneity. The entire methodology is implemented by the associated R package `PACURM`. We establish large-sample properties on the operational characteristics of the developed algorithm and also describe its computational complexity.

We apply PACURM on consumer responses to digital coupons from the apparel industry. We demonstrate the importance of user segmentation, the frequency of communication messages and the utility of messaging components for effective digital couponing.

1. Introduction. Digital communication has become a fundamental aspect of modern interactions, increasingly favored for its convenience, immediacy, and extensive reach (Statista, 2023). This shift towards digital communication is fueled by its ability to transcend geographical boundaries, enabling seamless interaction across diverse demographics (Nelson, 2023). As individuals increasingly turn to online channels for information and interaction, organizations including businesses, non-profits, and political entities recognize the essential need to adopt digital communication strategies to remain relevant and competitive in their respective domains. Political candidates are increasingly leveraging social media, email campaigns, and targeted messaging to engage voters, rally support, and effectively communicate their platforms (Glazer and Krouse, 2020; Pandit, 2022). Similarly, non-profit organizations are progressively utilizing digital channels to raise awareness, garner donations, and galvanize backing for their causes (Tabas, 2021). Through data analytics and tailored messaging, these entities can personalize their outreach efforts, effectively resonating with their target audiences. Digital communication in retail is consistently becoming crucial for boosting sales, building consumer loyalty, and improving the shopping experience, offering a competitive advantage to companies that prioritize personalized interactions (Statista, 2022). In

Keywords and phrases: Digital communications, joint modeling, logistic regression, Polya-Gamma, generalized clusterwise linear regression, path algorithm.

the realm of e-health, digital communication facilitates seamless collaboration among healthcare professionals, enabling faster information exchange, consultation, and decision-making, thereby enhancing efficiency and quality of patient care (Appleby and Kumar, 2023).

The effectiveness of digital communication lies in its inherent capacity for tailoring and framing messages to suit individual characteristics and stages of behavior change. Petty et al. (1986) highlight that effective communication not only captures attention but also optimizes its impact on how people perceive issues. Recent research underscores the importance of aligning messages with the distinct stages of a user's behavior change: detection, decision, and implementation (Burkholder and Evers, 2002). Through targeted messaging and persuasive techniques, digital platforms tailor content to appeal to recipients' preferences, beliefs, and desires, thus influencing their motivations (Kunda, 1990). Moreover, digital communication allows for sophisticated message framing, presenting it in terms of intrinsic versus extrinsic gains or losses, offering a nuanced persuasion approach (Petty et al., 1986; Snyder and DeBono, 2014). For instance, in political campaigns, messages are tailored to appeal to different voter groups by framing policies in various ways, such as presenting healthcare reform as providing better access to medical care or reducing overall healthcare costs; or communicating tax policies as promoting social justice or fostering economic growth. The ease of such adaptability in digital communication ensures that messages resonate with diverse audiences, maximizing their persuasive impact.

1.1. *Messaging Component analysis: Background & Formulation.* Messaging component analysis harnesses users' responses to previous communications to study their reactions to various attributes of interest contained in the messages. The influence of various messages on user psychology has been extensively studied in the existing literature (Petty et al., 1986; Detweiler et al., 1999; Mann, Sherman and Updegraff, 2004; Rimer and Glassman, 1999). However, these psychological methods have primarily focused on studying univariate aspects of the messages, such as their themes. Due to the explosive growth in digital messaging, most contemporary messages are blended assortments of features. Tokenizing the message into components and estimating their separate effect on users improves understanding and increases flexibility to tailor future communication to heterogeneous preferences of recipients. Further, due to the ease and reach of digital communication, huge observational datasets containing users' responses to digital messages have been archived, which naturally presents an opportunity to utilize messaging component analysis for improved messaging.

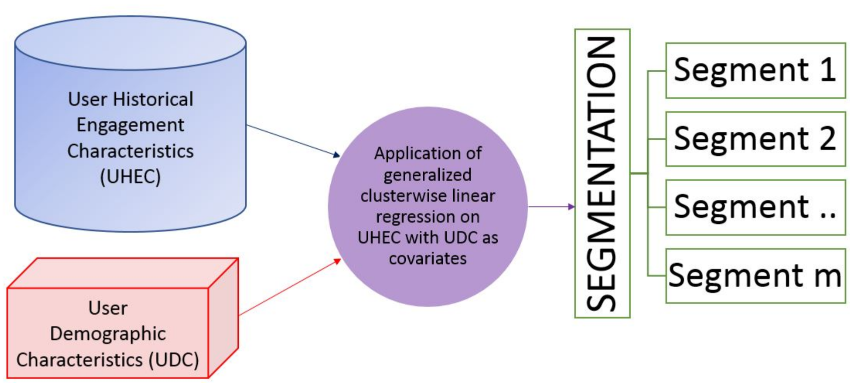


Fig 1: Segmenting users based on users' historical engagement characteristics (UHEC) and user' demographic characteristics (UDC) in Stage I of CURM.

In this paper, we develop a disciplined data analysis pipeline for understanding the effectiveness of digital messaging by studying users' responses in large observational datasets.

Statistical analysis is challenging here as we need to decouple the impacts of different constituents that influence users' decisions. There is usually wide heterogeneity in the user base, and thus, it is important to segment users based on similar attributes (Ch. 9 of Chabé-Ferret, 2022, Otter, Tüchler and Frühwirth-Schnatter, 2004, Malsiner-Walli, Pauger and Wagner, 2018) and then compare the effect of messaging within segments. This is Stage I of our pipeline, where we segment users by employing generalized clusterwise linear regression (GCLR) on users' historical engagement characteristics (UHEC) as well as users' demographic characteristics (UDC). Thereafter, in Stage II, we convert each digital message into messaging components and study how users from each segment respond to those low-dimensional messaging components. Our proposed *Clusterwise analysis of Users' Responses to communication Messages* (CURM) data analysis pipeline involves the development of novel scalable statistical algorithms for each of these two stages.

Figures 1 and 2 display the distinct challenges that are faced in CURM at Stages I and II respectively. In Stage II, we predict multiple nested responses of a user for an arbitrary message. The responses are nested in the messaging funnel starting from a user opening the message to culminating into positive action (conversion). while studying responses of the users' from each segments we need to consider (a) short-term changes in user engagement, and (b) the impact of recent messages (fatigue) in their responses to accurately estimate the effects of different messaging components used in the messages. Figure 2 shows how each digital communication message can be written as a weighted combination of features (components) from a prefixed dictionary. A user response to the message not only depends on their cluster and demographic characteristics, but also on their past experience and current engagement with the sender. Based on observing noisy user responses, we need to estimate the effect of these messaging components.

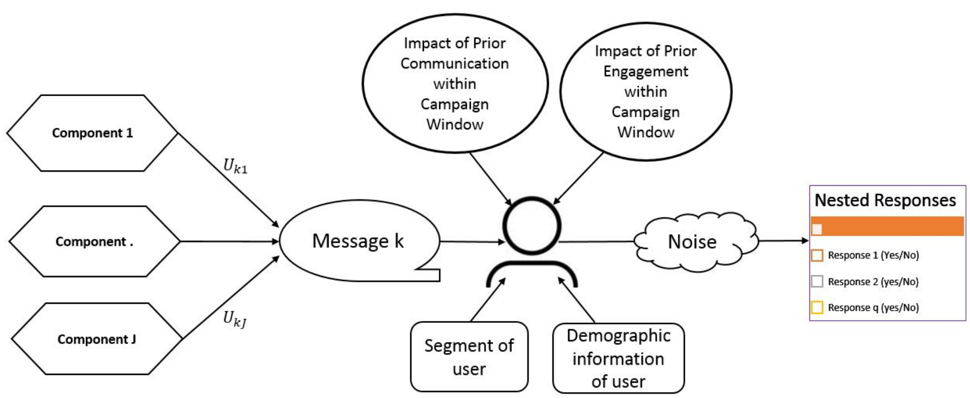


Fig 2: Schematic showing how a user's responses to a digital message are modeled in Stage II of CURM. A message comprises several components from a dictionary. Once a user gets the message, their responses are impacted by user and campaign specific characteristics.

In CURM, the two stages, (I) Segmentation of users, and (II) Analysis of users' responses, involve different statistical techniques. User responses are the primary data whereas UHEC and UDC for Stage I is often collected from third party surveyors. In order to understand the heterogeneous effect of digital communication among different users segments, we need to implement CURM as the number of segments varies. We provide a path-algorithm (Simon et al., 2011; Friedman, Hastie and Tibshirani, 2010; Park and Hastie, 2007) PACURM which provides a unified interface in R that can be used to study user responses and compare messaging effects at different heterogeneity levels. In this aspect, PACURM parallels recent developments of statistical methods in single-cell biology (Hickey et al., 2023; Chen et al., 2023; Anchang et al., 2016; Sen et al., 2014) where cells are first segmented and then subjected to expression analysis.

1.2. *Statistical Challenges and Our Contributions.* Developing a disciplined statistical framework for messaging component analysis in digital communications poses several statistical challenges and necessitates novel methodological developments. We describe the details below.

1. For segmentation in CURM, we conduct Generalized Cluster-Wise Linear Regression (GCLR) (Späth, 2014; Hennig, 2000; DeSarbo and Cron, 1988) which partitions users into a predefined number of clusters with each group's characteristics modeled by its own generalized regression model. For implementable policies, hard-clustering of the users is necessary (Frühwirth-Schnatter, 2006). GCLR has been successfully employed for hard-clustering in applications across various fields where the sample is from heterogeneous populations (Silva et al., 2019; Ng et al., 2006; Luo and Chou, 2006; Wedel and Kistemaker, 1989). In Section 4.1, we develop a modification of the popular iterative algorithm by Lloyd (1982) for K-means clustering to implement GCLR based on historical user characteristics, which may include discrete attributes. We provide theoretical support for our algorithm and demonstrate the importance of using overdispersion parameters in our application for efficient GCLR based segmentation of users.
2. In Stage II of CURM, we model multiple binary user responses to a digital message. In Section 4.2 (see Algorithm 2), using the Polya-Gamma data augmentation approach of Polson, Scott and Windle (2013), we develop a highly scalable (in sample size as well as in size of feature dictionary) algorithm for joint estimation of CURM parameters in Stage II. In CURM we use dynamic variables to account for a user's current engagement status and the impact of recent communications on their. Conducting scalable estimation in dynamic logistic models is very challenging (Park, Park and Schweidel, 2018). We develop a block Markov Chain Monte Carlo (MCMC) algorithm for efficient joint estimation in dynamic logistic models.
3. While it is extremely important to analyze the differences in the messaging-component effects as well as predictive performance of CURM at varying resolutions of heterogeneity in the user base, estimation of the parameters in Stage II of CURM is computationally expensive and it would be extremely resource-intensive to estimate each model separately for a different number of consumer segments. We develop a path algorithm PACURM that provides a path of estimates of the parameters in CURMas the number of clusters (m) varies over $1, \ldots, M$ for some large M. Akin to hierarchical clustering (Hartigan, 1985; Radchenko and Mukherjee, 2017; Tan and Witten, 2015; Chi and Lange, 2015), PACURM merges clusters in Stage I of CURM in a bottom up fashion based on minimizing the squared error loss. We provide fast, asymptotically efficient updates of the parameter estimates in Stage II of CURM as clusters are merged. In Section 5, we describe the proposed path algorithm along with detailed analysis of its computational complexity. We develop R package PACURM to implement the proposed algorithm.

We apply PACURM to digital communications data from the apparel industry where retailers use coupons to routinely communicate with their consumers. The main focus of our analysis is to understand consumers' responses regarding opening and then subsequently purchasing from a coupon. We study the impact of components used in the coupons after decoupling the effects of recent coupons and user specific historical and recent (in campaign) engagements. The key findings in our case-analysis reported in Section 6 are:

1. Price promotion components (Park, Park and Schweidel, 2018) used in coupons are, on average, more effective than non-priced promotions. These component effects vary across user segments (detailed in Section 6.2). The component dictionary used here following Ellickson, Kar and Reeder III (2022) well explains the variability in the coupon efficacy.
2. It is important to use overdispersion parameters in the GCLR model for segmenting users (detailed in Section 6.3). The carry over effects from previous coupons on a user's response to current coupon is highly significant.
3. Previous coupons have a significant negative effect and greatly decrease consumers' response propensity in terms of both opening and purchasing from coupons. It would be better if retailers reduced the frequency of the communications.

1.3. *Organization of the paper.* The rest of the paper is organized as follows. In Section 2, we describe our data. In Section 3, we present our two-stage dynamic mixture modeling framework and we detail the methodology used for estimating the proposed model in Section 4. In Section 5, we provide implementation details of the path algorithm. Using the dataset introduced in Section 2, we present the inference and implications of the results obtained by our proposed analysis in Section 6. The supplement contains the proofs of all the results stated in Sections 4 and 5, as well as further details on the model, its estimation procedure and results from the data analysis.

2. Data: Consumer responses to Apparel Coupons. In our application, we study digital coupons from an apparel retailer that actively employs frequent digital coupons over email to engage and incentivize its target users. Since their inception, coupons have always been among one of the most widely used communications for interventions (Lamb, Hair and McDaniel, 2011). Digital coupons provide several benefits over traditional coupons. They can be easily stored and searched by consumers, and they have much higher redemption rates than traditional coupons (Jung et al., 2010). They very low designing and dissemination costs. Through digital coupons retailers can not only assimilate different types of promotions in a single coupon but can also customize these coupons for different promotional needs. In our dataset, we analyze 25 digital coupons sent over a two-month period that we consider as the campaign period. Each user within the dataset received one or more of these digital coupons.

Variable	Mean	SD
Web Purchase (in $)	52.25	33.5
No. of Purchases	7.76	8.57
Days Since Opened	5.58	12.9
Days Since Purchased	82	91.59

Table 1: Summary Statistics of Historical Characteristics of Users

The retailer provided us with historical engagement characteristics for each user. Specifically, the data provided variables related to the average purchase amount (in US dollars) and the number of previous purchases before the campaign window, and the frequency of engagement with previous messages. Recency variables were constructed by noting the number of days since a user last opened and purchased, respectively, before the campaign window. Summary statistics of these variables are provided in Table 1. Additionally, the retailer provided user demographics, including age and income, which were categorical in nature.

The retailer granted us access to the subject lines of the digital coupons. Digital coupons offering any percentage discount were classified as price coupons, while the remaining coupons were designated as non-price coupons. Subsequently, following the method described by Ellickson, Kar and Reeder III (2022), we manually categorized each coupon into messaging components. Table 2 presents the number of digital coupons classified under price and non-price categories, along with their reach and overall effectiveness in eliciting user responses. Additionally, Table 3 provides more detailed information regarding reach and user response at the level of messaging components.

Promotion Type	No. of Emails	Open Rate	Purchase Rate	Reach
Price	18	0.410	0.012	1,016,835
Non-price	7	0.382	0.011	397,050

Table 2: Summary Statistics of Coupon Performance based on Coupon Type

Component Type	Components	No. of Emails	Open Rate	Purchase Rate	Reach
Price	30%Off Discount	4	0.400	0.011	148,757
	40%Off Discount	3	0.433	0.012	200,189
	50%Off Discount	4	0.398	0.013	253,682
	Product	13	0.414	0.011	782,583
	Clearance	2	0.388	0.013	81,399
Non-price	Mystery	6	0.390	0.014	349,188
	Free Gift	5	0.415	0.011	300,986
	Free Shipping	8	0.409	0.011	481,308
	Free Returns	3	0.376	0.012	201,369

Table 3: Summary Statistics of Coupon Performance based on Messaging Components

3. CURM model of users' responses to messages. Our model comprises broadly of two different statistical methods. Stage I clusters users based on a GLM (generalized linear model) framework. In Stage II we use the clusters generated in Stage I to estimate these cluster-specific effects. We provide a detailed analysis of the two stages below. In Section 5 later, we provide a computationally efficient path algorithm to estimate the effects as the number of consumer clusters varies in Stage I. To provide a concrete description of the model we explain CURM in the context of the data in Section 2. It can be easily adapted for other digital communication contexts.

3.1. *Stage I: Segmenting users based on historical characteristics.* To address the heterogeneity in the user base, we segment the consumers into G groups based on their historical characteristics and consider estimating the coupon effects for each segment. As G varies over 1 to n, we consider coupon effects over a wide range of resolutions, spanning from unsegmented consumer base to the highly personalized case with unique effects for each consumers.

We consider segmenting the consumers based on similarities in their recency, frequency and monetary (RFM) values (Ellickson, Kar and Reeder III, 2022) at the beginning of the communication campaign. For each consumer, we consider two recency variables $R^{(o)}$ and $R^{(p)}$, that respectively correspond to the days between the start of the promotional window and the last time they opened and purchased using any coupon from previous campaigns. For the frequency (F) and monetary (M) variables, we considered the number of instances of they purchased from the retailers in the last two years and the average spending in those transactions. We segment consumers based on these four RFM metrics.

Consider m clusters. Let $h : \{1,\dots,n\} \to \{1,\dots,m\}$ be the clustering function, i.e. $h(32) = 5$, implies that the 32^{nd} observation in the data belongs to the 5^{th} cluster. Let W and $\tilde{W}$ be two sets of covariates. We entertain global effects of W and cluster specific heterogeneous effects of $\tilde{W}$. We model the recency variables through linear models and the monetary variable using a log-normal distribution. For consumer i we denote the RFM attributes at the start of the campaigning window by $Z_i = (R_i^{(o)}, R_i^{(p)}, \log M_i, F_i)$. Then, for $l = 1, 2, 3$, we model,

$$Z_i(l) = w_i^T \delta^{(l)} + \tilde{w}_i^T \tau_{h(i)}^{(l)} + \sigma^{(l)} \epsilon_{il}, \text{ for } i = 1,\dots,n, \tag{1}$$

where, ϵ_{il} are independent standard normal distributions. We model the frequency variable through an overdispersed Poisson (Conway–Maxwell–Poisson) model:

$$\log P(Z_i(4) = z_i) = z_i \log \lambda_i - \left(\sigma^{(4)}\right)^2 \log(z_i!) - Z(\lambda, \left(\sigma^{(4)}\right)^2), \tag{2}$$

where, $\log \lambda_i = w_i^T \delta^{(4)} + \tilde{w}_i^T \tau_{h(i)}^{(4)}$ and $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \lambda^j/(j!)^\nu$. Using the overdispersion parameter is critical (as shown in Section 6.3) as the means and variances of frequency do not coincide. The variances are usually much larger than the means.

In (1)-(2), we use age, gender, income, and other demographic variables in W. These covariates are also used in Stage II models to analyze users' responses. The proposed GCLR framework clusters the RFM variables pertaining to historical user characteristics conditioned on the covarites. By conditioning on the covariates, we ensure that the segmentation procedure in Stage I does not impact the estimation of the covariate effects in the later stages. In (1)-(2), the clustering function h is invariant while the parameters change across the RFM variables. In Section 4.1, we estimate the clustering function h along with the parameters.

3.2. *Stage II: User Response Modeling.* We now model users' responses to digital messages. We use segments from Stage I along with the messaging component effects and previous coupon impacts to understand the heterogeneity in the responses. Let $\{C_1, C_2, \cdots, C_J\}$ be the dictionary of marketing components used in the promotional campaign. Let $\{\nu_1, \nu_2, \cdots, \nu_J\}$ be their corresponding effects. We explain the effect μ_k of coupon C_k, $k \in \{1,\dots,K\}$ by a linear combination of these components as: $\mu_k = \sum_{j=1}^{J} u_{kj}\, \nu_j + \gamma_k$, where, $u_{kl} \in \{0,1\}$ based on the presence of component l in coupon k and γ_k is the residual effect that can not be explained by the linear combination of the components. While u_{kl} are known, the parameters $\{\nu_j : 1 \le j \le J\}$ and $\{\gamma_k : 1 \le k \le K\}$ need to be estimated.

For consumer i at the time t, define $e_{ij} = 1$ if they received a digital communication at time t. In our case, no user received more than one message in a day. We consider the daily timeline $t = 1,\dots,T = 45$, for the communication campaign. At time t, we model impact of previous messages through a decay effect: $\eta_{i,t} = \sum_{j=1}^{t-1} e_{i,t-j}\, \rho_\eta^{j-1}$, where, the memory parameter $\rho_\eta \in [0,1)$. Decay models are widely used in quantifying the carry-over effect in communications (Leone, 1995; Venkatesan and Farris, 2012). For example, Park, Park and Schweidel (2018) uses *adstock* variables based on decay models to incorporate the prolonged effect of coupon advertising on consumer purchase behavior.

For modeling a user's current engagement status, we use a dynamic decay model. First note that we consider modeling user's multiple nested responses. In our application case, we model their responses to opening and purchasing from a digital coupon. These two responses are typically the initial and final steps of nested responses in digital communication. Consider the dynamic variable $\zeta_{i,t}$ as $\zeta_{i,t} = \sum_{j=1}^{t-1} e_{i,t-j}\, o_{i,t-j}\, \rho_\zeta^{j}$ where $o_{i,t} = 1$ if the user opened the coupon received at time t sometime before the end of the campaign window and 0 otherwise. The memory parameter $\rho_\zeta \in [0,1)$. This dynamic parameter thus considers the recent short-term engagement dynamics by giving exponentially decreasing weights to past engagements.

Next, consider the user's actions with regards to the purchase (or the final step) for previous communications: $\kappa_{i,t} = \max\{p_{i,t-j} : j = 1, \ldots, t-1\}$ where $p_{i,t} = 1$ if the user purchased from the coupon received at time t and 0 otherwise. By definition $\kappa_{i,t} \in \{0, 1\}$. For most communication campaigns, it is extremely rare for a user to respond affirmatively in the final stage of the conversion funnel more than once during the campaign. For our digital couponing case, we have a promotional window of only 45 days and having repeat purchases is very rare. Thus, if $\kappa_{i,t}$ is positive, we expect a significant decrease in purchase propensity for the subsequent coupons.

For $t = 1, \ldots, T$ and $i = 1, \ldots, n$ when $e_{it} = 1$, we model the log-odds for the ith consumer's positive response to the message received on day t as:

$$x_{i,k}^T \alpha + \tilde{x}_{i,k}^T \beta_h + \theta_1 \eta_{i,t} + \zeta_{i,t} + \theta_2 \kappa_{i,t} + \sum_{j=1}^{J} u_{kj} \nu_{j,h} + \gamma_{k,h} + a_h, \tag{3}$$

where, $h := h(i)$ is the segment that user i was classified in Stage I and $k(i,t)$ denotes the message that user i received on day t. For notational ease, we keep suffixes in $h(i)$ and $k(i,t)$ implicit. The parameter $\theta_1 \in \mathbb{R}$ determines whether the cumulative carry-over effect from previous messages is positive or negative. Note that, in (3), α captures user and message invariant covariate effects, β_h captures the effects varying across user segments but invariant across messages, $\eta_{i,t}$ capture the impact of previous messages, $\zeta_{i,t}$ captures the effect of recent user engagement, $\nu_{j,h}$ and $\gamma_{k,h}$ are the user segment specific messaging effects and a_h is a random intercept that varies across consumer segments.

We model $\text{logit}\big(\mathbb{P}(O_{it} = 1)\big)$–the log-odds for the ith consumer positive response to opening the coupon received on day t by (3) as

$$x_{ik}^T \alpha^{(1)} + \tilde{x}_{ik}^T \beta_h^{(1)} + \theta_1^{(1)} \eta_{it}^{(1)} + \zeta_{it}^{(1)} + \theta_2^{(1)} \kappa_{i,t} + \sum_{j=1}^{J} u_{kj} \nu_{jh}^{(1)} + \gamma_{kh}^{(1)} + a_h^{(1)}. \tag{4}$$

We model subsequent purchases conditionally by $\text{logit}\big(\mathbb{P}(P_{it} = 1 | O_{it} = 1)\big)$ equaling:

$$x_{ik}^T \alpha^{(2)} + \tilde{x}_{ik}^T \beta_h^{(2)} + \theta_1^{(2)} \eta_{it}^{(2)} + \zeta_{it}^{(2)} + \theta_2^{(2)} \kappa_{i,t} + \sum_{j=1}^{J} u_{kj} \nu_{jh}^{(2)} + \gamma_{kh}^{(2)} + a_h^{(2)}, \tag{5}$$

where, $k = k(i,t)$ and $h := h(i)$ are functions of t and i respectively and $\alpha^{(l)}$, $\beta_h^{(l)}$, $\theta_1^{(l)}$, $\eta_{it}^{(l)}$, $\theta_1^{(l)}$, $\zeta_{it}^{(l)}$, $\kappa_{it}^{(l)}$, $\nu_{jh}^{(l)}$, $\gamma_{kh}^{(l)}$ denote the base effects for the open stage for $l = 1$ and the conditional effects for the purchase stage for $l = 2$. As h varies the random intercepts $(a_h^{(1)}, a_h^{(2)})$ are independent and identically distributed from a bivariate normal distribution with mean 0 and covariance Σ.

4. Estimation of CURM.

4.1. *Segmentation analysis in Stage I.* Consider m clusters. For any known grouping index $\boldsymbol{h} = \{h_i : i = 1, \ldots, n\} \in \{1, \ldots, m\}^n$ of the consumers, the joint log-likelihood of Stage I is $\ell(\Lambda, \boldsymbol{h}) = \sum_{i=1}^n \ell_i(\Lambda, h_i)$, where, $\Lambda = (\delta^{(l)}, \tau_h^{(l)}, \sigma^{(l)} : l = 1, \ldots, 4; h = 1, \ldots, m)$ and $\ell_i(\lambda, \boldsymbol{h}) = -4 \sum_{l=1,2,3} \log \sigma^{(l)} - 2^{-1} \sum_{l=1,2,3} (\sigma^{(l)})^{-2} \left(Z_i^{(l)} - w_i^T \delta^{(l)} - \tilde{w}_i^T \tau_{h_i}^{(l)} \right)^2 + z_i \log \lambda_i - \big(\sigma^{(4)}\big)^2 \log(z_i!) - Z(\lambda, \big(\sigma^{(4)}\big)^2)$. For unknown $\boldsymbol{h}$, consider $\ell_*(\Lambda) := \max_{\boldsymbol{h}} \ell(\Lambda, \boldsymbol{h}) = \sum_{i=1}^n \max_{h=1,\ldots,L} \ell_i(\Lambda, h)$. We consider Gaussian priors on the parameters $\delta^{(l)}$, $\tau_h^{(l)}$ and Inverse-Gamma priors on variances $\big(\sigma^{(l)}\big)^2$: $\delta^{(l)} \sim \mathcal{N}(\mu_\delta^{(l)}, \Sigma_\delta^{(l)})$ $\quad \tau_h^{(l)} \sim \mathcal{N}(\mu_\tau^{(l)}, \Sigma_\tau^{(l)})$

$(\sigma^{(l)})^2 \sim \text{IG}(\alpha^{(l)}, \beta^{(l)})$. Detailed choice of prior hyper-parameters is provided in supplement. The logarithm of the posterior density is:

$$\pi(\Lambda, \boldsymbol{h}|\boldsymbol{Z}) = \sum_{i=1}^{n} \ell_i(\Lambda, \boldsymbol{h}) + \sum_{l} g(\delta^{(l)}) + \sum_{l,h} g(\tau_h^{(l)}) + \sum_{l} g(\sigma^{(l)}), \tag{6}$$

where, g denotes the log-density of the aforementioned priors. We want to estimate the MAP (maximum a posteriori probability) estimator: Λ_{MAP} and $\boldsymbol{h}_{\text{MAP}}$ that maximizes our posterior density in (6).

Given a set of parameters Λ, for any observation i, we define a distance

$$\mathcal{D}_\Lambda(Z_i) = - \max_{h=1,\dots,L} \ell_i(\Lambda, h)\,. \tag{7}$$

Drawing inspiration from Arthur and Vassilvitskii (2007), we use these distances to initialize our cluster centers τ_h's.

Based on the initialization, our algorithm follows the widely used Lloyd's algorithm to find the K-means centroids (Lloyd, 1982). Lloyd's algorithm iterates across two actions, estimating cluster centers and assigning observations to these centers. We use the Metropolis-Hastings algorithm for the first action of cluster center estimation while the second action of assigning clusters to observations is achieved by likelihood maximization for each observation. This likelihood maximization essentially mirrors distance minimization in the K-means if we use distances described in (7) instead of L_2 norm.

We develop the Metropolis-Hastings based clustering algorithm (MHC), to simultaneously estimate Λ as well as the cluster grouping $\boldsymbol{h}$. The algorithm is detailed below, with theoretical support provided in the subsequent lemma. The proof is presented in the supplement.

LEMMA 4.1. *Consider the log posterior likelihood in* (6). *If* $\hat{\Lambda}$ *be the estimated parameters and* $\hat{\boldsymbol{h}}$ *be the cluster grouping from Algorithm 1 with large* L. *If the size of the largest cluster in the optimal clustering* $\boldsymbol{h}_{MAP}$ *is* $\mathcal{O}(n \log m/m)$, *then,* $\mathbb{E}[\pi(\hat{\Lambda}, \hat{\boldsymbol{h}})|\boldsymbol{Z})] \geq c \cdot \log m \cdot \pi(\Lambda_{MAP}, \boldsymbol{h}_{MAP}|\boldsymbol{Z})$, *where, c is a constant that only depends on the design matrix.*

4.2. *Parameter Estimation in Stage II.* For any fixed m we use the estimated cluster grouping indices $\hat{\boldsymbol{h}}$ from the previous section. With $\hat{\boldsymbol{h}}$ fixed (henceforth we drop the hat for notational convenience) in the second stage of CURM we conduct Bayesian estimation of the parameters in (4)-(5). We assume Gaussian priors on $\alpha^{(l)}$, $\beta_h^{(l)}$, $\theta_1^{(l)}$, $\theta_2^{(l)}$, $\nu_{jh}^{(l)}$, $\gamma_{kh}^{(l)}$ for all k, h and $l = 1, 2$, and an Inverse-Wishart prior on Σ. The full specification of the Bayesian model is as follows. For $h = 1, \dots, m$; $j = 1, \dots, J$; $k = 1, \dots, K$ and $l = 1, 2$: $\alpha^{(l)} \sim \mathcal{N}(\mu^{(l)}, \Sigma^{(l)})$, $\beta_h^{(l)} \sim \mathcal{N}(\mu_h^{(l)}, \Sigma_h^{(l)})$, $\theta_1^{(l)} \sim \mathcal{N}(\mu_{\theta_1}^{(l)}, \Sigma_{\theta_1}^{(l)})$, $\theta_2^{(l)} \sim \mathcal{N}(\mu_{\theta_2}^{(l)}, \Sigma_{\theta_2}^{(l)})$, $\nu_{jk}^{(l)} \sim \mathcal{N}(\mu_{jk}^{(l)}, \Sigma_{jk}^{(l)}), \gamma_{kh}^{(l)} \sim \mathcal{N}(\mu_{kh}^{(l)}, \Sigma_{kh}^{(l)})$, $\Sigma \sim \mathcal{W}^{-1}(\Psi_0, \nu_0)$. We detail the choice of the prior hyper-parameters in supplement.

As h_i is fixed for each i in this second stage model, we rewrite the model in (4) by pooling the coefficients across clusters. Let $\beta^{(1)} = (\beta_1^{(1)}, \dots, \beta_m^{(1)})$ and $\nu^{(1)} = (\nu_{11}^{(1)}, \dots, \nu_{J1}^{(1)}, \dots, \nu_{1m}^{(1)} \dots, \nu_{Jm}^{(1)})$, $\gamma^{(1)} = (\gamma_{1,1}^{(1)}, \dots, \gamma_{K,1}^{(1)}, \dots, \gamma_{1m}^{(1)}, \dots, \gamma_{Km}^{(1)})$ and $a^{(1)} = (a_1^{(1)}, \dots, a_m^{(1)})$. The corresponding covariates in the data matrix are also rewritten and expanded by introducing new indicator variables. For example we rewrite $\tilde{x}_{ik}^T \beta_h^{(1)}$ as $x_{ik}^e \beta^{(1)}$ where $x_{ik}^e = (e_{h_i} \otimes \tilde{x}_{ik})^T$ where e_{h_i} is a $m \times 1$ unit vector with 1 at the h_ith coordinate. Similarly, we define the expanded covariate set of z_{ik}^e corresponding to the pool parameter set $\nu^{(1)}$ as $\boldsymbol{u}_k^T \otimes e_{h_i}^T$ where $\boldsymbol{u}_k = (u_{kj} : j = 1, \dots, J)$ is a known vector. The expanded covariate set $\tilde{z}_{ik}^e$ corresponding to $\gamma^{(1)}$ equals $e_k^T \otimes e_{h_i}^T$. Thus, we rewrite (4) as

$$\begin{aligned} \text{logit}\big(\mathbb{P}(O_{it} = 1)\big) = x_{ik}^T \alpha^{(1)} + \tilde{x}_{ik}^e \beta^{(1)} + \theta_1^{(1)} \eta_{it}^{(1)} + \kappa_{it}^{(1)} + \\ \theta_2^{(1)} \zeta_{it}^{(1)} + z_{ik}^e \nu^{(1)} + \tilde{z}_{ik}^e \gamma^{(1)} + e_{h_i}^T a^{(1)}. \end{aligned} \tag{8}$$

Algorithm 1: MHC: Metropolis-Hastings Clustering for Stage I

- *Initialization.*
  - Set initial parameter estimates $\hat{\delta}^{(l)}[0] = \mu_\delta^{(l)}$ and $\hat{\sigma}^{(l)}[0] = \beta^{(l)}/(\alpha^{(l)} - 1)$ for $l = 1, \ldots, 4$ based on the mean of their priors.
  - Initialize $\tau_h[0] := (\tau_h^{(l)}[0] : l = 1, \ldots, 4)$ for all clusters h as $C\mathbf{1}$ where C is a very large number.
  - Select an initial observation i randomly from all observations. Set the initial cluster center $\tau_1[0] := (\tau_1^{(l)}[0] : l = 1, \ldots, 4)$ by minimizing the distance $\mathcal{D}_\Lambda(Z_i)$.
  - Choose the next observation i' from the remaining with probability $\mathcal{D}_\Lambda(Z_{i'})/\sum_j \mathcal{D}_\Lambda(Z_j)$. Set the k^{th} cluster center $\tau_k[0]$ by minimizing $\mathcal{D}_\Lambda(Z_{i'})$.
  - Repeat the above step till all m cluster centers are chosen.
  - Compute $\hat{h}_i[0] = \text{argmax}_h\, \ell_i(\Lambda[0], h)$ for $i = 1, \cdots, n$ and using (6) evaluate and store the posterior log-likelihood of the estimated parameters $\hat{\pi}[0] = \pi(\hat{\Lambda}[0]|\boldsymbol{Z})$.
- *Updates.* At the $(t+1)$th iteration of the algorithm,
  - Set $\hat{\Gamma}[0] = \hat{\Lambda}[t]$
  - At the $(l+1)$th iteration,
    - $*$ Perturb the estimates with white noise where the variance of noise is set as $(1/100)^{th}$ the prior variance of each parameter, $\tilde{\Gamma}[l+1] = \hat{\Gamma}[l] + \boldsymbol{\epsilon}_{l+1}$ and compute the posterior log-likelihood $\tilde{\pi}[l+1] = \pi(\Gamma[l+1]|\boldsymbol{Z})$.
    - $*$ Generate uniform random variable $u \in [0, 1]$. If $\log(u) \leq \tilde{\pi}[l+1] - \hat{\pi}[l]$, then update parameter estimates to the perturbed values $\hat{\Gamma}[l+1] = \tilde{\Gamma}[l+1]$; else $\hat{\Gamma}[l+1] = \hat{\Gamma}[l]$
  - Generate a reasonably large brunt-in chain and average thinned samples of the brunt-in chain to obtain an estimates $\hat{\Lambda}[t+1]$ and assign the observations clusters based on $\hat{\boldsymbol{h}}[t+1] = \tilde{\boldsymbol{h}}$; else $\hat{\Lambda}[t+1] = \hat{\Lambda}[t]$ and $\hat{\boldsymbol{h}}[t+1] = \text{argmax}_h\, \ell_i(\Lambda[t+1], h)$ for $i = 1, \ldots, n$.
- *Final Estimates.* We keep on computing the above updates till the estimates converge to desired precision to obtain the final estimates $\hat{\Lambda}$ of the parameters. Based on it we report the cluster indices as $\hat{h}_i = \text{argmax}_h\, \ell_i(\hat{\Lambda}, h)$.

Similar rewriting (5) for the purchase stage we have:

$$\begin{aligned} \text{logit}\big(\mathbb{P}(P_{it} = 1|O_{it} = 1)\big) = x_{ik}^T \alpha^{(2)} + \tilde{x}_{ik}^e \beta^{(2)} + \theta_1^{(2)} \eta_{it}^{(2)} + \kappa_{it}^{(2)} + \\ \theta_2^{(2)} \zeta_{it}^{(2)} + z_{ik}^e \nu^{(2)} + \tilde{z}_{ik}^e \gamma^{(2)} + e_{h_i}^T a^{(2)}. \end{aligned} \tag{9}$$

For ρ_η, ρ_ζ, we do a grid search to estimate the parameter. Now, assume ρ_η, and ρ_ζ are fixed, and the adstock, short term dynamic variables $\eta_{it}^{(l)}$s, $\zeta_{it}^{(l)}$s are known. The parameters $\alpha^{(l)}$,$\beta^{(l)}$, $\theta_1^{(l)}$, $\theta_2^{(l)}$, $\nu^{(l)}$, $\gamma^{(l)}$, $a^{(l)}$ for $l = 1, 2$ are all jointly normal given Σ. The sets of parameters $\alpha^{(l)}$,$\beta^{(l)}$ $\theta_1^{(l)}$, $\theta_2^{(l)}$,$\nu^{(l)}$, $\gamma^{(l)}$, $a^{(l)}$ for $l = 1, 2$ are all mutually independent of each other, except $a^{(1)}$ and $a^{(2)}$. Also, $(\boldsymbol{a}|\Sigma) := (a^{(1)}, a^{(2)}|\Sigma)$ is a multivariate normal distribution with mean 0 and variance $\Sigma_a = \Sigma \otimes \mathbf{1}_m$. We concatenate the design matrix in (8)-(9) and consider estimating the GLMM framework based on the response vector $\boldsymbol{y} = (O_{it} : t = 1, \ldots, T_i; i = 1, \ldots, n), (P_{ik} : t = 1, \ldots, T_i; i = 1, \ldots, n))$. Note that, the prior distribution on the parameters $(\mathcal{A}|\Sigma) := (\alpha^{(1)}, \beta^{(1)}, \theta_1^{(1)}, \theta_2^{(1)}, \nu^{(1)}, \gamma^{(1)}, a^{(1)}, \alpha^{(2)}, \beta^{(2)}, \theta_1^{(2)}, \theta_2^{(2)}, \nu^{(2)}, \gamma^{(2)}, a^{(2)}|\Sigma)$ is a multivariate normal distribution. Let $\mathfrak{m}_0$ and $\mathcal{V}_0$ be its mean and covariance parameters respectively. Let X_* denote the combined design matrix whose bth row is x_b^* for $b = 1, \ldots, B$.

Following Polson, Scott and Windle (2013), we develop a Polya-Gamma (PG) data augmentation strategy to sample efficiently from the logistic regression likelihood in (8)-(9).

For each b we generate a PG variable ω_b, and $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_B)$. The likelihood of y_i is $L_i(\mathcal{A}|\Sigma) = \exp\{y_i(x_i^*\mathcal{A} + \kappa_i)\}(1 + \exp\{x_i^*\mathcal{A} + \kappa_i\})^{-1}$. Post data augmentation with PG random variables and the prior distributions, our posterior likelihood is:

(10)

$$\pi(\boldsymbol{\omega}, \mathcal{A}, \Sigma) = g(\mathcal{A})g(\Sigma)\frac{\exp\{\sum_{b=1}^{B} y_b(x_b^*\mathcal{A} + \kappa_b) - \omega_b(x_b^*\mathcal{A} + \kappa_b)^2/2\}\Pi_{b=1}^{B} p(\omega_b)}{\Pi_{b=1}^{B}(1 + \exp\{x_i^*\mathcal{A} + \kappa_i\})\int_0^\infty \exp\{-\omega(x_b^*\mathcal{A} + \kappa_b)^2/2\}\ p(\omega)\ d\omega},$$

where, $g(\mathcal{A})$ is a multivariate normal density with mean $\mathfrak{m}$ and variance $\mathcal{V}$, $g(\Sigma)$ is the inverse-Wishart prior with parameters Ψ_0, ν_0 and $p(\omega)$ is density of Polya-Gamma random variable $PG(1, 0)$. We derive a Gibbs sampler to estimate the parameters in user response model (GSURM) based on the posterior in (10). We describe the algorithm below. In lemma 2 whose proof is provided in the supplement, we show that GSURM is a valid sampler for (10).

Algorithm 2: GSURM: Gibbs Sampler for estimating user response model in Stage II

- *Initialization.*
  – In the beginning, the parameter estimates $\hat{\mathcal{A}}$ and $\hat{\Sigma}$ are set to the mean of their priors.
- *Updates.* At the $(t+1)$ th iteration of the algorithm,
  – Update $\big(\omega_b[t+1]|\mathcal{A}[t]\big) \sim \text{PG}\big(1, x_b^*\mathcal{A}[t] + \kappa_b\big)$
  – Set $\mathcal{V}_t$ by adjoining $\mathcal{V}_0$ and $\Sigma_t \otimes \mathbf{1}_m$ as block diagonal matrices.
  – Compute $\mathcal{V}[t+1] = (X_*^T\Omega[t+1]X_* + \mathcal{V}_t^{-1})^{-1}$, $\Omega[t+1]$ is a diagonal matrix of $\omega_b[t+1]$s
  – Compute $\mathfrak{m}[t+1]$ is given by $\mathcal{V}[t+1]\{X_*^T(Y - \Omega[t+1]\boldsymbol{\kappa} - 0.5) + \mathcal{V}_t^{-1}\mathfrak{m}_0\}$
  – Update $\big(\mathcal{A}[t+1]|\Sigma[t], \boldsymbol{\omega}[t+1]\big) \sim \mathcal{N}(\mathfrak{m}[t+1], \mathcal{V}[t+1])$
  – Update $\big(\Sigma[t+1]|\mathcal{A}[t+1]\big) \sim \mathcal{W}^{-1}\left(\Psi_0 + \sum_{h=1}^m a_h[t+1]\, a_h^T[t+1],\, \nu_0 + m\right)$
- *Final Estimates.* We keep on computing the above updates till the estimates converge to desired precision. Thereafter, we generate a reasonably large brunt-in chain and by averaging thinned samples of the brunt-in chain we obtain the final estimates $\hat{\mathcal{A}}$ and $\hat{\Sigma}$ of the parameters.

LEMMA 4.2. *GSURM presented in algorithm 2 is a valid Gibbs sampler for the posterior distribution (10) for modeling user responses in Stage II of CURM.*

5. PACURM: Path Algorithm for Integrated Analysis. The final stage of our algorithm is to provide an efficient mechanism to complete an analysis for varying cluster sizes using a bottom-up approach. We start with m clusters and at each stage provide a step-by-step procedure to go from h clusters to $h-1$ clusters. For each $h = m, \ldots, 2$, we (a) find the two closest clusters which we will merge to create a larger cluster, and (b) update cluster-specific parameters for the merged cluster.

The likelihood of our Stage I model, as specified in (1)-(2), generally decreases when two clusters are merged. We define the distance $D(v, \tilde{v})$ between two clusters (indexed by their centroids) v and $\tilde{v}$ as the decrease in the log-likelihood based on our model, once the two clusters are merged. We intend to find the closest pair among h clusters which is same as finding the pair of clusters whose merging causes least decrease in the log-likelihood.

Consider a pair of clusters $(v, \tilde{v})$ from $\{1, 2, \cdots, h\}$. Define, A_v as the set of all observations that belongs to the cluster with centroid at v. To calculate distance between these two clusters, we first need to estimate the cluster specific parameters for the merged cluster,

$\hat{\tau}^{(l)}_{(v,\tilde{v})}$'s which will replace the two set of parameters $\hat{\tau}^{(l)}_{v}$ and $\hat{\tau}^{(l)}_{\tilde{v}}$. $\hat{\tau}^{(l)}_{(v,\tilde{v})}$'s is the parameter which maximizes the likelihood given all the other parameters remain fixed. Observe that for a fixed $(v,\tilde{v})$ the changes in likelihood happens only for observations in $A_v \cup A_{\tilde{v}}$ and the prior of $\hat{\tau}^{(l)}_{(v,\tilde{v})}$'s. Also, the models in Section 3.1 are independent of each other given h, hence we can update $\hat{\tau}^{(l)}$ separately for each $l = 1,\dots,4$ as:

(11)
$$\hat{\tau}^{(l)}_{(v,\tilde{v})} = \max_{\tau^{(l)}} \sum_{i\in A_v\cup A_{\tilde{v}}} -\frac{1}{2}(\sigma^{(l)})^{-2}(Z_i(l) - w_i^T\delta^{(l)} - \tilde{w}_i^T\tau^{(l)})^2 + g(\tau^{(l)}) \quad \text{for } l = 1,2,3$$

(12)
$$\hat{\tau}^{(4)}_{(u,v)} = \max_{\tau^{(l)}} \sum_{i\in A_v\cup A_{\tilde{v}}} (\sigma^{(4)})^{-2}(Z_i(4)\log(w_i^T\delta^{(4)} + \tilde{w}_i^T\tau^{(4)}) - w_i^T\delta^{(4)} + \tilde{w}_i^T\tau^{(4)}) + g(\tau^{(4)}).$$

(11) is a Bayesian least square regression with the parameters $\tilde{Z}_{(v,\tilde{v})}(l)$ and $\tilde{W}_{(v,\tilde{v})}$, $\tilde{Z}_{(v,\tilde{v})}(l) := [Z_i(l) - w_i^T\delta^{(l)}]_{i\in A_v\cup A_{\tilde{v}}}$ and $\tilde{W}_{(v,\tilde{v})} := [w_i]_{i\in A_v\cup A_{\tilde{v}}}$. The posterior mean of $\hat{\tau}^{(l)}_{(v,\tilde{v})}$ is

$$(13)\qquad (\tilde{W}^T_{(v,\tilde{v})}\tilde{W}_{(v,\tilde{v})} + (\sigma^{(l)})^{-2}(\Sigma^{(l)}_{\tau})^{-1})^{-1}(\tilde{W}^T_{(v,\tilde{v})}\tilde{Z}_{(v,\tilde{v})}(l) + \mu^{(l)}_{\tau}\left((\sigma^{(l)})^{-2}(\Sigma^{(l)}_{\tau})^{-1}\right))$$

where as, (12) reduces to solving $\sum_{i\in A_v\cup A_{\tilde{v}}}(\sigma^{(4)})^{-2}\left(\frac{Z_i(4)\tilde{w}_i}{w_i^T\delta^{(4)}+\tilde{w}_i^T\tau^{(4)}} - \tilde{w}_i\right) - \Sigma_{\tau}^{-1}(\tau^{(4)} - \mu_{\tau}) = 0$, for which we use gradient descent. With estimates $\hat{\tau}^{(l)}_{(v,\tilde{v})}$'s, we compute $\mathcal{D}(v,\tilde{v})$ as the negative of the change in log-likelihood. Explicit expression is provided in (20) of supplement. Based on these distances, we find the two clusters closest to each other and merge them.

Upon merging two closest clusters, we update the parameters for our Stage II logistic models. Based on $\hat{\beta}^{(j)}_{v}$ and $\hat{\beta}^{(j)}_{\tilde{v}}$ an asymptotically consistent estimate of the parameters of the combined cluster is: $\hat{\beta}^{(j)}_{(v,\tilde{v})} = ((\hat{\Sigma}^{(j)}_{v})^{-1/2} + (\hat{\Sigma}^{(j)}_{\tilde{v}})^{-1/2})^{-1}\left((\hat{\Sigma}^{(j)}_{v})^{-1/2}\hat{\beta}^{(j)}_{v} + (\hat{\Sigma}^{(j)}_{\tilde{v}})^{-1/2}\hat{\beta}^{(j)}_{\tilde{v}}\right)$ for j= 1,2, where, $\hat{\Sigma}^{(j)}_{v}$ and $\hat{\Sigma}^{(j)}_{\tilde{v}}$ are estimated using the MCMC samples of the corresponding parameter. Note that for each cluster h, we need to use samples of all the parameters from the original m clusters parameter estimation. Suppose h is one of the clusters we want to merge and h has been created by merging k clusters $\{h_1, h_2, \cdots, h_k\}$ in the original m clusters, then the estimation of $\hat{\Sigma}^{(j)}_{h}$ should utilize all the MCMC samples from $\beta^{(j)}_{h_1}, \beta^{(j)}_{h_2}, \cdots, \beta^{(j)}_{h_k}$.

5.1. *Implementation and R package* `PACURM`. The developed R code for implementing the proposed algorithm can be downloaded from GitHub repository https://github.com/rrbhuyan/PACURM. The repository also contains the data, plots and results presented in the paper.

5.2. *Computational Complexity.* To study the computational efficiency of CURM, we analyze Algorithms 1 and 2 for the two stages separately. In addition, we also analyze the the path-algorithm to derive the complexity for PACURM.

Recall that the number of users is n, the number of coupons is K. Let p represent the total number of covariates, which includes both global and cluster-specific variables. In Algorithm 1, define T_1, L as the number of iterations for the outer and inner loops, respectively, and T_2 is the number of iterations for Algorithm 2.

Consider m clusters. Then, in Algorithm 1, the expensive step is the likelihood computation inside the inner loop. In each computation, for a single user, the number of operations is

$\mathcal{O}(mp)$. Therefore, the total computational effort for executing both the inner and outer loops is $\mathcal{O}(T_1 Lnmp)$.

In Stage II, estimating the conditional mean of $(\mathcal{A}[t+1]|\Sigma[t], \boldsymbol{\omega}[t+1])$, $\mathfrak{m}[t+1]$ is computationally the most expensive step. The logistic models in (4)-(5) have $\mathcal{O}(nK)$ samples and $\mathcal{O}(mp)$ covariates. Hence, conditional mean estimation is $\mathcal{O}(nK(mp)^2)$. Since, we need to compute the conditional mean once for each iteration of the Gibbs sampler, the complexity of Algorithm 2 is $\mathcal{O}(T_2 nKm^2p^2)$.

Finally, analyzing the path algorithm, we examine the complexity of reducing the cluster count from h to $h-1$. The computationally expensive step here is solving (13), which has order $\mathcal{O}(np^2)$ and we have to perform this step $h-1$ times. Thus, to move from h to $h-1$ clusters, the complexity is $\mathcal{O}(hnp^2)$ and the complexity of starting at m clusters and aggregating till two clusters is $\mathcal{O}(m^2np^2)$. This latter complexity is significantly lower compared to the complexity of Stage II for m clusters. We summarize the computational complexity result below.

LEMMA 5.1. *The computational complexity of PACURM for providing the path of solutions from* 2 *to* m *clusters is* $\mathcal{O}(nKm^2p^2\mathcal{T})$, *where,* p *is the number of covariates and* $\mathcal{T}$ *is the burn-in time for GSURM.*

6. Results. In our study, we selected a test set comprising 10000 randomly chosen consumers from a total of 77986 consumers in our dataset. We begin by presenting results from Stage I of CURM. Subsequently, we demonstrate the impact of considering a consumer's recent engagement and the predictive power of the messaging components. We also highlight the significant role of overdispersion parameters in CURM and finally, we conclude with the results from the path algorithm.

6.1. *Segmentation analysis based on Stage I of CURM.* In Stage I, CURM segments consumers based on the historical user characteristics. As described in Section 3.1, we use GCLR via (1)-(2) on the RFM variables with demographic variables age and income as the regressors.

Age is treated as an ordinal variable with five levels, while income is categorized into eight nominal groups. Along with these, we use cluster varying intercepts. We estimate the GCLR model for both 3 and 5 clusters as outlined in Section 4.1 and report their centroids in Tables 4 and 5 respectively.

Cluster	Cluster Size	Recency$^{(o)}$	Recency$^{(p)}$	Frequency	Monetary
I	34147	12.5	53.0	6.0	3.7
II	30753	5.5	68.1	11.1	3.8
III	13086	12.2	249.3	3.7	3.6

Table 4: Average value of the RFM variables in the 3 clusters

In table 4, the three clusters can be tagged as high, medium, and low. The low engagement group (cluster III) seldom buys as they have a significantly small frequency (no. of orders) and large recency variable (days since last purchased). Contrast to that, the medium (cluster I) and high engagement (cluster II) groups have larger frequency variables and smaller recency variables. The monetary variable follow the same order but have a much lower spread. Expanding to 5 clusters provides a more nuanced understanding of consumer behaviors. For example, cluster IV in table 5 consists of the frequent shoppers who have ordered an average

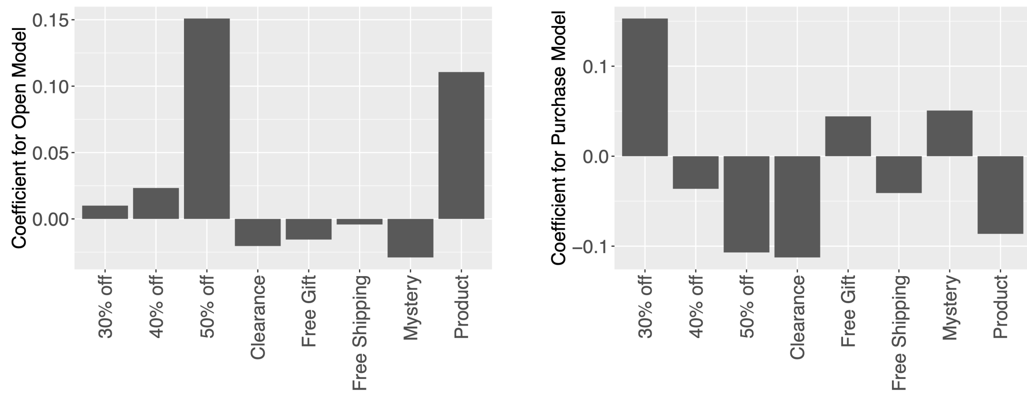


Fig 3: Barplot of overall effects (coefficients) of different messaging components on consumer responses. Component ‘Free Returns’ is the baseline with zero effect.

of 12.8 times where as cluster II consumers open coupons very often (days since last open 5.7), but they seldom buy as average frequency for the cluster in last two years is 2.4, and last purchases were made on average 158.8 days ago.

Cluster	Cluster Size	Recency$^{(o)}$	Recency$^{(p)}$	Frequency	Monetary
I	23194	6.8	65.1	4.6	3.8
II	7419	5.7	158.8	2.4	3.4
III	17359	11.6	152.7	6.9	3.8
IV	25631	7.5	58.9	12.8	3.7
V	4383	36.0	79.3	5.2	3.7

Table 5: Average value of the RFM variables in the 5 clusters

6.2. *Explaining segmentation, engagement and component effects.* To understand the utility of different attributes in CURM on its predictive performance, we consider three sub-models of CURM. As a baseline, we consider CURM with a single cluster, which we refer to as URM. URM is just fitting Stage II of CURM.

In Figure 3, we see the varying effects of these digital components on the open and purchase propensity of consumers. As we observe the open rates, we notice a pattern of increasing impact across the ‘30% off’, ‘40% off’ and ‘50% off’ messaging components. This progression logically aligns with the notion that higher discount percentages tend to be more lucrative, prompting consumers to open the email. Additionally, we observe that the priced messaging components such as ‘30% off’, ‘40% off’, ‘50% off’, ‘Clearance’ and ‘Product’ provide on average a higher rate of opening compared to the non-priced components –‘Free Gift’, ‘Free Shipping’, ‘Mystery’ and ‘Free returns’. In the supplemental Tables A1 and A2, we provide the 95% credible intervals for the coefficients of messaging components for the open and purchase stages respectively. In the open stage, all except two components are significantly different from the baseline effect of ‘Free Returns’. Specifically, the priced promotions involving ‘50% off’ and ‘Product’ have significantly higher impact than ‘Free Returns’ whereas ‘Free Gift’, ‘Free Shipping’, ‘Mystery’ are provably sub-optimal than ‘Free Returns’ at 5% level.

Next, we take into account the clustering we performed with 5 cluster in the previous section and examine the cluster-specific coefficients for the messaging components. In Figure 4, we see the varying effects of these messaging components across consumer clusters. Supplemental Tables A3 and A4 provide the coefficient estimates and their 95% credible intervals. We witness that the monotone relationship in the estimates of the '30% off', '40% off' and '50% off' components is consistent across all the clusters. Additionally, we observe significant differences in the efficacy of these components across clusters. For example, 'Clearance' is highly effective in the open stage for Cluster IV, while for the remaining clusters it is not as effective. Examining the RFM values for Cluster IV in Table 5, it's evident that this group places the highest number of orders with the shortest time since their last purchase. Given these insights, it is sensible to target Cluster IV with coupons featuring clearance offers, taking advantage of their high frequency of purchases.

These cluster-specific effects are instrumental in understanding the effectiveness of digital coupons and in tailoring new coupons for specific clusters. For example, while a '30% off' coupon will be effective for a highly engaged user, a 'free return' coupon might entice a low engagement user to use the platform to make a purchase. These highly customized component effects for each cluster is important in personalizing coupons for each cluster.

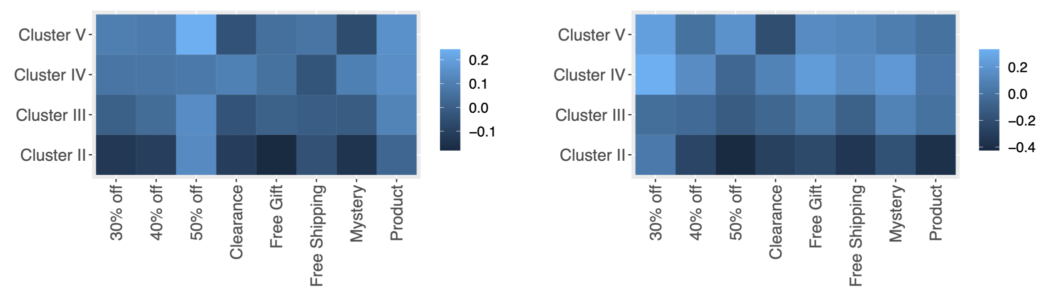
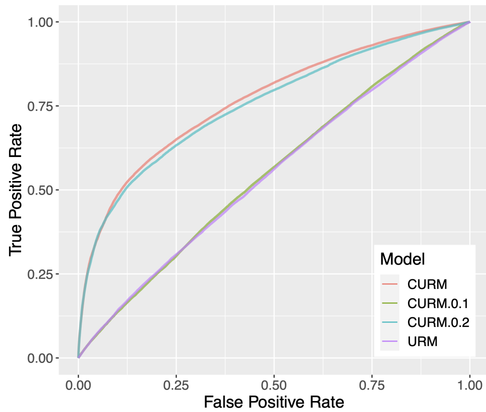


Fig 4: Heatmap of effects (coefficients) of different messaging components on consumer responses across cluster segments. The left pane displays the coefficients in the open stage model and the right pane shows the coefficients in the purchase stage model. Component 'Free Returns' and cluster I is the baseline.

The estimated coefficients for θ_1 are -0.58 and -0.26 for the open and purchase stages, respectively. These negative coefficients indicate that previous coupons have adverse effects on consumer's response to current coupon, leading to a significant reduction in the likelihood of both opening the coupon and making a purchase based on the current coupon. In our application case, a consumer received on average 20 promotions in a 45 days window. It would thus be advisable to reassess the frequency with which these communications are sent. By reducing the frequency, retailers can potentially mitigate the adverse effects of consumer fatigue and enhance consumer response rates.

Similarly, we also observe negative estimates of -0.31 and -0.52 for κ coefficients corresponding to the purchase engagement variable in open and purchase models. In the apparel industry, consumers seldom make multiple purchases within a 45-day short campaign window. Thus, the negative values of the estimates are in concordance as it suggest a significant decrease in consumer engagement for the subsequent coupons once a purchase is made. In CURM.0.2, the γ coefficients, pertaining to the residual impact of each coupon is set to 0. As the γ coefficients are the coupon effects unexplained by the dictionary, the difference in predictive performance between the CURM and CURM.0.2 models will reflect the explanatory power of the dictionary – lower the difference, higher the power. In CURM.0.1, the θ and ζ coefficients pertaining to the impact of recent communications and user engagements within the campaign window are also set to 0. In Figure 5 for URM as well as the three CURM variants with 5 clusters, we report the receiver operations characteristics (ROC) curves for both



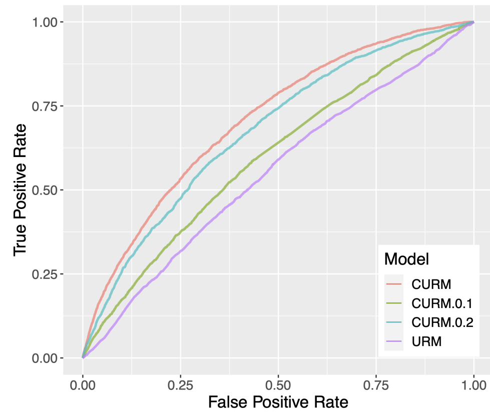


Fig 5: Receiver operations characteristics (ROC) curves for the 4 CURM based models on test data. The left pane represents the open response prediction and the plot on the right side represents the purchase response prediction.

the open and purchase decisions. In Table 6, we provide area under the curve (AUC) for these ROCs. In the supplemental Tables A5-A8, we provide the coefficient estimates and their 95% credible intervals for URM as well as the three CURM variants. Compared to our base model URM, CURM performs almost 28% better in the purchase stage while in the open stage it performs around 40% better.There is a noticeable improvement in predictive performance on adding the short-term dynamic consumer engagement variables which is reflected by discernible gap between the ROCs from CURM.0.1 and CURM.0.2. The difference between CURM.0.1 and URM is negligible in open stage and innappreciable compared to short-term engagement impact in the purchase stage, which suggest that the dictionary used here based on the catalogue in Ellickson, Kar and Reeder III (2022) very well explains the 25 coupons present in the data.

Model	Open Rates	Purchase Rates
URM	0.547	0.552
CURM.0.1	0.550	0.598
CURM.0.2	0.753	0.676
CURM	0.767	0.709

Table 6: Area under the curve (AUC) for 4 models in test data

6.3. *Role of Overdispersion in Stage I.* It is important to use an overdispersed Poisson GLM model specified in (2), since the mean and variance of frequency of purchases is different. We estimate $\hat{\sigma}^{(4)} = 0.57$ which means that variance is approximately 3 times the mean. In Table 7, we illustrate the differences between the clusters generated by our model on test data, with and without overdispersion (in this scenario, we ran Algorithm 1 setting $\sigma^{(4)} = 1$).

While most of the consumers remained in the same cluster for both the models, around 17% were not in the same cluster. To properly compare whether the overdispersion added any benefit we compare the log-likelihood for both the models. In Table 8 we provide the average log-likelihoods for both the models for both training as well as testing data.

The average log-likelihood in the test data is approximately 35% lower in the model without overdispersion compared to our CURM model. In Figure 6, we plot the difference in

		CURM (No overdispersion)				
		I	II	III	IV	V
CURM	I	1781	0	8	1119	0
	II	17	504	457	14	0
	III	0	0	2128	82	0
	IV	0	0	0	3350	0
	V	0	0	14	60	466

Table 7: Comparison of cluster overlap in test sample between CURM with and without overdispersion

	Train Data	Test Data
CURM	-20.88	-20.78
CURM (No overdispersion)	-27.58	-27.63

Table 8: Average log-likelihood of CURM vs CURM without overdispersion in training and testing data

log-likelihood across all the 10000 consumers in test data. For most users, the predictive log-likelihood from CURM is significantly higher than that of the model without overdispersion.

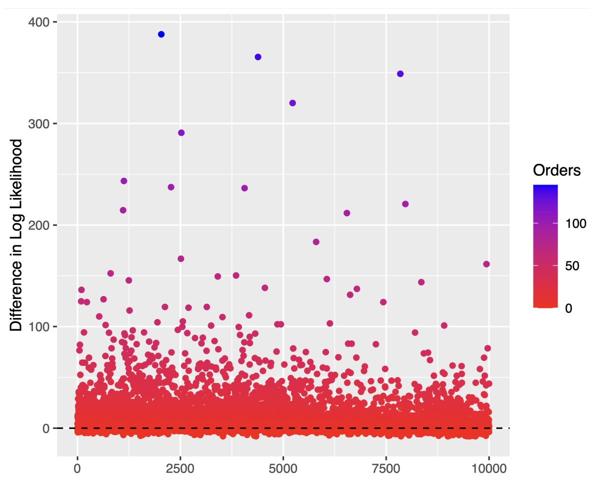


Fig 6: Difference in log-likelihood of CURM with and without overdispersion for each users in test data. The color signifies the historical frequency (number of orders) from the user.

Figure 6 shows that there are significant consumers for whom the log-likelihood of the model without overdispersion is better (points below the $y = 0$ line) but the difference is quite small. However, the log-likelihood for CURM is consistently higher than that for CURM without overdispersion for the majority of consumers, and the differences are more pronounced. Also, we can see from the Figure that the major differences occur when the number of orders is large since Poisson without overdispersion can't take very large values.

6.4. *Path Analysis.* Now, we apply our path algorithm to provide description of clusters as the number of clusters vary from 10 to 1. We initially segment the training data into 10 distinct clusters. In Table 9 presents the cluster centroids.

Cluster	Cluster Size	Recency$^{(o)}$	Recency$^{(p)}$	Frequency	Monetary
I	17447	6.9	85.0	6.8	3.7
II	9476	5.7	45.9	6.6	3.8
III	2739	8.3	308.2	5.2	3.5
IV	9202	5.3	51.6	5.1	3.8
V	217	242.3	159.4	9.8	3.5
VI	15201	10.8	35.0	13.4	3.7
VII	56	11.8	1677.6	0.0	0.0
VIII	4495	10.5	59.5	16.7	3.8
IX	4920	5.5	170.9	2.7	3.9
X	14233	15.3	153.9	4.0	3.7

Table 9: Average value of the RFM variables in the 10 clusters

This segmentation allows for a detailed understanding of consumer behaviors. For example, Cluster VII in Table 9 comprises a small group of consumers who have never made a purchase. On the other hand, consumers in Cluster VIII exhibit a distinct behavior; while they open coupons with a similar frequency (days since last open 10.5), they stand out for their remarkably frequent purchases, an average 16.7 times in the last two years. We also observe Cluster IX, akin to Cluster II in Table 5, which displays high coupon opening rates (days since last open 5.5), yet exhibits infrequent purchasing behavior. The average frequency of purchase for consumers in this cluster is 2.7 purchases, with the last purchase made on average 170.9 days ago.

Starting with ten clusters, we employ the path algorithm to progressively reduce the number of clusters by merging the closest ones and re-estimating the parameters. The PACURM algorithm utilizes changes in likelihood post-merging to decide which clusters to merge. Initially, the smallest two clusters, Cluster V and Cluster VII, are merged. Subsequently, Cluster III is merged into this combined cluster. The dendrogram in supplemental figure 7 illustrates how the clusters evolve and consolidate over time.

This approach provides an integrated framework to study consumer responses using the CURM model at various levels of granularity. It also facilitates the judicious selection of the optimal resolution level that balances potential revenue gains from increasing consumer segmentation against the implementation and operational costs associated with maintaining a higher number of clusters.

7. Discussion. In this paper, we studied the impact of messaging components in digital communications. The components were based on a communications theory based catalogue used in Ellickson, Kar and Reeder III (2022) for indexing text based email communications. There has been an explosive recent growth in electronic communications and cataloguing unstructured, complex messages needs usage of high-dimensional dictionary. For instance, the analysis of communication images and the development of techniques to automatically identify and measure human-interpretable visual characteristics using large dictionaries represent a vibrant area of research. Often, with these high-dimensional dictionaries we encounter high proportions of rare features (Yan and Bien, 2021). Following Makalic and Schmidt (2016); Biswas, Mackey and Meng (2022), spike-and-slab priors can be used in Polya-Gamma based data augmentation schemes for estimation in high-dimensional sparse logistic models. As future work, it will be interesting to extend the PACURM framework to high-dimensional setups with rare features.

REFERENCES

ANCHANG, B., HART, T. D., BENDALL, S. C., QIU, P., BJORNSON, Z., LINDERMAN, M., NOLAN, G. P. and PLEVRITIS, S. K. (2016). Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nature protocols* **11** 1264–1279.

APPLEBY, C. and KUMAR, I. K. (2023). Integrating Digital Health Tools to Improve the Consumer Experience. Available at https://deloitte.wsj.com/cmo/integrating-digital-health-tools-to-improve-the-consumer-experience-9c80fb24.

ARTHUR, D. and VASSILVITSKII, S. (2007). k-means++: The advantages of careful seeding. **7**.

BISWAS, N., MACKEY, L. and MENG, X.-L. (2022). Scalable spike-and-slab. In *International Conference on Machine Learning* 2021–2040. PMLR.

BURKHOLDER, G. J. and EVERS, K. A. (2002). *Application of the Transtheoretical Model to several problem behaviors.* Springer Publishing Company.

CHABÉ-FERRET, S. (2022). Statistical Tools for Causal Inference. Available at https://chabefer.github.io/STCI/.

CHEN, S., ZHU, B., HUANG, S., HICKEY, J. W., LIN, K. Z., SNYDER, M., GREENLEAF, W. J., NOLAN, G. P., ZHANG, N. R. and MA, Z. (2023). Integration of spatial and single-cell data across modalities with weakly linked features. *Nature Biotechnology* 1–11.

CHI, E. C. and LANGE, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics* **24** 994–1013.

DESARBO, W. S. and CRON, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification* **5** 249–282.

DETWEILER, J. B., BEDELL, B. T., SALOVEY, P., PRONIN, E. and ROTHMAN, A. J. (1999). Message framing and sunscreen use: gain-framed messages motivate beach-goers. *Health psychology* **18** 189.

ELLICKSON, P. B., KAR, W. and REEDER III, J. C. (2022). Estimating Marketing Component Effects: Double Machine Learning from Targeted Digital Promotions. *Marketing Science*.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33** 1.

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

GLAZER, E. and KROUSE, S. (2020). Dear Voter, Here's Why Political Texts Are Blowing Up Your Phone. Available at https://www.wsj.com/articles/dear-voter-heres-why-political-texts-are-blowing-up-your-phone-11582210800.

HARTIGAN, J. A. (1985). Statistical theory in clustering. *Journal of classification* **2** 63–76.

HENNIG, C. (2000). Identifiablity of Models for Clusterwise Linear Regression. *Journal of classification* **17**.

HICKEY, J. W., BECKER, W. R., NEVINS, S. A., HORNING, A., PEREZ, A. E., ZHU, C., ZHU, B., WEI, B., CHIU, R., CHEN, D. C. et al. (2023). Organization of the human intestine at single-cell resolution. *Nature* **619** 572–584.

JUNG, K., LEE, B. Y. et al. (2010). Online vs. offline coupon redemption behaviors. *International Business & Economics Research Journal (IBER)* **9**.

KUNDA, Z. (1990). The case for motivated reasoning. *Psychological bulletin* **108** 480.

LAMB, C. W., HAIR, J. F. and MCDANIEL, C. (2011). *Essentials of marketing*. Cengage Learning.

LEONE, R. P. (1995). Generalizing what is known about temporal aggregation and advertising carryover. *Marketing Science* **14** G141–G150.

LLOYD, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory* **28** 129–137.

LUO, Z. and CHOU, E. Y. (2006). Pavement condition prediction using clusterwise regression. *Transportation research record* **1974** 70–77.

MAKALIC, E. and SCHMIDT, D. F. (2016). High-dimensional Bayesian regularised regression with the BayesReg package. *arXiv preprint arXiv:1611.06649*.

MALSINER-WALLI, G., PAUGER, D. and WAGNER, H. (2018). Effect fusion using model-based clustering. *Statistical Modelling* **18** 175–196.

MANN, T., SHERMAN, D. and UPDEGRAFF, J. (2004). Dispositional motivations and message framing: a test of the congruency hypothesis in college students. *Health Psychology* **23** 330.

NELSON, E. (2023). Breaking barriers: Enabling innovation with global communications. Available at https://www.ericsson.com/en/blog/1/2023/breaking-barriers-enabling-innovation-with-global-communications.

NG, S.-K., MCLACHLAN, G. J., WANG, K., BEN-TOVIM JONES, L. and NG, S.-W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics* **22** 1745–1752.

OTTER, T., TÜCHLER, R. and FRÜHWIRTH-SCHNATTER, S. (2004). Capturing consumer heterogeneity in metric conjoint analysis using Bayesian mixture models. *International Journal of Research in Marketing* **21** 285–297.

PANDIT, A. (2022). How Governments Are Using Digital Communications for Richer Engagement. Available at https://deloitte.wsj.com/cmo/how-governments-are-using-digital-communications-for-richer-engagement-01658861557.

PARK, M. Y. and HASTIE, T. (2007). L 1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69** 659–677.

PARK, C. H., PARK, Y.-H. and SCHWEIDEL, D. A. (2018). The effects of mobile promotions on customer purchase dynamics. *International Journal of Research in Marketing* **35** 453–470.

PETTY, R. E., CACIOPPO, J. T., PETTY, R. E. and CACIOPPO, J. T. (1986). *The elaboration likelihood model of persuasion*. Springer.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association* **108** 1339–1349.

RADCHENKO, P. and MUKHERJEE, G. (2017). Convex clustering via l1 fusion penalization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 1527–1546.

RIMER, B. K. and GLASSMAN, B. (1999). Is there a use for tailored print communications in cancer risk communication? *JNCI Monographs* **1999** 140–148.

ROBERTS, G. O. and TWEEDIE, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110.

SEN, N., MUKHERJEE, G., SEN, A., BENDALL, S. C., SUNG, P., NOLAN, G. P. and ARVIN, A. M. (2014). Single-cell mass cytometry analysis of human tonsil T cell remodeling by varicella zoster virus. *Cell reports* **8** 633–645.

SILVA, A., ROTHSTEIN, S. J., MCNICHOLAS, P. D. and SUBEDI, S. (2019). A multivariate Poisson-log normal mixture model for clustering transcriptome sequencing data. *BMC bioinformatics* **20** 1–11.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software* **39** 1.

SNYDER, M. and DEBONO, K. G. (2014). A functional approach to attitudes and persuasion. In *Social Influence* 107–125. Psychology Press.

SPÄTH, H. (2014). *Mathematical algorithms for linear regression*. Academic Press.

STATISTA (2022). Topic: Digital Advertising in the U.S. Available at: https://www.statista.com/topics/1176/online-advertising/#topicHeader__wrapper.

STATISTA (2023). Time spent with digital vs. traditional media in the U.S. 2011-2025. Available at https://www.statista.com/statistics/565628/time-spent-digital-traditional-media-usa/.

TABAS, J. (2021). How Nonprofits Can Use Social Media To Increase Donations And Boost Visibility. Available at https://www.forbes.com/sites/allbusiness/2021/03/06/how-nonprofits-can-use-social-media-to-increase-donations-and-boost-visibility/?sh=113e61292bb7.

TAN, K. M. and WITTEN, D. (2015). Statistical properties of convex clustering. *Electronic journal of statistics* **9** 2324.

VENKATESAN, R. and FARRIS, P. W. (2012). Measuring and managing returns from retailer-customized coupon campaigns. *Journal of marketing* **76** 76–94.

WEDEL, M. and KISTEMAKER, C. (1989). Consumer benefit segmentation using clusterwise linear regression. *International Journal of Research in Marketing* **6** 45–59.

YAN, X. and BIEN, J. (2021). Rare feature selection in high dimensions. *Journal of the American Statistical Association* **116** 887–900.

SUPPLEMENTARY MATERIAL

APPENDIX A: ORGANIZATION OF THE SUPPLEMENT

In Section B of the supplement, we present the proofs of the lemmas discussed in Section 4, along with details for the selection of prior hyper-parameters for Algorithms 1 and 2. In Section C, we offer a comprehensive proof of Lemma 5.1 concerning the computational complexity of PACURM, as well as a precise definition for computing the distance between two clusters. Lastly, in Section D, we include all the relevant tables referenced in Section 6 of the main paper.

APPENDIX B: FURTHER DETAILS ON CURM ESTIMATION

B.1. Proof of Lemma 1. Consider the particular set-up where $W = 0$, $\tilde{W} = 1$, $\sigma^{(l)} = 1$ for all l in (1) and the cluster centers have standard Gaussian prior. Further we assume that all the models are linear. The proof can easily be extended to the generic set-up. Under these assumption, for $l = 1, 2, 3$, we model,

$$Z_i(l) = \tau^{(l)}_{h(i)} + \varepsilon_{il}, \text{ for } i = 1, \cdots, n, \tag{14}$$

where $\tau^{(l)}_h \sim \mathcal{N}(0, 1)$.

Note that in Algorithm 1, at time $t+1$, we run a random walk based Hastings-Metropolis with fixed cluster $\hat{h}_i[t]$. Once this clustering is fixed, the model boils down to linear models with Gaussian priors. The posterior distribution is thus from an exponential family. This means that following Roberts and Tweedie (1996), the Markov chain is geometrically Ergodic. Since L is large, the mean of the thinned samples after the burn in period converges to the posterior mean. Thus, at each stage, the log-likelihood $\pi(\hat{\Lambda})$ increases. Thus, we need to prove

$$\mathbb{E}[\pi(\hat{\Lambda}[0], \hat{\boldsymbol{h}}[0])|\boldsymbol{Z})] \geq c\,(2 + \log m)\,\pi(\Lambda_{\text{MAP}}, \boldsymbol{h}_{\text{MAP}}|\boldsymbol{Z})\,.$$

In this simple set-up, the distances $\mathcal{D}_\Lambda(Z_i) \propto \min_h \|Z_i - \tau_h\|^2$, the initialization is exactly the same as k++ means in Arthur and Vassilvitskii (2007). For ease of notation $\hat{h} = \hat{h}[0]$, $\hat{\tau} = \hat{\tau}[0]$ and $h = h_{\text{MAP}}$, $\tau = \tau_{\text{MAP}}$. Also let $A_1, A_2, \cdots A_m$ be the sets of points in the m clusters corresponding to h and $C(A) = |A|^{-1} \sum_{i \in A} Z_i$. Thus using Theorem 1 from Arthur and Vassilvitskii (2007), we can rewrite the same for our purpose as

$$\sum_i \|Z_i - \hat{\tau}_{\hat{h}_i}\|^2 \leq 8(\log m + 2) \sum_{h \in \{1,2,\cdots,m\}} \sum_{Z_j \in A_h} \|Z_j - C(A_h)\|^2\,. \tag{15}$$

For a fixed cluster h, the optimal cluster center τ_h depends only on the points in A_h. Thus, $\tau_h = \text{argmax}_\tau \sum_{Z_j \in A_h} \|Z_j - \tau\|^2 + \|\tau\|^2$. The first term corresponds to the likelihood of all the points in A_h and the second term is for the prior on τ_h. The optimal solution for τ_h is thus

$$\tau_h = (|A_h| - 1)^{-1} \sum_{Z_j \in A_h} Z_j = C(A_h) \frac{|A_h|}{|A_h| - 1}\,.$$

We can rewrite $\sum_{Z_i \in A} \|Z_i - C(A)\|^2$ in terms of $C(A)|A|/(|A| - 1)$ as

$$= \sum_{Z_i \in A} \left\| Z_i - C(A)\frac{|A|}{|A| - 1} + C(A)\frac{1}{|A| - 1} \right\|^2$$

$$= \sum_{Z_i \in A} \left\| Z_i - C(A)\frac{|A|}{|A| - 1} \right\|^2 + \left\| C(A)\frac{1}{|A| - 1} \right\|^2$$

$$+\sum_{Z_i\in A}\left(Z_i - C(A)\frac{|A|}{|A|-1}\right)^T\left(C(A)\frac{1}{|A|-1}\right)$$

$$=\sum_{Z_i\in A}\left\|Z_i - C(A)\frac{|A|}{|A|-1}\right\|^2 + \|C(A)\|^2\frac{1-2|A|}{(|A|-1)^2} \le \sum_{Z_i\in A}\left\|Z_i - C(A)\frac{|A|}{|A|-1}\right\|^2,$$

where the last inequality follows since $1-2|A|<0$. Using this we can rewrite (15) as

$$\mathbb{E}\left[\sum_i \|Z_i - \hat{\tau}_{\hat{h}_i}\|^2\right] \le 8(\log m + 2)\sum_{h\in\{1,2,\cdots,m\}}\sum_{Z_j\in A_h}\|Z_j - \tau_h\|^2,$$

since the optimal solution is $\tau_h = C(A_h)|A_h|/(|A_h|-1)$.

Next, we use the fact that in the cluster centers are selected at random, then

$$\mathbb{E}\left[\sum_h \|\hat{\tau}_{\hat{h}}\|^2\right] = m\mathbb{E}[\|\hat{\tau}_{\hat{h}}\|^2] = m\frac{1}{n}\sum_i \|Z_i\|^2 = \frac{m}{n}\sum_h\sum_{Z_j\in A_h}\|Z_j\|^2$$

$$\ge \frac{m}{n}\sum_h \|C(A_h)\|^2|A_h| = \frac{m}{n}\sum_h \frac{\|C(A_h)\|^2|A_h|^2}{(|A_h|-1)^2}\frac{(|A_h|-1)^2}{|A_h|}$$

where the inequality follows using the power-mean inequality. Using the optimal solution for τ_h described above, we can simplify the above term as $\frac{m}{n}\sum_h |A_h|^{-1}(|A_h|-1)^2\|\tau_h\|^2$. Finally, we can simplify this as

$$\frac{m}{n}\sum_h \|\tau_h\|^2\frac{(|A_h|-1)^2}{|A_h|} \ge \frac{m}{4n}\sum_h \|\tau_h\|^2|A_h| \ge \frac{m}{4n}\max_h |A_h|\sum_h \|\tau_h\|^2,$$

where the first inequality follows from $|A_h|\ge 2$. Since, $\max_h |A_h| = \mathcal{O}(n\log m/m)$, thus, we have, for some constant c,

$$\mathbb{E}[\|\sum_h \|\hat{\tau}_{\hat{h}}\|^2\|] \ge c\log m\sum_h \|\tau_h\|^2 \tag{16}$$

Since posterior density $\pi(\Lambda, \boldsymbol{h}|\boldsymbol{Z}) = k - \sum_i \|Z_i - \tau_{h_i}\|^2/2 - \sum_h \|\tau_h\|^2/2$ with a fixed constant k. Thus, combing both (15) and (16), we get,

$$\mathbb{E}[\pi(\hat{\Lambda}, \hat{\boldsymbol{h}})|\boldsymbol{Z})] \ge c\,\log m\,\pi(\Lambda, \boldsymbol{h}|\boldsymbol{Z}).$$

This proves the lemma since, we used $\hat{h} = \hat{h}[0]$, $\hat{\tau} = \hat{\tau}[0]$ and $h = h_{\text{MAP}}$, $\tau = \tau_{\text{MAP}}$.

B.2. Proof of Lemma 2. The full posterior distribution is

$$\pi(\boldsymbol{\omega}, \mathcal{A}, \Sigma) = g(\mathcal{A})g(\Sigma)\frac{\exp\{\sum_{b=1}^B y_b(x_b^*\mathcal{A}+\kappa_b) - \omega_b(x_b^*\mathcal{A}+\kappa_b)^2/2\}\Pi_{b=1}^B p(\omega_b)}{\Pi_{b=1}^B(1+\exp\{x_i^*\mathcal{A}+\kappa_i\})\int_0^\infty \exp\{-\omega(x_b^*\mathcal{A}+\kappa_b)^2/2\}\,p(\omega)\,d\omega}, \tag{17}$$

where $g(\mathcal{A})$ is a multivariate normal density with mean $\mathfrak{m}$ and variance $\mathcal{V}$, $g(\Sigma)$ is the inverse-Wishart prior with parameters Ψ_0, ν_0 and $p(\omega)$ is density of Polya-Gamma random variable $PG(1,0)$.

We need to simplify three conditional densities for the Gibbs sampling

1. $\pi(\omega|\mathcal{A}[t], \Sigma[t])$
2. $\pi(\mathcal{A}|\omega[t+1], \Sigma[t])$
3. $\pi(\Sigma|\omega[t+1], \mathcal{A}[t+1])$

We use the following theorem from Polson, Scott and Windle (2013) for the first two densities;

THEOREM B.1. *(Polson, Scott and Windle, 2013, Theorem 1) Let $p(\omega)$ denote the density of the random variable $\omega \sim PG(b, 0)$, $b > 0$. Then the following integral identity holds for all $a \in \mathbb{R}$:*

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{\gamma\psi} \int_0^{\infty} e^{-\omega\psi^2/2}\, p(\omega)\, d\omega\,, \tag{18}$$

where $\gamma = a - b/2$.

Moreover, the conditional distribution

$$p(\omega \mid \psi) = \frac{e^{-\omega\psi^2/2}\, p(\omega)}{\int_0^{\infty} e^{-\omega\psi^2/2}\, p(\omega)\, d\omega}\,,$$

where $(\omega \mid \psi) \sim PG(b, \psi)$.

Since all the ω_b's are independent, we can find the conditional densities for all of them separately. Hence,

$$\pi(\omega_b|\mathcal{A}, \Sigma) \propto \frac{e^{-\omega_b(x_b^*\mathcal{A}+\kappa_b)^2/2}\, p(\omega)}{\int_0^{\infty} e^{-\omega_b(x_b^*\mathcal{A}+\kappa_b)^2/2}\, p(\omega)\, d\omega}$$

since the other terms are fixed conditional on $\mathcal{A}$ and Σ. Since $p(\omega) \sim PG(1, 0)$ here, using Theorem B.1, $\pi(\omega_b|\mathcal{A}, \Sigma)$ is $\mathrm{PG}\big(1, x_b^*\mathcal{A} + \kappa_b\big)$. Hence, the first step of the Gibbs sampler is

$$(\omega_b[t+1]|\mathcal{A}[t], \Sigma[t]) \sim \mathrm{PG}\big(1, x_b^*\mathcal{A}[t] + \kappa_b\big).$$

Next for the conditional distribution of $\mathcal{A}$, we have

$$\begin{aligned}
\pi(\mathcal{A}|\Sigma, \omega) &\propto \Pi_{b=1}^{B} \left(\frac{\{\exp(x_b^*\mathcal{A} + \kappa_b)\}^{y_b}}{1 + \exp(x_b^*\mathcal{A} + \kappa_b)} \frac{e^{-\omega_b(x_b^*\mathcal{A}+\kappa_b)^2/2}\, p(\omega)}{\int_0^{\infty} e^{-\omega_b(x_b^*\mathcal{A}+\kappa_b)^2/2}\, p(\omega)\, d\omega} \right) g(\mathcal{A}) \\
&\propto \Pi_{b=1}^{B} \Bigg(\exp(\gamma_b(x_b^*\mathcal{A} + \kappa_b)) \int_0^{\infty} \exp(-\omega(x_b^*\mathcal{A} + \kappa_b)^2/2) p(\omega) d\omega \\
&\qquad \frac{e^{-\omega_b(x_b^*\mathcal{A}+\kappa_b)^2/2}\, p(\omega)}{\int_0^{\infty} e^{-\omega_b(x_b^*\mathcal{A}+\kappa_b)^2/2}\, p(\omega)\, d\omega} \Bigg) g(\mathcal{A}) \\
&= \Pi_{b=1}^{B} \left(\exp\left(\gamma_b(x_b^*\mathcal{A} + \kappa_b) - \omega_b(x_b^*\mathcal{A} + \kappa_b)^2/2 \right) \right) g(\mathcal{A}) \\
&\propto \Pi_{b=1}^{B} \left(\exp\left(-\frac{\omega_b}{2} \left((x_b^*\mathcal{A} + \kappa_b) - \frac{\gamma_b}{\omega_b} \right)^2 \right) \right) g(\mathcal{A}) \\
&= \exp\left((z - X_*\mathcal{A})^T \Omega (z - X_*\mathcal{A}) \right) g(\mathcal{A})
\end{aligned}$$

where $\gamma_b = y_b - 0.5$, $z_b = \gamma_b/\omega_b - \kappa_b$ and Ω is a diagonal matrix of ω_b's. The second equation follows directly from (18) of the theorem. Since $g(\mathcal{A})$ is multivariate Gaussian distribution with mean $\mathfrak{m}_0$ and variance $\mathcal{V}_0$, it follows directly that $\pi(\mathcal{A}|\Sigma, \omega)$ is Gaussian as well with variance $\mathcal{V}_\omega = (X_*^T \Omega X_* + \mathcal{V}_t^{-1})^{-1}$ and mean $\mathfrak{m}_\omega = \mathcal{V}_\omega \big\{ X_*^T \big(Y - \Omega\boldsymbol{\kappa} - 0.5 \big) + \mathcal{V}_t^{-1}\mathfrak{m}_0 \big\}$. The second Gibbs step is thus

$$\big(\mathcal{A}[t+1]|\Sigma[t], \boldsymbol{\omega}[t+1]\big) \sim \mathcal{N}(\mathfrak{m}[t+1], \mathcal{V}[t+1]),$$

where, $\mathcal{V}[t+1]) = (X_*^T\Omega[t+1]X_* + \mathcal{V}_t^{-1})^{-1}$ and $\Omega[t+1]$ is a diagonal matrix of $\omega_b[t+1]$s; the mean $\mathfrak{m}[t+1]$ is given by $\mathcal{V}[t+1]\{X_*^T(Y - \Omega[t+1]\boldsymbol{\kappa} - 0.5) + \mathcal{V}_t^{-1}\mathfrak{m}_0\}$. For the conditional distribution of Σ, we first observe that Σ only appears in $g(\mathcal{A})$ and $g(\Sigma)$ in the full posterior (17). Also, since Σ is independent of all the priors, Σ appears in $g(\mathcal{A})$ only through $a^{(l)}$'s. Since $a_h = (a_h^{(1)}, a_h^{(2)})$'s are independent, normally distributed with mean 0 and variance Σ, we can write

$$\pi(\Sigma|\omega, \mathcal{A}) \propto \Pi_{h=1}^m \phi_2(\Sigma^{-\frac{1}{2}}(a_h^{(1)}, a_h^{(2)})^T)g(\Sigma)$$

Since, the inverse Wishart prior $g(\Sigma)$ is a conjugate prior for random Gaussian samples, $\pi(\Sigma|\omega, \mathcal{A}) \sim \mathcal{W}^{-1}(\Psi_0 + \sum_{h=1}^m a_h\, a_h^T,\, \nu_0 + m)$. This gives rise to the last Gibbs step

$$\big(\Sigma[t+1]|\mathcal{A}[t+1]\big) \sim \mathcal{W}^{-1}\left(\Psi_0 + \sum_{h=1}^m a_h[t+1]\, a_h^T[t+1],\, \nu_0 + m\right).$$

B.3. Choice of Prior Hyper-parameters. In Section 4.1 and Section 4.2, we described the family of prior for the Metropolis-Hastings MCMC in Stage-I, and the Gibbs sampler for joint logistic regression in Stage-II. Here, we describe the choice of hyper-parameters for the prior distribution.

Stage-I hyper-parameters. We considered Gaussian priors $\delta^{(l)}$, $\tau_h^{(l)}$ and Inverse-Gamma priors on the variances $\left(\sigma^{(l)}\right)^2$:

$$\delta^{(l)} \sim \mathcal{N}(\mu_\delta^{(l)}, \Sigma_\delta^{(l)});\quad \tau_h^{(l)} \sim \mathcal{N}(\mu_\tau^{(l)}, \Sigma_\tau^{(l)})\quad (\sigma^{(l)})^2 \sim \text{IG}(\alpha^{(l)}, \beta^{(l)}).$$

Consider the model (1)-(2), if we ignore the clusters, the models are simplified. For $l = 1, 2, 3, 5$, the model turns into a linear model:

$$Z_i(l) = w_i^T\delta^{(l)} + \tilde{w}_i^T\tau^{(l)} + \sigma^{(l)}\epsilon_{il}.$$

While for the count variable, our model becomes an overdispersed Poisson model:

$$\log P(Z_i(4) = z_i) = z_i \log\lambda_i - \left(\sigma^{(4)}\right)^2\log(z_i!) - Z(\lambda, \left(\sigma^{(4)}\right)^2),$$

where, $\log\lambda_i = w_i^T\delta^{(4)} + \tilde{w}_i^T\tau_{h(i)}^{(4)}$ and $Z(\lambda, \nu) = \sum_{j=0}^\infty \lambda^j/(j!)^\nu$.

These models can be solved using existing likelihood maximization techniques in a frequentist setting. We use the means and the variances of the estimated $\delta^{(l)}$'s and $\tau^{(l)}$'s and set them as the hyper-parameters $\mu_\delta^{(l)}, \Sigma_\delta^{(l)}$'s and $\mu_\tau^{(l)}, \Sigma_\tau^{(l)}$'s respectively. For $\left(\sigma^{(l)}\right)^2$, we set the scale parameter as 1 and the shape parameter as $(\left(\hat{\sigma}^{(l)}\right)^2 + 1)/(\left(\hat{\sigma}^{(l)}\right)^2$. The mean of the inverse gamma prior is thus set as the marginal estimate of $\left(\sigma^{(l)}\right)^2$.

Stage-II hyper-parameters. Using the construction of the augmented response vector $\boldsymbol{y}$, the combined design matrix X_*, the clubbed dynamic and ad-stock variable κ and the parameter set $\mathcal{A}$ in Section 4.2, we have the logistic model as:

$$\text{logit}(y_b) = x_b^*\mathcal{A} + \kappa_b. \tag{19}$$

Though, in our model in Stage-II, the parameters have complex dependencies through a multivariate normal distribution, for the purpose of setting hyper-parameters, we solve (19) with a frequentist method. For $\alpha^{(l)}$'s, $\beta_h^{(l)}$'s, $\nu_{jk}^{(l)}$'s and $\gamma_{jk}^{(l)}$'s we set the estimates and the variance estimates as the prior hyper-parameters.

Let, $\hat{\Sigma}$ denotes the estimated covariance based on the estimates of $(a_h^{(1)}, a_h^{(2)})$'s we receive from the logistic regression. We would like to set the hyper-parameters of Σ such that the mean is $\hat{\Sigma}$. Since there are m pairs of $(a_h^{(1)}, a_h^{(2)})$'s, we set the degrees of freedom parameter as m. Next, using the fact that mean of $\mathcal{W}_p^{-1}(\Psi, \nu)$ is $\Psi/(\nu - p - 1)$. Thus, we choose the inverse-Wishart hyper-parameters as $\Psi_0 = (m-3)\hat{\Sigma}$ and $\nu_0 = m$ so that the mean of our prior is indeed $\hat{\Sigma}$.

APPENDIX C: FURTHER DETAILS ON PATH ALGORITHM PACURM

The distance between two clusters with centroids at v and $\tilde{v}$ is defined as absolute change in log-likelihood after merging them. The distance $\mathcal{D}(v,\tilde{v})$ is given by:

$$
\begin{aligned}
&-\sum_{l,i\in A_v}\frac{1}{2}(\sigma^{(l)})^{-2}\left(Z_i^{(l)}-w_i^T\delta^{(l)}-\tilde{w}_i^T\tau_v^{(l)}\right)^2-\sum_{l,i\in A_{\tilde{v}}}\frac{1}{2}(\sigma^{(l)})^{-2}\left(Z_i^{(l)}-w_i^T\delta^{(l)}-\tilde{w}_i^T\tau_{\tilde{v}}^{(l)}\right)^2\\
&+\sum_{l,i\in A_v\cup A_{\tilde{v}}}\frac{1}{2}(\sigma^{(l)})^{-2}\left(Z_i^{(l)}-w_i^T\delta^{(l)}-\tilde{w}_i^T\tau_{(v,\tilde{v})}^{(l)}\right)^2\\
&+\sum_{i\in A_v}(\sigma^{(4)})^{-2}\big\{Z_i^{(4)}\log\left(w_i^T\delta^{(l)}+\tilde{w}_i^T\tau_v^{(l)}\right)-\left(w_i^T\delta^{(l)}+\tilde{w}_i^T\tau_v^{(l)}\right)\big\}\\
&+\sum_{i\in A_{\tilde{v}}}(\sigma^{(4)})^{-2}\big\{Z_i^{(4)}\log\left(w_i^T\delta^{(l)}+\tilde{w}_i^T\tau_{\tilde{v}}^{(l)}\right)-\left(w_i^T\delta^{(l)}+\tilde{w}_i^T\tau_{\tilde{v}}^{(l)}\right)\big\}\\
&-\sum_{i\in A_v\cup A_{\tilde{v}}}(\sigma^{(4)})^{-2}\big\{Z_i^{(4)}\log\left(w_i^T\delta^{(l)}+\tilde{w}_i^T\tau_{(v,\tilde{v})}^{(l)}\right)-\left(w_i^T\delta^{(l)}+\tilde{w}_i^T\tau_{(v,\tilde{v})}^{(l)}\right)\big\}
\end{aligned}
$$

(20)

$$+g(\tau_v^{(l)})+g(\tau_{\tilde{v}}^{(l)})-g(\tau_{(v,\tilde{v})}^{(l)}).$$

C.1. Proof of Lemma 3.

Recall that the number of users is n, the number of coupons is K. p be the total number of covariates including global and cluster specific variables.

In Algorithm 1, T_1, L are the number of iterations for the outer and inner loops respectively, and, T_2 is the number of iterations for Algorithm 2.

Algorithm 1 For initialization, we run linear models assuming only one single cluster, hence we have to compute OLS solution for n observations and p covariates, the complexity for which is $\mathcal{O}(p^2n)$.

During the MHC procedure, for each iteration of the inner loop we first perturb all the mp parameters, one for each cluster. So we have T_1mp computations. Then for each consumer, we compute the likelihood for a fixed cluster which makes p computations. Thus for all consumers, clusters and iterations of the inner and outer loop, the complexity is of the order $(O)(T_1Lmnp)$.

The last stage in MHC is the acceptance/rejection stage with only constant computation and hence of $\mathcal{O}(T_1)$. The first stage complexity is thus $(O)(T_1Lmnp)$.

Algorithm 2 For Stage II, the total number of transactions is $\mathcal{O}(nK)$, i.e. number of times a coupon was opened as well as the number of recorded purchase. The initialization is training a logistic regression with $\mathcal{O}(nK)$ samples and mp covariates, the complexity of which is $\mathcal{O}(nKmp)$.

Next for each iteration of the Gibb's sampling, we need to sample from the three conditional densities based on our proposed Gibbs sampler. The expensive step in the process is the computation of the conditional mean of $(\mathcal{A}[t+1]|\Sigma[t],\boldsymbol{\omega}[t+1])$, $\mathfrak{m}[t+1]$. Since inverting an $n\times n$ matrix has order $\mathcal{O}(n^3)$ and matrix multiplication of an $m\times n$, $n\times p$ matrix is $\mathcal{O}(mnp)$, hence computing $\mathfrak{m}[t+1]$ is of the order $\mathcal{O}((mp)^3+nK(mp)^2)$. In our fixed dimension problem $nK>mp$, hence the computation complexity of Stage II is $\mathcal{O}(T_2nLm^2p^2)$.

Finally, consider the path algorithm. Consider the complexity of decreasing one cluster from h to $h-1$. The computationally expensive step here is solving (13), which has order $\mathcal{O}(np^2)$ and we have to perform this step $h-1$ times. Thus, to move from h to $h-1$ clusters, the complexity is $\mathcal{O}(hnp^2)$ and the complexity of starting at m clusters and aggregating till

two clusters is $\mathcal{O}(m^2np^2)$. This is dwarfed by the complexity of Stage II for m clusters. We summarize the computational complexity result below.
Path Algorithm. Finally we also have the path algorithm. We use the same notations for the size of the parameters from the previous two stages.

The first task here was to estimate the merged cluster parameter. The computationally expensive step here is solving (13), which has order $\mathcal{O}(np^2 + p^3 + np) = \mathcal{O}(np^2)$.

The second task was to update the logistic regression parameters. Computing the likelihood has complexity $\mathcal{O}(np)$ and the asymptotic update computation has complexity $\mathcal{O}(p^3)$. So, the updates to combine any two cluster is $\mathcal{O}(np^2)$.

When we start with m clusters, and we have to perform this step $\binom{m}{2}$ times which is $\mathcal{O}(m^2)$. Next, we consider the complexity of decreasing one cluster from h to $h-1$. We only need to compute the parameters for only the newly formed cluster with the other clusters. So we need to combine clusters $h-1$ times. The complexity for starting at m clusters and aggregate till 2 cluster is $\mathcal{O}(m^2)$. Thus, the complexity of PACURM is $\mathcal{O}(m^2np^2)$.

APPENDIX D: FURTHER RESULTS ON COUPON RESPONSE ANALYSIS

In this section, we present the dendogram mentioned in Sec. 6.4 of the main paper as well as present all the tables referenced in the Section 6 of the main paper. We provide a list of the tables, arranged in the order they are referenced in the main paper.

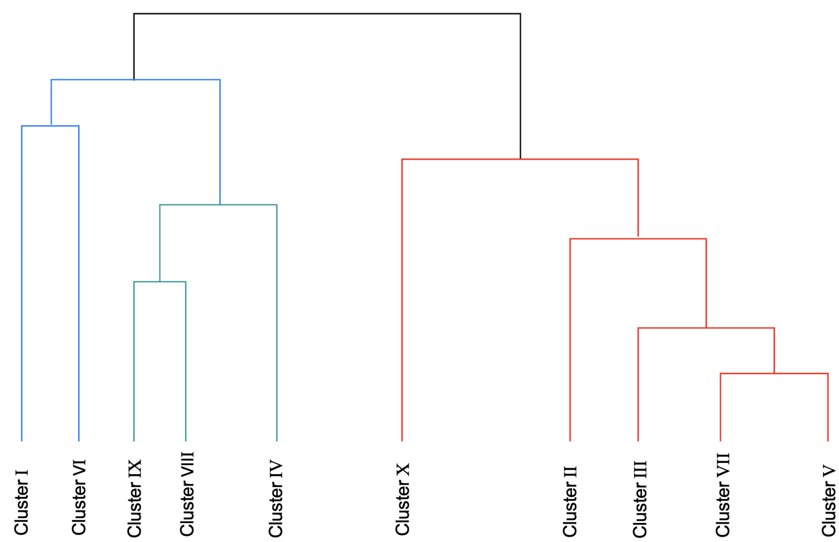


Fig 7: Dendogram for the Path Algorithm with 10 clusters.

List of Tables.

Component	Mean Estimate	Lower CI	Upper CI
30% off	0.01	-0.037	0.061
40% off	0.023	-0.022	0.063
50% off	0.151	0.089	0.216
Product	0.111	0.049	0.167
Clearance	-0.02	-0.087	0.046
Mystery	-0.029	-0.079	0.021
Free Gift	-0.016	-0.067	0.031
Free Shipping	-0.004	-0.057	0.051

Table A1: CURM open stage messaging component effects with 95% credible interval

Component	Mean Estimate	Lower CI	Upper CI
30% off	0.153	-0.041	0.381
40% off	-0.036	-0.195	0.138
50% off	-0.107	-0.387	0.221
Product	-0.086	-0.325	0.151
Clearance	-0.113	-0.381	0.136
Mystery	0.051	-0.207	0.297
Free Gift	0.044	-0.193	0.272
Free Shipping	-0.041	-0.316	0.175

Table A2: CURM purchase stage messaging component effects with 95% credible interval

Cluster & Component	Mean Estimate	Lower CI	Upper CI
Cluster 2 30% off	-0.132	-0.205	-0.054
Cluster 3 30% off	0.005	-0.047	0.061
Cluster 4 30% off	0.068	0.012	0.116
Cluster 5 30% off	0.098	0.004	0.198
Cluster 2 40% off	-0.106	-0.174	-0.045
Cluster 3 40% off	0.039	-0.010	0.089
Cluster 4 40% off	0.072	0.033	0.106
Cluster 5 40% off	0.089	0.002	0.167
Cluster 2 50% off	0.135	0.034	0.245
Cluster 3 50% off	0.142	0.061	0.213
Cluster 4 50% off	0.083	0.012	0.159
Cluster 5 50% off	0.244	0.142	0.385
Cluster 2 Product	0.022	-0.060	0.099
Cluster 3 Product	0.116	0.048	0.178
Cluster 4 Product	0.148	0.089	0.213
Cluster 5 Product	0.157	0.053	0.269
Cluster 2 Clearance	-0.111	-0.210	0.005
Cluster 3 Clearance	-0.037	-0.121	0.043
Cluster 4 Clearance	0.108	0.043	0.183
Cluster 5 Clearance	-0.042	-0.173	0.088
Cluster 2 Mystery	-0.145	-0.210	-0.067
Cluster 3 Mystery	-0.010	-0.074	0.055
Cluster 4 Mystery	0.103	0.044	0.160
Cluster 5 Mystery	-0.064	-0.161	0.028
Cluster 2 Free Gift	-0.181	-0.258	-0.112
Cluster 3 Free Gift	0.009	-0.050	0.066
Cluster 4 Free Gift	0.058	0.006	0.110
Cluster 5 Free Gift	0.053	-0.040	0.147
Cluster 2 Free Shipping	-0.046	-0.135	0.043
Cluster 3 Free Shipping	-0.007	-0.071	0.058
Cluster 4 Free Shipping	-0.033	-0.095	0.023
Cluster 5 Free Shipping	0.070	-0.028	0.166

Table A3: CURM open stage messaging component and cluster cross effects with 95% credible interval

Cluster & Component	Mean Estimate	Lower CI	Upper CI
Cluster 2 30% off	0.047	-0.403	0.392
Cluster 3 30% off	-0.016	-0.222	0.215
Cluster 4 30% off	0.332	0.162	0.552
Cluster 5 30% off	0.249	-0.076	0.600
Cluster 2 40% off	-0.256	-0.557	0.075
Cluster 3 40% off	-0.040	-0.225	0.144
Cluster 4 40% off	0.148	-0.005	0.287
Cluster 5 40% off	0.003	-0.318	0.351
Cluster 2 50% off	-0.426	-0.957	0.108
Cluster 3 50% off	-0.123	-0.407	0.184
Cluster 4 50% off	-0.058	-0.334	0.245
Cluster 5 50% off	0.179	-0.388	0.731
Cluster 2 Product	-0.383	-0.904	0.111
Cluster 3 Product	0.004	-0.240	0.299
Cluster 4 Product	0.030	-0.194	0.264
Cluster 5 Product	0.003	-0.533	0.500
Cluster 2 Clearance	-0.284	-0.872	0.227
Cluster 3 Clearance	-0.062	-0.334	0.234
Cluster 4 Clearance	0.101	-0.107	0.391
Cluster 5 Clearance	-0.206	-0.643	0.219
Cluster 2 Mystery	-0.182	-0.658	0.239
Cluster 3 Mystery	0.103	-0.210	0.374
Cluster 4 Mystery	0.214	-0.011	0.440
Cluster 5 Mystery	0.068	-0.280	0.439
Cluster 2 Free Gift	-0.229	-0.626	0.141
Cluster 3 Free Gift	0.033	-0.177	0.262
Cluster 4 Free Gift	0.231	0.021	0.428
Cluster 5 Free Gift	0.142	-0.180	0.486
Cluster 2 Free Shipping	-0.348	-0.836	0.094
Cluster 3 Free Shipping	-0.087	-0.353	0.175
Cluster 4 Free Shipping	0.151	-0.101	0.374
Cluster 5 Free Shipping	0.120	-0.296	0.544

Table A4: CURM purchase stage messaging component and cluster cross effects with 95% credible interval

Coefficients	Open Stage			Purchase Stage		
	Posterior Mean	Lower CI	Upper CI	Posterior Mean	Lower CI	Upper CI
Intercept	-0.807	-0.840	-0.776	-4.835	-5.011	-4.676
Age	0.094	0.089	0.098	0.050	0.032	0.072
Income: 15K - 25K	-0.005	-0.025	0.013	-0.025	-0.096	0.070
Income: 25K - 35K	0.014	-0.007	0.036	-0.021	-0.112	0.047
Income: 35K - 50K	0.022	0.006	0.037	0.000	-0.081	0.076
Income: 50K - 75K	0.015	0.002	0.029	-0.007	-0.090	0.069
Income: 75K - 100K	0.044	0.026	0.058	0.004	-0.075	0.076
Income: 100K - 120K	0.070	0.055	0.088	-0.015	-0.099	0.065
Income: 120K - 150K	0.073	0.054	0.098	0.015	-0.077	0.108
Income: 150K plus	0.098	0.074	0.118	-0.003	-0.117	0.098
30% off	-0.182	-0.198	-0.159	0.515	0.443	0.599
40% off	0.038	0.026	0.055	0.565	0.493	0.630
50% off	0.039	0.027	0.052	-0.086	-0.128	-0.031
Product	0.178	0.166	0.190	0.612	0.560	0.668
Clearance	0.451	0.432	0.474	1.306	1.196	1.413
Mystery	-0.033	-0.054	-0.012	1.434	1.333	1.556
Free Gift	0.025	0.009	0.043	0.737	0.653	0.813
Free Shipping	0.103	0.091	0.117	0.047	0.009	0.097
Free Returns	-0.425	-0.441	-0.410	-	-	-

Table A5: URM: coefficients for open and purchase stages with 95% credible intervals

Coefficients	Open Stage			Purchase Stage		
	Posterior Mean	Lower CI	Upper CI	Posterior Mean	Lower CI	Upper CI
Intercept	-0.721	-0.796	-0.643	-3.831	-3.955	-3.724
Age	0.095	0.091	0.100	0.044	0.020	0.066
Income: 15K - 25K	0.239	0.218	0.262	-0.201	-0.308	-0.129
Income: 25K - 35K	0.009	-0.009	0.030	0.015	-0.061	0.129
Income: 35K - 50K	0.145	0.128	0.161	0.010	-0.095	0.099
Income: 50K - 75K	0.007	-0.009	0.023	0.074	0.014	0.145
Income: 75K - 100K	0.033	0.016	0.052	-0.266	-0.324	-0.206
Income: 100K - 120K	0.082	0.067	0.098	0.032	-0.066	0.106
Income: 120K - 150K	0.080	0.050	0.105	-0.199	-0.286	-0.091
Income: 150K plus	0.138	0.121	0.160	-0.067	-0.157	0.013
Cluster 2	0.109	-0.008	0.217	-1.305	-1.626	-0.941
Cluster 3	-0.247	-0.313	-0.177	-0.968	-1.168	-0.741
Cluster 4	-0.081	-0.168	0.041	-0.687	-0.876	-0.531
Cluster 5	-0.456	-0.563	-0.366	-0.748	-1.093	-0.338
Cluster 2 30% off	-0.241	-0.296	-0.196	0.509	0.187	0.831
Cluster 3 30% off	-0.216	-0.253	-0.187	0.471	0.280	0.642
Cluster 4 30% off	-0.177	-0.206	-0.146	0.630	0.514	0.737
Cluster 5 30% off	-0.225	-0.306	-0.159	0.645	0.376	0.925
Cluster 2 40% off	-0.030	-0.071	0.016	0.413	0.075	0.672
Cluster 3 40% off	0.024	-0.007	0.051	0.564	0.397	0.711
Cluster 4 40% off	0.056	0.027	0.082	0.588	0.476	0.691
Cluster 5 40% off	-0.017	-0.088	0.048	0.467	0.167	0.752
Cluster 2 50% off	0.052	0.016	0.092	-0.113	-0.282	0.104
Cluster 3 50% off	0.047	0.022	0.074	-0.174	-0.294	-0.062
Cluster 4 50% off	0.037	0.015	0.056	-0.091	-0.158	-0.008
Cluster 5 50% off	0.071	0.012	0.125	0.009	-0.261	0.219
Cluster 2 Product	0.129	0.099	0.171	0.408	0.189	0.650
Cluster 3 Product	0.192	0.164	0.220	0.607	0.490	0.713
Cluster 4 Product	0.211	0.193	0.229	0.635	0.534	0.754
Cluster 5 Product	0.202	0.152	0.255	0.517	0.238	0.821
Cluster 2 Clearance	0.390	0.310	0.479	1.062	0.692	1.466
Cluster 3 Clearance	0.382	0.333	0.429	1.296	1.084	1.503
Cluster 4 Clearance	0.507	0.472	0.544	1.314	1.166	1.497
Cluster 5 Clearance	0.373	0.284	0.460	1.069	0.710	1.472
Cluster 2 Mystery	-0.126	-0.196	-0.051	1.189	0.808	1.587
Cluster 3 Mystery	-0.052	-0.103	-0.011	1.537	1.339	1.751
Cluster 4 Mystery	0.006	-0.029	0.039	1.454	1.298	1.625
Cluster 5 Mystery	-0.145	-0.245	-0.054	1.414	1.018	1.739
Cluster 2 Free Gift	-0.091	-0.145	-0.028	0.433	0.026	0.806
Cluster 3 Free Gift	0.009	-0.035	0.040	0.785	0.597	0.967
Cluster 4 Free Gift	0.047	0.013	0.078	0.789	0.668	0.906
Cluster 5 Free Gift	-0.024	-0.114	0.044	0.818	0.521	1.079
Cluster 2 Free Shipping	0.061	0.014	0.096	-0.146	-0.393	0.087
Cluster 3 Free Shipping	0.083	0.065	0.104	-0.033	-0.165	0.085
Cluster 4 Free Shipping	0.089	0.068	0.113	0.109	0.017	0.179
Cluster 5 Free Shipping	0.102	0.047	0.147	0.112	-0.112	0.294
Cluster 2 Free Returns	-0.429	-0.472	-0.381	0.865	0.542	1.160
Cluster 3 Free Returns	-0.438	-0.467	-0.410	0.729	0.549	0.882
Cluster 4 Free Returns	-0.427	-0.448	-0.401	0.556	0.453	0.663
Cluster 5 Free Returns	-0.518	-0.610	-0.433	-	-	-

Table A6: CURM.0.1: coefficients for open and purchase stages with 95% credible intervals

Coefficients	Open Stage			Purchase Stage		
	Posterior Mean	Lower CI	Upper CI	Posterior Mean	Lower CI	Upper CI
Intercept	-0.331	-0.466	-0.185	-3.861	-3.982	-3.737
Age	0.051	0.046	0.057	0.087	0.067	0.105
Income: 15K - 25K	0.163	0.137	0.186	-0.103	-0.194	-0.001
Income: 25K - 35K	0.030	0.009	0.047	0.057	-0.030	0.138
Income: 35K - 50K	0.112	0.090	0.134	0.081	-0.001	0.162
Income: 50K - 75K	0.030	0.011	0.047	0.115	0.057	0.179
Income: 75K - 100K	0.043	0.026	0.064	-0.203	-0.261	-0.137
Income: 100K - 120K	0.074	0.056	0.096	0.099	0.023	0.173
Income: 120K - 150K	0.067	0.039	0.095	-0.127	-0.213	-0.028
Income: 150K plus	0.107	0.085	0.126	0.016	-0.061	0.075
Cluster 2	0.061	-0.161	0.239	-1.683	-2.019	-1.304
Cluster 3	-0.154	-0.319	0.009	-0.933	-1.251	-0.563
Cluster 4	-0.209	-0.395	-0.055	-0.677	-0.901	-0.468
Cluster 5	-0.278	-0.528	-0.079	-0.528	-0.977	0.006
Cluster 2 30% off	-0.252	-0.326	-0.191	0.759	0.429	1.039
Cluster 3 30% off	-0.165	-0.198	-0.125	0.469	0.315	0.619
Cluster 4 30% off	-0.087	-0.135	-0.051	0.699	0.571	0.829
Cluster 5 30% off	-0.109	-0.199	-0.010	0.499	0.175	0.771
Cluster 2 40% off	0.083	0.033	0.127	0.577	0.273	0.908
Cluster 3 40% off	0.224	0.195	0.257	0.504	0.380	0.638
Cluster 4 40% off	0.305	0.265	0.334	0.574	0.445	0.696
Cluster 5 40% off	0.234	0.173	0.313	0.331	0.033	0.579
Cluster 2 50% off	0.242	0.201	0.277	-0.109	-0.297	0.106
Cluster 3 50% off	0.289	0.262	0.320	-0.170	-0.258	-0.084
Cluster 4 50% off	0.257	0.237	0.281	-0.099	-0.174	-0.024
Cluster 5 50% off	0.342	0.285	0.398	-0.040	-0.224	0.220
Cluster 2 Product	0.117	0.067	0.152	0.614	0.420	0.780
Cluster 3 Product	0.184	0.160	0.216	0.572	0.421	0.687
Cluster 4 Product	0.238	0.217	0.262	0.560	0.473	0.657
Cluster 5 Product	0.177	0.115	0.237	0.351	0.084	0.581
Cluster 2 Clearance	0.334	0.262	0.429	1.410	1.066	1.738
Cluster 3 Clearance	0.353	0.297	0.422	1.256	1.027	1.485
Cluster 4 Clearance	0.553	0.504	0.615	1.270	1.128	1.422
Cluster 5 Clearance	0.302	0.190	0.415	0.844	0.340	1.318
Cluster 2 Mystery	-0.288	-0.361	-0.214	1.511	1.189	1.852
Cluster 3 Mystery	-0.221	-0.260	-0.168	1.443	1.165	1.662
Cluster 4 Mystery	-0.063	-0.106	-0.018	1.401	1.250	1.557
Cluster 5 Mystery	-0.315	-0.426	-0.209	1.133	0.644	1.461
Cluster 2 Free Gift	-0.186	-0.251	-0.135	0.771	0.438	1.091
Cluster 3 Free Gift	-0.037	-0.075	0.003	0.756	0.561	0.908
Cluster 4 Free Gift	0.053	0.013	0.095	0.809	0.687	0.959
Cluster 5 Free Gift	-0.021	-0.101	0.067	0.643	0.295	0.887
Cluster 2 Free Shipping	0.034	-0.005	0.075	-0.087	-0.315	0.170
Cluster 3 Free Shipping	0.068	0.042	0.097	-0.014	-0.117	0.092
Cluster 4 Free Shipping	0.069	0.042	0.092	0.146	0.074	0.232
Cluster 5 Free Shipping	0.099	0.045	0.157	0.070	-0.149	0.282
Cluster 2 Free Returns	-0.648	-0.703	-0.592	1.109	0.770	1.444
Cluster 3 Free Returns	-0.656	-0.697	-0.621	0.789	0.616	0.978
Cluster 4 Free Returns	-0.604	-0.630	-0.575	0.702	0.586	0.827
Cluster 5 Free Returns	-0.696	-0.779	-0.615	0.608	0.312	0.930
Carry-over Effect (θ_1)	-0.663	-0.669	-0.658	-0.225	-0.204	-0.248
Purchase Engagement (θ_2)	-0.300	-0.314	-0.288	-0.469	-0.424	-0.517

Table A7: CURM.0.2: coefficients for open and purchase stages with 95% credible intervals

Coefficients	Open Stage			Purchase Stage		
	Posterior Mean	Lower CI	Upper CI	Posterior Mean	Lower CI	Upper CI
Intercept	-0.690	-0.844	-0.525	-4.373	-4.519	-4.206
Age	0.053	0.048	0.058	0.087	0.070	0.108
Income: 15K - 25K	0.132	0.112	0.158	-0.109	-0.224	-0.008
Income: 25K - 35K	0.015	-0.010	0.035	0.050	-0.035	0.127
Income: 35K - 50K	0.091	0.069	0.110	0.079	-0.001	0.146
Income: 50K - 75K	0.022	0.001	0.039	0.110	0.027	0.195
Income: 75K - 100K	0.017	-0.006	0.035	-0.197	-0.280	-0.121
Income: 100K - 120K	0.059	0.037	0.076	0.093	-0.030	0.188
Income: 120K - 150K	0.045	0.018	0.072	-0.129	-0.226	-0.051
Income: 150K plus	0.083	0.058	0.106	0.029	-0.062	0.104
Cluster 2	0.337	0.032	0.706	-0.549	-1.140	0.122
Cluster 3	-0.068	-0.297	0.156	0.122	-0.333	0.544
Cluster 4	-0.095	-0.369	0.178	0.205	-0.237	0.581
Cluster 5	-0.211	-0.438	-0.008	0.405	-0.360	1.153
Campaign Residuals 2 (γ_2)	0.394	0.374	0.417	1.295	1.203	1.381
Campaign Residuals 3 (γ_3)	0.811	0.768	0.860	0.706	0.536	0.849
Campaign Residuals 4 (γ_4)	1.149	1.113	1.185	-0.156	-0.307	-0.011
Campaign Residuals 5 (γ_5)	0.437	0.396	0.476	0.287	0.122	0.457
Campaign Residuals 6 (γ_6)	0.185	0.155	0.220	0.685	0.489	0.829
Campaign Residuals 7 (γ_7)	-0.191	-0.232	-0.151	-0.156	-0.335	0.036
Campaign Residuals 8 (γ_8)	-0.016	-0.052	0.024	0.453	0.239	0.640
Campaign Residuals 9 (γ_9)	-0.214	-0.247	-0.185	0.525	0.371	0.684
Campaign Residuals 10 (γ_{10})	0.246	0.190	0.300	0.723	0.483	1.014
Campaign Residuals 11 (γ_{11})	0.273	0.240	0.316	0.289	0.099	0.445
Campaign Residuals 12 (γ_{12})	-0.006	-0.043	0.036	0.612	0.462	0.769
Campaign Residuals 13 (γ_{13})	0.922	0.888	0.960	0.388	0.232	0.532
Campaign Residuals 14 (γ_{14})	-0.220	-0.255	-0.182	1.053	0.850	1.215
Campaign Residuals 15 (γ_{15})	-0.174	-0.217	-0.127	-0.411	-0.666	-0.209
Campaign Residuals 16 (γ_{16})	0.330	0.297	0.368	0.431	0.247	0.584
Campaign Residuals 17 (γ_{17})	0.733	0.695	0.773	0.047	-0.078	0.194
Campaign Residuals 18 (γ_{18})	0.388	0.351	0.436	0.225	0.018	0.440
Campaign Residuals 19 (γ_{19})	0.300	0.268	0.332	-1.548	-1.759	-1.295
Campaign Residuals 20 (γ_{20})	-0.012	-0.044	0.022	0.735	0.566	0.877
Campaign Residuals 21 (γ_{21})	-0.133	-0.164	-0.094	0.740	0.557	0.897
Campaign Residuals 22 (γ_{22})	0.689	0.651	0.731	0.399	0.211	0.539
Campaign Residuals 23 (γ_{23})	-0.257	-0.292	-0.211	0.192	0.002	0.366
Campaign Residuals 24 (γ_{24})	0.604	0.561	0.663	-2.011	-2.249	-1.668
Campaign Residuals 25 (γ_{25})	0.734	0.702	0.765	-0.028	-0.229	0.148
Cluster 2 30% off	-0.210	-0.282	-0.143	0.354	-0.101	0.742
Cluster 3 30% off	-0.053	-0.122	0.002	-0.014	-0.278	0.199
Cluster 4 30% off	0.038	-0.021	0.086	0.315	0.130	0.529
Cluster 5 30% off	-0.010	-0.106	0.079	0.111	-0.272	0.468
Cluster 2 40% off	-0.185	-0.252	-0.122	0.051	-0.373	0.430
Cluster 3 40% off	-0.019	-0.070	0.029	-0.038	-0.212	0.133
Cluster 4 40% off	0.041	-0.004	0.082	0.131	-0.052	0.305
Cluster 5 40% off	-0.019	-0.111	0.065	-0.136	-0.397	0.226
Cluster 2 50% off	0.057	0.011	0.102	-0.119	-0.358	0.104
Cluster 3 50% off	0.084	0.047	0.111	-0.121	-0.264	0.003
Cluster 4 50% off	0.052	0.016	0.082	-0.075	-0.224	0.070
Cluster 5 50% off	0.136	0.086	0.212	0.041	-0.267	0.297
Cluster 2 Product	-0.057	-0.112	0.001	-0.076	-0.397	0.196
Cluster 3 Product	0.058	0.020	0.095	0.006	-0.184	0.177
Cluster 4 Product	0.117	0.080	0.150	0.013	-0.134	0.177
Cluster 5 Product	0.049	-0.018	0.114	-0.135	-0.398	0.164

Coefficients	Open Stage			Purchase Stage		
	Posterior Mean	Lower CI	Upper CI	Posterior Mean	Lower CI	Upper CI
Cluster 2 Clearance	-0.189	-0.272	-0.082	0.023	-0.501	0.503
Cluster 3 Clearance	-0.095	-0.171	-0.018	-0.060	-0.347	0.233
Cluster 4 Clearance	0.077	0.007	0.139	0.085	-0.145	0.344
Cluster 5 Clearance	-0.149	-0.290	-0.035	-0.344	-0.806	0.140
Cluster 2 Mystery	-0.223	-0.315	-0.126	0.125	-0.427	0.677
Cluster 3 Mystery	-0.068	-0.134	-0.002	0.105	-0.160	0.365
Cluster 4 Mystery	0.072	0.011	0.122	0.197	-0.109	0.482
Cluster 5 Mystery	-0.172	-0.308	-0.067	-0.070	-0.490	0.389
Cluster 2 Free Gift	-0.259	-0.323	-0.185	0.078	-0.479	0.569
Cluster 3 Free Gift	-0.050	-0.110	0.007	0.035	-0.171	0.243
Cluster 4 Free Gift	0.027	-0.024	0.082	0.214	-0.019	0.464
Cluster 5 Free Gift	-0.055	-0.152	0.036	0.003	-0.326	0.432
Cluster 2 Free Shipping	-0.124	-0.178	-0.073	-0.041	-0.369	0.243
Cluster 3 Free Shipping	-0.065	-0.103	-0.032	-0.085	-0.280	0.072
Cluster 4 Free Shipping	-0.064	-0.101	-0.027	0.134	-0.017	0.246
Cluster 5 Free Shipping	-0.038	-0.101	0.033	-0.018	-0.249	0.192
Cluster 2 Free Returns	-0.078	-0.147	-0.011	0.307	-0.092	0.722
Cluster 3 Free Returns	-0.058	-0.110	-0.013	0.002	-0.227	0.181
Cluster 4 Free Returns	-0.031	-0.085	0.024	-0.017	-0.228	0.170
Cluster 5 Free Returns	-0.108	-0.189	-0.030	-0.138	-0.517	0.277
Carry-over Effect (θ_1)	-0.583	-0.592	-0.573	-0.264	-0.227	-0.305
Purchase Engagement (θ_2)	-0.312	-0.324	-0.300	-0.519	-0.472	-0.564

Table A8: CURM : coefficients for open and purchase stages with 95% credible intervals