Transfer Learning for Targeted Marketing: A Bayesian Matrix Factorization Approach

Marat Ibragimov[*], Duncan Simester[†], and Artem Timoshenko[‡]

[*]Goizueta Business School, Emory University

[†]Sloan School of Management, Massachusetts Institute of Technology

[‡]Kellogg School of Management, Northwestern University

June 5, 2025

Firms can enhance the profitability of marketing campaigns by targeting different marketing actions to different customer segments. However, when designing targeting policies, firms typically use training data that closely resembles the focal campaign, and this data is often scarce. We propose a method to augment focal training data with information from different marketing campaigns, even when the campaigns differ in content, audience, and timing. Our approach accounts for campaign and customer differences and explicitly incorporates uncertainty in the measurement of treatment effects using a Bayesian probabilistic matrix factorization framework. Scalability and computational efficiency are achieved through a combination of closed-form conditional distributions and gradient-based MCMC sampling. We validate the method using simulated data and a sequence of field experiments at a luxury fashion retailer, and demonstrate substantial improvements in predictive accuracy and customer response. By leveraging information across campaigns, the proposed approach amplifies the value of firms' marketing data and offers a scalable and robust solution for designing targeting policies.

1 Introduction

Many firms want to increase the profitability of their marketing campaigns by targeting different customers with different marketing actions. Retailers send personalized coupons based on customers' purchase histories; social networks show advertisements that vary depending on users' profiles or search behavior; financial services firms offer different products according to customers' investment records. To maximize the incremental effects, these actions are often directed by targeting (uplift) models that predict customers' responsiveness to different marketing actions.

Targeting policies are typically trained using data that closely matches the focal campaign, such as experiments with a similar campaign in a previous year or a small pilot experiment. The more similar the customers and marketing actions in the training data are to the focal campaign, the more accurately the firm can predict the responsiveness to the focal campaign. However, training data that closely matches the focal campaign is often scarce. This scarcity can be costly, as the expected performance of a targeting policy depends upon the size of the training data. We investigate how to augment training data using data from different marketing campaigns that do not closely match the focal campaign.

For example, a luxury fashion retailer distributes a large fashion catalog to a selected group of customers at the start of the Christmas season (the "focal campaign"). It chooses which customers to mail to using experimental data collected when mailing the previous year's Christmas catalog (the "focal data"). We propose supplementing the limited experimental data available from the previous year's Christmas catalog with information from experiments conducted with the retailer's other marketing campaigns. These could include Mother's Day promotions and back-to-school offers (the "source campaigns").

Our paper focuses on two challenges in combining information across marketing campaigns. The first challenge recognizes that marketing campaigns often involve different mar-

keting actions. The marketing actions could have different value propositions: the focal campaign may offer price promotions, while the source campaigns could communicate information about product features. The marketing channels may differ: the focal campaign may use direct mail, whereas the source campaigns used digital communications. The timing of the campaigns could differ, with the focal campaign distributed at Christmas and the source campaigns distributed at other times of the year. There could also be differences in creative content, such as wording, sound, color, graphics, or other visual content. A complicating factor is that these differences are not always accurately documented. For example, the luxury fashion retailer in our empirical application carefully documents the timing and mailing cost of each catalog, but does not document the products and prices featured in each catalog.

The second challenge involves the measurement of treatment effects. The responsiveness to each campaign is measured with uncertainty, and the amount of uncertainty varies across source campaigns. Different marketing campaigns produce different variation in customer responses, and different campaigns also have different sample sizes. Transferring information about treatment effects across marketing campaigns will be more valuable if we are able to account for the uncertainty in the measurement of treatment effects.

We propose an approach that combines the information from source and focal campaigns to improve the firm's targeting policy for the focal campaign. The proposed approach takes into account the differences between the customers and the campaigns, and explicitly incorporates uncertainty in the measurement of treatment effects. Our approach extends the Bayesian probabilistic matrix factorization (Salakhutdinov and Mnih, 2008). Intuitively, the approach represents customer segments and marketing campaigns using low-dimensional embeddings, and models treatment effects as a function of these embeddings. Customer segments with similar embeddings respond similarly to different marketing actions. Campaigns with similar embeddings yield similar patterns of treatment effects across segments. Our extension incorporates information about uncertainty in the treatment effect measures to

improve the embeddings and treatment effect estimates for the focal campaign.

As inputs, our approach requires an existing customer segmentation, and segment-level estimates of treatment effects (with corresponding uncertainty) for the focal and source campaigns. The segmentation and treatment effect measures replace the need to maintain detailed records of the design characteristics of the source campaigns and customer segments.[1] Instead, the model uses joint variation in treatment effects across customer segments to identify which source campaigns are most similar to the focal campaign, and which customer segments respond similarly to the firm's marketing actions. The richness of this information allows the model to learn these similarities quickly and accurately.

Different marketing campaigns often target different audiences. For example, a luxury department store may only send cosmetics promotions to customers who purchased cosmetics in the prior 12 months, and only send menswear promotions to customers who recently purchased from the menswear department. Our model can accommodate eligibility criteria that vary across campaigns, and learns from source campaigns even when the eligibility criteria are different than in the focal campaign.

Our approach targets customers at the segment level; all customers within the same segment receive the same action, but customers in different segments can receive different actions. The proposed solution scales to a large number of segments, and could potentially be applied to individual-level targeting.[2] However, customer-level applications would require measures of treatment effects and uncertainty at the individual level.

We evaluate the performance of the model using a combination of simulated data, and data from a sequence of field experiments conducted by a luxury fashion retailer. We demonstrate that information from source campaigns can greatly improve the performance

[1] We provide an extension that incorporates observable campaign design features and customer characteristics, but the model generally does not require them.

[2] The scalability of our method is achieved because, with one exception, the conditional distributions of the unobserved variables are derived as closed-form expressions. For this exception, we use gradient-based MCMC sampling.

of targeting policies that rank customer segments according to their responsiveness to the focal campaign. Improving marketing decisions by leveraging information in past campaigns amplifies the value of the information in those past campaigns. The method we propose will be most valuable for firms that have experimented frequently on their past marketing campaigns, and carefully maintained a catalog of their past experiments.

The paper continues in Section 2 with a review of the related literature. In Section 3 we formally state the business problem and introduce the Probabilistic Matrix Factorization model. We illustrate the properties of the model using synthetic data in Section 4 and using field data in Section 5. The paper concludes in Section 6.

2 Related Literature

Our findings contribute to a growing literature in marketing focusing on the training of targeting (uplift) policies. Recent research has studied: the design and evaluation of targeting methods (Simester et al., 2020; Hitsch et al., 2024), welfare implications of personalized pricing (Dubé and Misra, 2023), explainable targeting policies and unintended bias (Ascarza and Israeli, 2022; Zhang, 2024). The business setting is similar to the problem studied in Simester et al. (2025), who study the size of the focal experiment required to train and certify a targeting policy. Like this study, their training method assumes an existing customer segmentation and combines Bayesian priors with information from an initial experiment conducted for the focal campaign. However, unlike this study, they do not use information from multiple source campaigns. We propose methods to supplement the focal data with data from past campaigns, so that firms may be able to greatly reduce the amount of focal data required to train a targeting policy.

In concurrent research, Ellickson et al. (2023) and Huang et al. (2024) also propose using data from source campaigns to target a new focal campaign. Ellickson et al. (2023)

study email promotions and decompose the impact of different keywords and different types of promotional offers. Huang et al. (2024) model the incremental effects of coupons using observable characteristics of consumer packaged goods (CPG) brands. There are two important differences between their studies and our study. First, they use observable design features of marketing campaigns, and identify similarities between campaigns from these features. In contrast, our approach identifies campaign similarities from the variation in treatment effects across customer segments and campaigns. Our method can incorporate observable covariates, but does not require these covariates, and so does not depend upon accurate documentation of customer and campaign design characteristics. On the other hand, we require that a sample of training data is available for the focal campaign, which their approaches do not need.

The second major distinction is that we account for uncertainty in the treatment effect estimates. When transferring information across campaigns, our approach recognizes that treatment effects in different campaigns capture customer responsiveness to potentially different marketing actions (design characteristics), and this responsiveness is measured with different precision. For example, the precision can depend on the size of the focal and source campaigns. We show that accounting for the varying precision across campaigns improves the accuracy of treatment effect predictions for the focal campaign. Our method extends the probabilistic matrix factorization framework (PMF; Salakhutdinov and Mnih, 2008) by explicitly accounting for the precision of the treatment effect measures used as inputs. This extension is important for marketing applications, and requires a change in both the definition and estimation of the PMF model. We demonstrate using both simulations and an empirical study that incorporating the precision information sharply improves targeting performance compared to the standard PMF framework.

Within the PMF literature, our proposed extension to the PMF model can be compared with two papers. Lakshminarayanan et al. (2011) introduce Robust Bayesian Matrix Factor-

ization to handle outliers and atypical customer behavior in the Netflix problem (predicting movie ratings for Netflix users). Their paper models heteroskedastic noise across user:movie observations by multiplying a global precision parameter by user-specific and movie-specific scaling factors. The latent scaling factors are estimated together with the user and movie embeddings. In contrast, Yang (2017) incorporate heteroskedasticity information as an input to the PMF framework (instead of estimating it). The paper aims to predict the conversion rates of digital ads, and to mitigate the sparsity in responses, defines the precision of available conversion rates as a function of the average conversion rates and the number of impressions. Similar to these two papers, our proposed extension focuses on heteroskedastic noise. We decompose the variation in observed treatment effects into two components: random deviation from the embedding structure and imprecision in the measurement. We assume that the first component is homoskedastic, and estimate it in a Bayesian framework along with customer and campaign embeddings. The second component varies across observations. We assume that uncertainties associated with the treatment effect measurements are known and provide them as inputs to the model.

3 Proposed Method: Bayesian Matrix Factorization

Our approach extends a matrix factorization framework to estimate treatment effects across marketing campaigns and customer segments. We parameterize the treatment effects using campaign and customer embeddings, and infer these embeddings from the available (noisy) treatment effect measures. We extend the standard PMF framework to account for the precision of these measures. We use the estimated treatment effects to design a targeting policy for the focal campaign.

3.1 Model Setup

We consider a firm that operates in a market with S customer segments and distributes C marketing campaigns to these customers. We index segments by $s \in \{1, .., S\}$ and marketing campaigns by $c \in \{1, .., C\}$. We assume that customer segmentation is known and is not updated during model training. The effectiveness of marketing campaigns varies across segments. We denote a conditional average treatment effect (CATE) of campaign c in segment s as m_{cs}.

The firm's goal is to predict CATEs for each customer segment that is eligible for the focal campaign. To design a targeting policy, the firm can use the predicted CATEs to identify segments for which the expected response exceeds the cost of the marketing action. Alternatively, in our empirical application, we consider ranking customer segments based upon their responsiveness to the focal campaign. Our method is agnostic to how firms use the predicted treatment effects to design a targeting policy, and we focus instead on estimating the CATEs.

For the focal and source campaigns, the firm has information about the CATEs in different segments. We denote the observed treatment effect measure for campaign c in segment s as $\widehat{m}_{cs}$. These measures could be obtained from pilot experiments, experiments conducted the previous year, natural experiments, or observational methods. We assume that the measurement precision λ_{cs} is known, and that the error has a zero-mean Gaussian distribution:

$$p(\widehat{m}_{cs}|m_{cs}) = \mathcal{N}(\widehat{m}_{cs}|m_{cs}, \lambda_{cs}^{-1}) \tag{1}$$

The notation $\mathcal{N}(x|\mu, \lambda^{-1})$ indicates a Gaussian random variable x with mean μ and variance λ^{-1}, where λ is a precision parameter. Hereafter, we denote three $C \times S$ matrices for the true treatment effects $\mathrm{M} = (m_{cs})$, the treatment effect estimates $\widehat{\mathrm{M}} = (\widehat{m}_{cs})$, and the precision of these estimates $\Lambda = (\lambda_{cs})$.

The available measurements are limited to the customer segments that were eligible for a given campaign. We capture the eligibility criteria by $q_{cs} \in \{0, 1\}$, which equals 1 if segment s was eligible for campaign c, and zero otherwise.

Our proposed approach infers the unobserved true treatment effects M from the (noisy) treatment effect measures $\widehat{\mathrm{M}}$ and precision Λ. It recognizes that different customer segments may respond similarly to a firm's marketing campaigns, and that different marketing campaigns may yield similar responses. We will infer these similarities from $\widehat{\mathrm{M}}$ and Λ.

We provide a complete list of notations in Appendix A.

3.2 Model Structure

We parametrize the distribution of conditional average treatment effects m_{cs} by segment embeddings $U_s \in \mathbb{R}^{K\times 1}$ and campaign embeddings $V_c \in \mathbb{R}^{K\times 1}$:

$$p\left(m_{cs}|U_s, V_c, \lambda_m\right) = \mathcal{N}\left(m_{cs}|U_s'V_c, \lambda_m^{-1}\right) \tag{2}$$

where $U_s'V_c$ and λ_m^{-1} indicate the mean and variance of the normal distribution.

We place Gaussian priors on the embeddings for customer segments and marketing campaigns:

$$\begin{aligned} p(U|\mu_U, \Lambda_U) &= \prod_s \mathcal{N}(U_s|\mu_U, \Lambda_U^{-1}) \\ p(V|\mu_V, \Lambda_V) &= \prod_c \mathcal{N}(V_c|\mu_V, \Lambda_V^{-1}) \end{aligned} \tag{3}$$

where $U = (U_s) \in \mathbb{R}^{K\times S}$, $V = (V_c) \in \mathbb{R}^{K\times C}$, and (μ_U, Λ_U) and (μ_V, Λ_V) are mean and precision matrices. For ease of exposition, we denote sets of parameters $\Omega_U \equiv (\mu_U, \Lambda_U)$ and $\Omega_V \equiv (\mu_V, \Lambda_V)$.

We further place Gaussian-Wishart priors on parameters (μ_U, Λ_U) and (μ_V, Λ_V), and

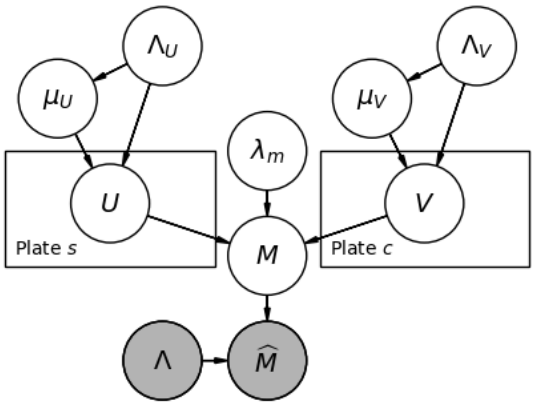
a Gamma prior on the parameter λ_m:

$$
\begin{aligned}
p(\Omega_U) &= p(\mu_U|\Lambda_U)p(\Lambda_U) = \mathcal{N}\left(\mu_U|\eta_0, (\kappa_0\Lambda_U)^{-1}\right) \cdot \mathcal{W}(\Lambda_U|W_0, \nu_0) \\
p(\Omega_V) &= p(\mu_V|\Lambda_V)p(\Lambda_V) = \mathcal{N}\left(\mu_V|\eta_0, (\kappa_0\Lambda_V)^{-1}\right) \cdot \mathcal{W}(\Lambda_V|W_0, \nu_0) \\
p(\lambda_m) &= \mathcal{G}(\lambda_m|\alpha_0, \beta_0)
\end{aligned}
\tag{4}
$$

where κ_0 is a scale factor of normal-Wishart distribution; $\mathcal{W}$ is the Wishart distribution with ν_0 degrees of freedom and $K \times K$ scale matrix W_0; $\mathcal{G}$ is a Gamma distribution with shape α_0 and rate β_0.

Figure 1 illustrates our model using plate notations. An important characteristic of the proposed model is the separability of customer embeddings U_s and campaign embeddings V_c, which allows for efficient inference. The parametrization of the treatment effects with a dot product of embeddings does not limit the accuracy of the model. Because the embeddings are sampled from a K-dimensional vector space, the model is flexible enough to approximate non-linear patterns in treatment effects across segments and campaigns. We demonstrate this point in Section 4.

Figure 1: Graphical Representation of the Model



The figure summarizes the design of the model using plate notation.

We next describe how we estimate the proposed model and the corresponding inference

for the treatment effects M.

3.3 Inference

We estimate the conditional distribution $p(m_{cs}|\widehat{M})$:

$$p(m_{cs}|\widehat{M}) = \iint p(m_{cs}|U, V, \lambda_m, \widehat{M}) p(U, V, \lambda_m, \Omega_U, \Omega_V|\widehat{M}) dU dV d\lambda_m d\Omega_U d\Omega_V \tag{5}$$

where we used the law of total probability and the fact that m_{cs} is independent of Ω_U, Ω_V conditional on U, V.

In Appendix B we derive a closed form expression for $p(m_{cs}|U, V, \lambda_m, \widehat{M})$:

$$p(m_{cs}|U, V, \lambda_m, \widehat{M}) = \begin{cases} \mathcal{N}\left(m_{cs}|U_s' V_c, \lambda_m^{-1}\right) & \text{if } q_{cs} = 0 \\ \mathcal{N}\left(m_{cs}\middle|\frac{U_s' V_c \lambda_m + \widehat{m}_{cs}\lambda_{cs}}{\lambda_{cs}+\lambda_m}, \frac{1}{\lambda_{cs}+\lambda_m}\right) & \text{if } q_{cs} = 1 \end{cases} \tag{6}$$

The joint distribution $p(U, V, \lambda_m, \Omega_U, \Omega_V|\widehat{M})$ and the integral in Equation (5) do not have a closed form. We approximate the integral using an average over a set of independent random samples $\{U^t, V^t, \lambda_m^t, \Omega_U^t, \Omega_V^t\}_{t=1}^T$:

$$p(m_{cs}|\widehat{M}) \simeq \frac{1}{T} \sum_t p(m_{cs}|U^t, V^t, \lambda_m^t, \widehat{M}) \tag{7}$$

To draw independent samples $\{U, V, \lambda_m, \Omega_U, \Omega_V\}_{t=1}^T$, we implement the Gibbs sampling algorithm. Gibbs sampling draws hidden variables iteratively, while conditioning on all other hidden variables. This approach is computationally efficient if the conditional distributions have a closed form. Our choice of priors allows us to derive an analytical form for the distribution of all hidden variables except λ_m. This substantially decreases the computation

time and hardware requirements.[3]

Conditional Distributions for U and V

We start our derivations with the conditional distribution of U:

$$\begin{aligned} p(U|V,\lambda_m,\Omega_U,\Omega_V,\widehat{M}) = \frac{p(U,V,\lambda_m,\Omega_U,\Omega_V,\widehat{M})}{\int p(U,V,\lambda_m,\Omega_U,\Omega_V,\widehat{M})dU} &= \frac{p(\widehat{M}|U,V,\lambda_m)p(U|\Omega_U)}{\int p(\widehat{M}|U,V,\lambda_m)p(U|\Omega_U)dU} \\ &\propto \left[\prod_{c,s}\mathcal{N}\left(\widehat{m}_{cs}|U_s'V_c,\lambda_m^{-1}+\lambda_{cs}^{-1}\right)\right]\left[\prod_s \mathcal{N}(U_s|\mu_U,\Lambda_U)\right] \\ &\propto \prod_s\left[\mathcal{N}(U_s|\mu_U,\Lambda_U)\prod_c\mathcal{N}\left(\widehat{m}_{cs}|U_s'V_c,\lambda_m^{-1}+\lambda_{cs}^{-1}\right)\right] \end{aligned} \tag{8}$$

Equation (8) demonstrates that all components U_s are conditionally independent, so the sampling can be done in parallel for each U_s separately. Furthermore, for each component U_s, we can use the properties of the conjugate priors to derive the sampling distribution:

$$p(U|V,\lambda_m,\Omega_U,\Omega_V,\widehat{M}) = \prod_s p(U_s \mid V,\lambda_m,\Omega_U,\widehat{M}) = \prod_s \mathcal{N}\left(U_s \mid \widetilde{\mu}_{U_s},\widetilde{\Lambda}_{U_s}\right) \tag{9}$$

where the parameters of the distribution can be expressed analytically:

$$\begin{aligned} \widetilde{\mu}_{U_s} &= \widetilde{\Lambda}_{U_s}^{-1}\left[\sum_c \frac{q_{cs}}{\lambda_{cs}^{-1}+\lambda_m^{-1}}V_c\widehat{m}_{cs}+\Lambda_U\mu_U\right] \\ \widetilde{\Lambda}_{U_s} &= \Lambda_U + \sum_c \frac{q_{cs}}{\lambda_{cs}^{-1}+\lambda_m^{-1}}V_cV_c' \end{aligned} \tag{10}$$

By symmetry, we derive the closed-form expression for the campaign embeddings V_c:

$$p(V|U,\lambda_m,\Omega_U,\Omega_V,\widehat{M}) = \prod_c p(V_c \mid U,\lambda_m,\Omega_V,\widehat{M}) = \prod_c \mathcal{N}\left(V_c \mid \widetilde{\mu}_{V_c},\widetilde{\Lambda}_{V_c}\right) \tag{11}$$

[3]In our empirical application, we estimate the model using a 2023 MacBook Pro (32GB), achieving a processing speed of approximately 20 seconds for every 1,000 samples.

with the following parameters of the posterior distribution:

$$\begin{aligned}\widetilde{\mu}_{V_c} &= \widetilde{\Lambda}_{V_c}^{-1}\left[\sum_s \frac{q_{cs}}{\lambda_{cs}^{-1}+\lambda_m^{-1}} U_s \widehat{m}_{cs} + \Lambda_V \mu_V\right] \\ \widetilde{\Lambda}_{V_c} &= \Lambda_V + \sum_s \frac{q_{cs}}{\lambda_{cs}^{-1}+\lambda_m^{-1}} U_s U_s'\end{aligned} \tag{12}$$

Conditional Distributions for Ω_U and Ω_V

Similarly, we can compute the distribution $\Omega_U = (\mu_U, \Lambda_U)$:

$$\begin{aligned}p(\Omega_U \mid U, V, \lambda_m, \Omega_V, \widehat{M}) &\propto p(U \mid \Omega_U) p(\Omega_U) \\ &\propto \left[\prod_s \mathcal{N}(U_s \mid \mu_U, \Lambda_U)\right] \mathcal{N}\left(\mu_U \mid \eta_0, (\kappa_0 \Lambda_U)^{-1}\right) \mathcal{W}(\Lambda_U \mid W_0, \nu_0) \\ &= \mathcal{N}\left(\mu_U \mid \widetilde{\eta}_U, (\widetilde{\kappa}_U \Lambda_U)^{-1}\right) \mathcal{W}(\Lambda_U \mid \widetilde{W}_U, \widetilde{\nu}_U)\end{aligned} \tag{13}$$

where the parameters of the posterior distribution can be expressed analytically:

$$\begin{aligned}&\widetilde{\kappa}_U = \kappa_0 + S; \quad \widetilde{\nu}_U = \nu_0 + S; \quad \widetilde{\eta}_U = \frac{\eta_0 \kappa_0 + S\overline{U}}{\kappa_0 + S}; \quad \overline{U} = \frac{\sum_s U_s}{S} \\ &\widetilde{W}_U^{-1} = W_0^{-1} + \sum_s (U_s - \overline{U})(U_s - \overline{U})' + \frac{\kappa_0 S}{\kappa_0 + S}(\overline{U} - \eta_0)(\overline{U} - \eta_0)'\end{aligned} \tag{14}$$

The structure of the Normal-Wishart distribution suggests a two-stage sampling of Ω_U: (1) sample Λ_U according to the Wishart distribution; (2) given the realization of Λ_U sample μ_U according to the normal distribution. The sampling distribution only depends upon the segment embedding U, which contains all of the information about Ω_U (see Figure 1).

We can derive the distribution for the parameters Ω_V by symmetry:

$$p(\Omega_V \mid U, V, \lambda_m, \Omega_U, \widehat{M}) = \mathcal{N}\left(\mu_V \mid \widetilde{\eta}_V, (\widetilde{\kappa}_V \Lambda_V)^{-1}\right) \mathcal{W}(\Lambda_V \mid \widetilde{W}_V, \widetilde{\nu}_V) \tag{15}$$

where the parameters of the posterior distribution can be expressed analytically:

$$\begin{aligned}
&\widetilde{\kappa}_V = \kappa_0 + C; \quad \widetilde{\nu}_V = \nu_0 + C; \quad \widetilde{\eta}_V = \frac{\eta_0 \kappa_0 + C\overline{V}}{\kappa_0 + C}; \quad \overline{V} = \frac{\sum_c V_c}{C} \\
&\widetilde{W}_V^{-1} = W_0^{-1} + \sum_c (V_c - \overline{V})(V_c - \overline{V})' + \frac{\kappa_0 C}{\kappa_0 + C}(\overline{V} - \eta_0)(\overline{V} - \eta_0)'
\end{aligned} \tag{16}$$

Conditional Distribution for λ_m

Lastly, we derive the conditional distribution for λ_m:

$$p(\lambda_m | U, V, \Omega_U, \Omega_V, \widehat{M}) \propto \left[\prod_{c,s} \mathcal{N}\left(\widehat{m}_{cs} | U_s' V_c, \lambda_m^{-1} + \lambda_{cs}^{-1}\right)^{q_{cs}} \right] \mathcal{G}(\lambda_m | \alpha_0, \beta_0) \tag{17}$$

Equation (17) does not support a closed-form expression for a posterior distribution. Moreover, we could not identify a conjugate prior for λ_m that would absorb variation in λ_{cs} across the campaign-segment combinations to ensure an analytical form for the posterior distribution. We thus rely on the Markov Chain Monte Carlo (MCMC) approach to draw samples from the conditional distribution $p(\lambda_m | U, V, \Omega_U, \Omega_V, \widehat{M})$.

We can improve the efficiency of the MCMC sampling for Equation (17) by incorporating the derivative of the log-likelihood function. In particular, the logarithm of Equation (17) and its derivative can be expressed as follows:

$$\begin{aligned}
\log p(\lambda_m | U, V, \Omega_U, \Omega_V, \widehat{M}) = &- \sum_{s,c} \frac{q_{cs}}{2} \left[\frac{(\widehat{m}_{cs} - U_s' V_c)^2}{\lambda_{cs}^{-1} + \lambda_m^{-1}} + \log\left(\lambda_{cs}^{-1} + \lambda_m^{-1}\right) \right] \\
&+ \log Z + \log \mathcal{G}(\lambda_m | \alpha_0, \beta_0)
\end{aligned} \tag{18}$$

$$\begin{aligned}
\frac{\partial \log p(\lambda_m | U, V, \Omega_U, \Omega_V, \widehat{M})}{\partial \lambda_m} = &- \sum_{s,c} \frac{q_{cs}}{2} \left[\frac{(\widehat{m}_{cs} - U_s' V_c)^2}{(\lambda_m / \lambda_{cs} + 1)^2} - \frac{1}{\lambda_m (\lambda_m / \lambda_{cs} + 1)} \right] \\
&+ \frac{\alpha_0 - 1}{\lambda_m} + \beta_0
\end{aligned} \tag{19}$$

where Z is the normalization constant.

The derivative in Equation (19) does not include the normalization constant Z, which allows us to use the Hamiltonian Monte Carlo (HMC) sampling approach. Similarly to the Metropolis-Hastings algorithm, the HMC generates samples from a target distribution, $p(x) \sim \pi(x)/Z$, by first sampling random variables from a simpler distribution, $q(x^t|x^{t-1})$, and then accepting (or rejecting) new samples using a stochastic rule. The stochastic rule is defined so that over time, a sequence of accepted samples closely approximates the target distribution $p(x)$. HMC improves on the Metropolis-Hastings algorithm by utilizing the derivative of the log-likelihood function. This allows for more efficient exploration of the parameter space, and leads to faster convergence (burn-in) and better approximation of the target distribution. We provide details in Appendix C.

Connecting the Steps

We summarize the algorithm for estimating our proposed model in Algorithm 1. Our algorithm yields a distribution $p(m_{cs}|\widehat{M})$. This is a Gaussian mixture distribution with T components with equal weights, where each component follows the normal distribution specified in Equation (6). This distribution has two helpful properties. First, the mean of this Gaussian mixture is an average of the components' means. We can thus write a point prediction as follows:

$$m^*_{cs} = \begin{cases} \frac{1}{T}\sum_t U_s^{t\prime} V_c^t & \text{if } q_{cs} = 0 \\ \frac{1}{T}\sum_t \left(\lambda_{cs} + \lambda_m^t\right)^{-1} \cdot \left(\lambda_m^t U_s^{t\prime} V_c^t + \lambda_{cs}\widehat{m}_{cs}\right) & \text{if } q_{cs} = 1 \end{cases} \tag{20}$$

Second, we can sample from a posterior distribution $(m_{cs}|\widehat{M})$ by first sampling a component (with equal probabilities), and then sampling from the corresponding normal distribution. This allows us to obtain a complete distribution of the predicted values rather than a point estimate.

Algorithm 1 Proposed Method: Bayesian Matrix Factorization

Inputs: $(\widehat{m}_{cs}, \lambda_{cs}, q_{cs})$ for $c = 1, .., C$ and $s = 1, .., S$

Initialize: (U^0, V^0, λ_m^0)

for $t = 1$ to $\tau + T$ **do**

Sample λ_m^t using HMC algorithm from Appendix C:

$$\lambda_m^t \sim p(\lambda_m | U^{t-1}, V^{t-1}, \widehat{M})$$

Sample (Ω_U^t, Ω_V^t) using Equations (13)-(16):

$$\Omega_U^t \sim p(\Omega_U | U^{t-1}), \quad \Omega_V^t \sim p(\Omega_V | V^{t-1})$$

for $s = 1$ to S **do**

Sample segment embeddings U_s^t using Equations (8)-(10):

$$U_s^t \sim p(U_s | V^{t-1}, \lambda_m^t, \Omega_U^t, \widehat{M})$$

for $c = 1$ to C **do**

Sample campaign embeddings V_c^t using Equations (11)-(12)

$$V_c^t \sim p(V_c | U^t, \lambda_m^t, \Omega_V^t, \widehat{M})$$

Discard the first τ samples of (U^t, V^t, λ_m^t), and use the remaining T samples to approximate the posterior distribution $p(m_{cs}|\widehat{M})$:

$$p(m_{cs}|\widehat{M}) \simeq \frac{1}{T} \sum_t p(m_{cs} | U^t, V^t, \lambda_m^t, \widehat{M})$$

where $p(m_{cs}|U^t, V^t, \lambda_m^t, \widehat{M})$ is defined in Equation (6).

The estimator in Equation (20) has an intuitive interpretation. If measurement $\widehat{m}_{cs}$ is not available for a segment-campaign, the model uses $\frac{1}{T}\sum_t U_s^{t\prime} V_c^t$ for prediction. If measurement $\widehat{m}_{cs}$ is available ($q_{cs} = 1$), the model combines $\widehat{m}_{cs}$ and $U_s^{t\prime} V_c^t$ weighted by the precision of each component, and then averages these predictions across T samples. If the measurement in the segment is very precise ($\lambda_{cs} \to +\infty$), the prediction converges to the measurement: $m^*_{cs} \to \widehat{m}_{cs}$. And vice versa, $\lim_{\lambda_{cs}\to 0} m^*_{cs} = \frac{1}{T}\sum_t U_s^{t\prime} V_c^t$.

The proposed model allows learning representations for a new marketing campaign V_{new} without completely retraining the model. In practice, firms can estimate the model with a large database of source campaigns once and save the MCMC draws. For any new focal campaign, we can use the closed-form conditional distributions in Equations (11)-(12) to draw embeddings for the new campaign, without updating the draws of segment embeddings.

3.4 Discussion

We have framed the firm's problem as predicting CATEs at the segment level, but the proposed method also supports estimation at the individual customer level. Firms could use the customer-level estimates to design targeting policies that vary marketing actions for individual customers rather than segments. A challenge is that the firm would require information about treatment effects for focal and source campaigns at the individual customer level, including both point estimates and precision. If the firm has this information, then our method scales and can be applied to this task.

The proposed method can be extended to incorporate observable covariates for customer segments and marketing campaigns. For example, firms often have the past purchasing histories for customer segments, or maintain records of the mailing dates (seasonality) for past marketing campaigns. One approach for incorporating this information is to represent the embeddings as a linear combination of the covariates with segment-specific or

campaign-specific weights:

$$
\begin{aligned}
p(U|\beta_U, \Lambda_U) &= \prod_s \mathcal{N}(U_s|\beta_U X_s, \Lambda_U^{-1}) \\
p(V|\beta_V, \Lambda_V) &= \prod_c \mathcal{N}(V_c|\beta_V X_c, \Lambda_V^{-1})
\end{aligned}
\tag{21}
$$

where X_s and X_c are the covariates for segment s and campaign c, and β_U and β_V correspond to the (random) coefficients. The hierarchical model preserves the analytical forms of the conditional distributions, and estimation of the model remains efficient. An alternative approach is to use nonlinear functional forms. For example, Adams et al. (2010) incorporates covariates into the standard PMF model using Gaussian processes.

In Section 2, we acknowledged that our approach builds on the PMF formulation (Salakhutdinov and Mnih, 2008). When the PMF model was initially proposed, the 'Netflix Problem' was used as motivation. The goal is to predict Netflix movie ratings for a new user : movie combination, based on the ratings by other users, and ratings by that user for other movies. There are two important differences between this problem and our problem. First, the training data in our problem represents noisy measures of the treatment effects, instead of the "true" treatment effects. To incorporate the associated uncertainty, we extended the PMF framework by introducing the λ_{cs} terms, and consequently adjusting the inference process (for λ_m). Second, in the Netflix problem the goal is to predict movie ratings where no rating exists. In our setting, the focal data contains noisy measures of the treatment effects, which changes the objective from value imputation to value updating.

4 Synthetic Data

We use simulation analysis to illustrate and validate the properties of the proposed method with a known data generating process. We show that the proposed method can effectively

combine information across marketing campaigns and customer segments, and scale to many campaigns and many segments.

4.1 Data Generating Process

We summarize the data generating process (DGP) in this section, and provide additional details in Appendix D. Throughout the simulation analysis, we maintain the same default parameter values, and vary parameters one-by-one to illustrate model performance. We emphasize that our goal is to predict treatment effects in different segments for a focal campaign, rather than to recover the parameters of the DGP.

Our DGP has three important components. First, marketing campaigns and customer segments have (latent) characteristics that define the treatment effects. Treatment effects are similar for marketing campaigns with similar characteristics, and similar segments have similar treatment effects. We denote the segment characteristics as $\mathcal{U}_s$, and campaign characteristics as $\mathcal{V}_c$. For illustrative purposes, we simulate clusters of customer segments, and clusters of campaigns. Notice a difference in notations: we use $\mathcal{U}_s$ and $\mathcal{V}_c$ for latent characteristics, and U_s and V_c for estimated embeddings for segments and campaigns. In Section 4.3 we will illustrate that these latent characteristics and estimated embeddings are not equivalent (by design).

Second, the treatment effect variation is not fully explained by the latent segment and campaign characteristics. Our data generating process assumes that for each segment-campaign combination, the "true" treatment effect m_{cs}^{DGP} is a function of latent characteristics $f(\mathcal{U}_s, \mathcal{V}_c)$ and an additive Gaussian noise with zero mean and precision λ_f. We simulate the noise terms independently across observations. We initially consider $f(\mathcal{U}_s, \mathcal{V}_c) = \mathcal{U}_s \cdot \mathcal{V}_c$, but also consider a nonlinear function in Section 4.3.

The third component of the data generating process recognizes that the firm does

not observe "true" treatment effects m_{cs}^{DGP}, and instead observes noisy information about the treatment effects. We will model this information as coming from experiments. Our simulation randomly draws experimental sample sizes for each campaign N_c, and assumes that the firm allocates these samples proportional to segment sizes w_s in balanced experiments. We further assume that the idiosyncratic noise in the outcomes for individual observations is λ_ϵ^{-1}, so the difference-in-means estimates follow the distribution:

$$\widehat{m}_{cs} \sim \mathcal{N}\left(m_{cs}^{DGP}, \lambda_{cs}^{-1}\right) \tag{22}$$

$$\lambda_{cs} = w_s N_c \cdot \lambda_\epsilon \tag{23}$$

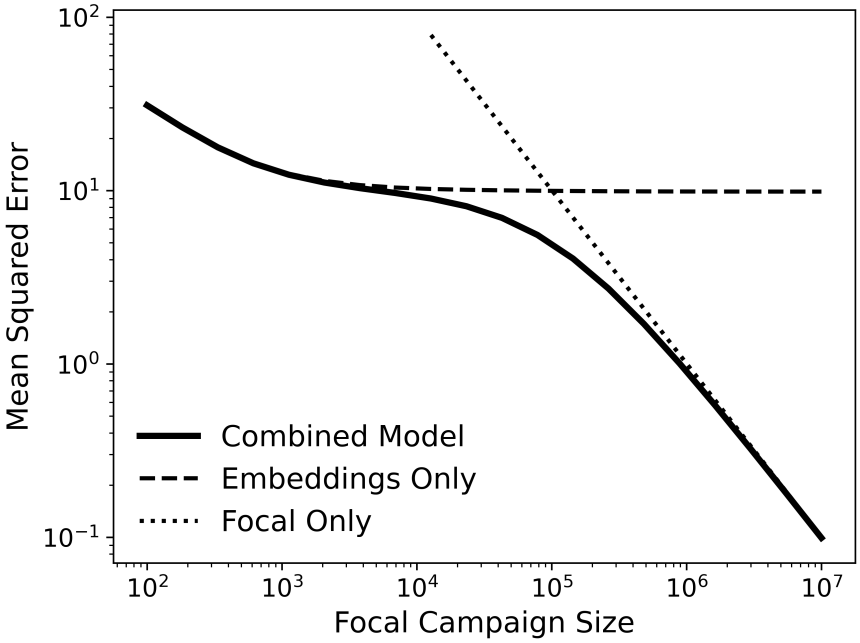
We estimate the proposed model using the simulated $(\widehat{m}_{cs}, \lambda_{cs})$ for $c = 1, .., C$ and $s = 1, .., S$. Notice that the model calibration and inference does not use information about the latent characteristics $\mathcal{U}_s$ and $\mathcal{V}_c$, or the "true" treatment effects m_{cs}^{DGP}. We will use these known parameters from the DGP to evaluate the model performance.

4.2 Predictive Performance

In Figure 2, we evaluate the predictive performance of the proposed model. On the x-axis, we vary the size of the training data in the focal campaign. On the y-axis, we plot the mean squared error (MSE) when comparing the predicted treatment effects to the true treatment effects. We report the performance of three models. The performance of the "Focal Only" model is reported using the dotted line. This is a baseline model that uses treatment effects from the training data as predicted values, and does not use any information from the source campaigns. The performance of the Focal Only model improves linearly (on the log-log scale) with the size of the training data.

The dashed line corresponds to the "Embeddings Only" model, which predicts treatment effects using the segment and campaign embeddings alone, without combining the

Figure 2: Overall Predictive Performance



The figure reports the accuracy of the treatment effect predictions from different models. The x-axis reports the sample size of the focal data and the y-axis measures prediction accuracy using mean squared error (MSE).

embedding-based predictions with the Focal Only measures. Specifically, the Embeddings Only model predicts treatment effects using the first line in Equation (20) and isolates the information contained in the estimated embeddings. With little focal training data, the Embedding Only model performs better than the Focal Only model, because it leverages the segment embeddings (even though the embedding for the focal campaign is just random). Intuitively, the U embedding captures information about the generic responsiveness of different customer segments to the firm's marketing actions. With additional focal training data, the Embeddings Only model initially improves because estimates of U and V both improve, and the focal campaign embedding is no longer random. However, the performance of the Embeddings Only model eventually plateaus as the sample size of the focal experiment continues to increase. This is because the DGP incorporates unexplained noise (λ_f), which cannot be captured without direct measurement of the focal treatment effects.

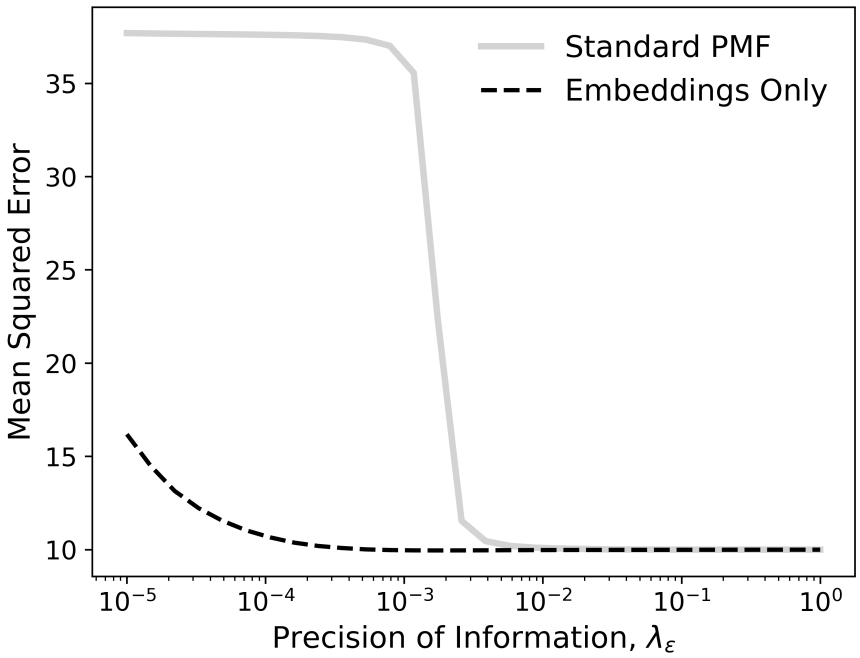
Our proposed model (the solid line) combines the predictions from the Embeddings Only and Focal Only models. We label this the "Combined Model". Due to the Bayesian updating structure, the combination is optimal, and the performance of the model is better than either of the individual approaches. When the sample size of the focal data is small, the Combined Model closely follows the Embeddings Only model and out-performs the Focal Only model. When the focal experiment is large, the Focal Only model provides precise estimates of the response to the focal campaign in each segment. As a result, the Focal Only model converges to the Combined Model.

Recall that our proposed model is motivated by and extends the PMF model. We introduce an adjustment to this model to account for the precision of the treatment effect measurements in the source and focal data. This modification has two implications for model performance. First, the Combined Model combines the (noisy) treatment effect estimates from the focal data with the predictions from the Embeddings Only model. These two components are weighted by the corresponding uncertainty, and Figure 2 confirms that the addition of the focal data estimates in the Combined Model outperforms the Embeddings Only model.

Second, the estimation of the embeddings themselves accounts for the precision in the source and focal data. The Standard PMF assumes that the available source and focal information is correct. For example, in the Netflix problem, the movie ratings provided by customers are known with certainty. However, in our context, treatment effects are measured with different precision across marketing campaigns and customer segments. Our proposed approach adjusts the embedding estimation for this varying precision.

Figure 3 demonstrates that accounting for the precision of the inputs yields better embedding estimates. In particular, we compare the performance of the Embeddings Only model to the Standard PMF model. Both models estimate treatment effects for the focal campaign from the embeddings alone. However, our Embeddings Only model adjusts for

Figure 3: Predictive Performance When Varying Noise in the Source and Focal Data



The figure compares the predictive performance of the Embeddings Only model and the Standard PMF. On the x-axis, we vary the amount of noise in the source and focal data, and the y-axis measures prediction accuracy using mean squared error (MSE).

uncertainty when estimating the embeddings, while the Standard PMF model does not. The y-axis measures the accuracy of the treatment effect predictions for the focal campaign. The x-axis varies the precision of the source and focal treatment effects used as inputs by both models. When the inputs are measured precisely (the right side of the x-axis) both models perform equally well. However, as the precision of the inputs deteriorates, the Embeddings Only model provides more accurate predictions because it accounts for the imprecision. In our empirical application in Section 5, which uses data from a sequence of direct mail field experiments, we compare the Standard PMF model with the Focal Only, Embeddings Only and Combined models. The findings reinforce the conclusion that accounting for imprecision in the inputs sharply improves accuracy.

4.3 Functional Form

Recall that we model treatment effects using the dot product of U and V (see Equation 2). This should not be interpreted as an implicit assumption that treatment effects are generated using the same functional form. If treatment effects have a different structure, the model can approximate the treatment effects by adapting the U and V embeddings, without changing the dot product specification. Intuitively, the flexibility in the Combined Model extends beyond the dot product of U and V to include the structure inherent in the embeddings themselves. We can adjust the model's flexibility by varying the dimensionality of these embeddings.

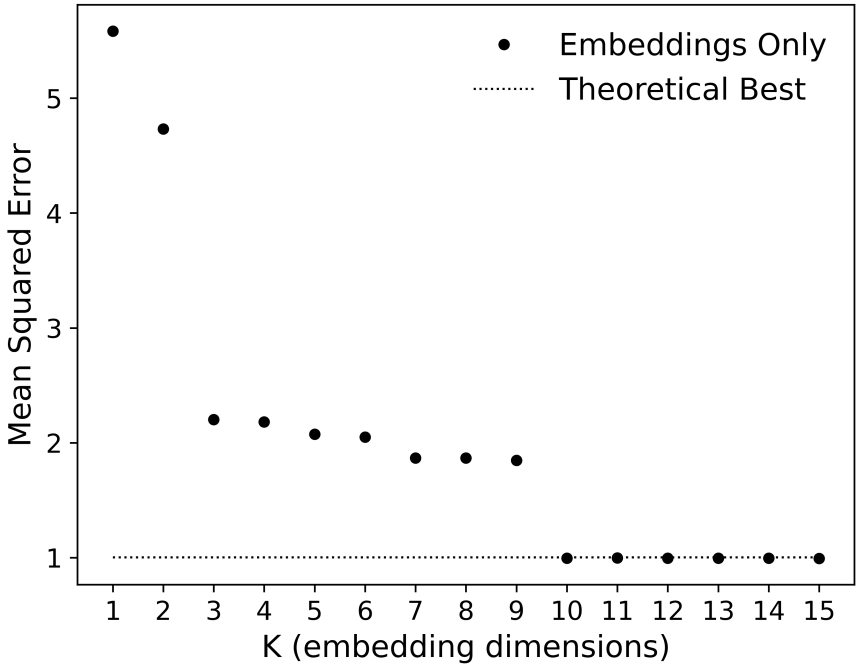
We illustrate this point by changing the functional form in the DGP. In particular, we modify the DGP to include a nonlinear relationship between the latent customer and segment characteristics:

$$f(\mathcal{U}_s, \mathcal{V}_c) = \mathcal{U}_s'\mathcal{V}_c + \text{sign}(\mathcal{U}_s'\mathcal{V}_c) \cdot (\mathcal{U}_s'\mathcal{V}_c)^2 + (\mathcal{V}_c'\mathcal{V}_c)^2 \tag{24}$$

We estimate the Embeddings Only model using data from this modified DGP and report its performance in Figure 4. On the x-axis, we vary the dimensionality of the U and V embeddings. This allows us to approximate treatment effects with a more flexible model. The rest of the model stays unchanged, including the dot-product specification in Equation (2). In Figure 4 we report the performance of both the Embeddings Only model and the "Theoretical Best". The Theoretical Best represents the predictive performance when using the true functional form and latent characteristics from the DGP (Equation (24)), so that the only error remaining in the model is variation in the treatment effects that the embeddings do not explain (λ_f).

When the embeddings each have just two dimensions, there is insufficient flexibility to capture the non-linearities in the DGP. However, with just three dimensions, the performance of the Embeddings Only model quickly improves. This improvement continues as the

Figure 4: Model Performance with Non-Linear DGP



The figure reports the predictive accuracy of the Embeddings Only model when using a non-linear functional form in the DGP (Equation 24). The x-axis reports the number of dimensions (K) in the segment and campaign embeddings. The y-axis measures prediction accuracy using mean squared error (MSE). The "Theoretical Best" indicates the predictive performance when using the true functional form from Equation (24).

flexibility of the embeddings grows. With ten dimensions, the Embeddings Only model perfectly fits the systematic variation in treatment effects across the segments and campaigns, and achieves its theoretical maximum performance. Even though the dot product functional form may appear mis-specified when the true model is Equation (24), with sufficient flexibility in the embeddings, the model is able to perfectly approximate the treatment effects. We conclude that the dot product specification should not be interpreted as a restrictive assumption in the model.

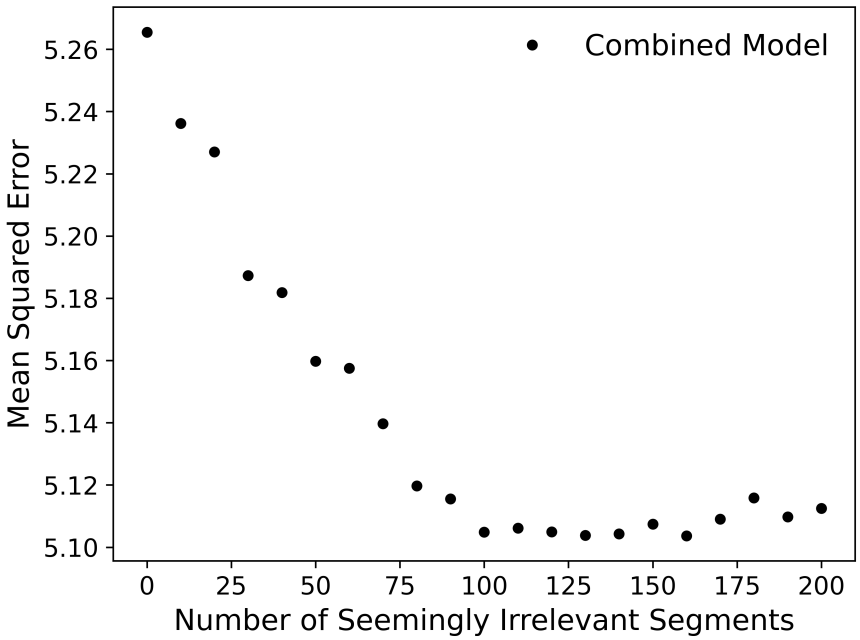
4.4 (Ir)relevant Source Data and Scalability

A source campaign can provide valuable information for predicting treatment effects in the focal campaign, even if the set of customers eligible for the source campaign is different than

the customers in the focal campaign. The basis of this argument is that the source data can help to estimate the segment and campaign embeddings, and this in turn will improve predictions for the focal campaign. To investigate this conjecture we incrementally added segments to the source campaigns in the estimation sample, where *none* of these additional segments are eligible to receive the focal campaign.[4]

Figure 5 illustrates the performance of the Combined Model when adding "Seemingly Irrelevant" segments to the source campaigns. We do not include these segments when evaluating MSE for the focal campaign. The findings are consistent with our conjecture. The incremental segments improve the performance of the Combined Model on the focal campaign, even though none of the additional segments are eligible for that campaign.

Figure 5: Seemingly Irrelevant Segments Can Provide Information



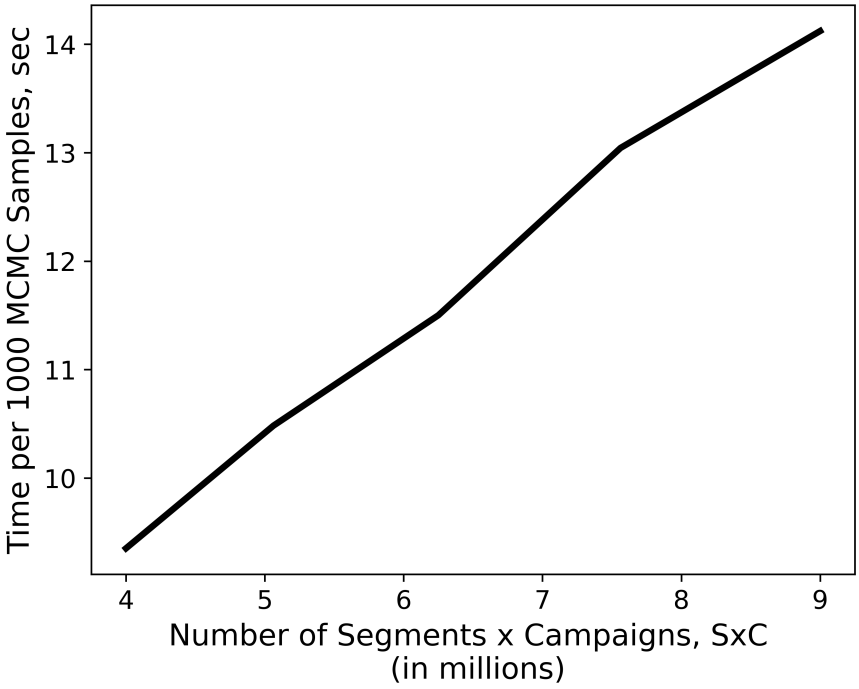
The figure reports the accuracy of the Combined Model when incrementally adding segments to the source campaigns that are not eligible for the focal campaign. The y-axis measures prediction accuracy using mean squared error (MSE) for the focal campaign.

An implication of the findings in Figure 5 is that firms should add all of their available

[4]For illustrative purposes, we increase sparsity of information in the default DGP. In particular, we create synthetic data in which only 5% of the segments are eligible for each source campaign.

source data when estimating the Combined Model. There is no need to trim the source data to remove seemingly irrelevant customers. This introduces a new question: how well does the estimation scale as the number of campaigns and segments grows? We investigate model scalability in Figure 6, where we report the time required to estimate each iteration of the Combined Model as we increase $S \times C$.[5] The results reveal a linear relationship between estimation time and the number of segments $\times$ campaigns. Bayesian models do not always scale well, but we are able to (1) derive analytical expressions for the conditional distributions used in the Gibbs sampling, and (2) reduce the MCMC sampling to a single variable with a known derivative of the log-likehood.

Figure 6: Scalability to Segment x Campaigns



The figure reports the tradeoff between computation time and size of the problem. The y-axis measures computation time in seconds per 1,000 MCMC samples. The x-axis measures the total number of segment : campaign combinations ($S \times C$).

This computational efficiency is reassuring. In our simulation analysis (Figures 2 -

[5]In each iteration, we increase both parameters S and C, while maintaining the same ratio $S = C$.

4), we estimate the model with 1,000 segments and 1,000 campaigns, and the estimation converges in 45 seconds for each set of parameters. In our empirical application, the model yields 1,000 samples from the posterior distribution in 12 seconds on a MacBook Pro 2023 (32GB). We discuss this empirical application next.

5 Empirical Application

We use data from field experiments conducted by a luxury fashion retailer to demonstrate the performance of the proposed method in an empirical setting. The retailer distributed direct mail campaigns to its customers, and for each campaign, randomly assigned customers into *mail* and *no mail* conditions. In this empirical setting, the proposed method substantially improves targeting decisions compared to the Focal Only baseline, as well as the standard PMF approach (which does not account for uncertainty in the source campaigns). Moreover, it estimates embeddings that capture meaningful relationships between segments and campaigns that were unobserved to the model.

5.1 Data Overview

Our empirical application uses data provided by a major US luxury fashion retailer. The retailer operates physical stores in many cities, together with an online website. The assortment spans men's and women's shoes and apparel, jewelry, accessories, and beauty products.

The retailer regularly distributes catalogs, postcards and other direct mail campaigns to its existing customers to announce and promote new products. The focus of these campaigns varies, and could include specific brands, specific product categories, specific seasons, or special events. To measure the overall impact (ATE) of each direct mail campaign on incremental purchases and profits, the retailer conducted A/B tests. The firm randomly

selected a 'no mail' control sample of qualified households for each campaign. We received the campaign descriptions and the circulation files identifying households randomly assigned to the *mail* and *no mail* conditions. The campaign descriptions include (cryptic) campaign names, mailing format (catalog, postcard, or brochure), mailing costs, and in-home dates.[6]

Our analysis focuses on direct mail campaigns distributed from January to August 2017. The focal campaign is a Q3 Beauty Postcard, a postcard promoting beauty products mailed to over 230,000 customers in August 2017.The 60 source campaigns include all direct mail experiments that involved at least 5,000 customers and were distributed prior to the focal campaign. We document the variation in campaign characteristics, and report randomization checks in Appendix F. In particular, the sample sizes in source campaigns varied substantially, from about 5,000 to over 1,000,000 customers. This variation affects how much we can learn about customer responsiveness from each of these source campaigns. It is this observation that motivated our extension of the standard PMF framework.

We define customer segments using geographic areas. In particular, each segment in our analysis corresponds to a 2-digit zip-code. While zip-code segmentation might seem simplistic, it is practical for the retailer. Its stores are almost all located in separate 2-digit zip codes. The retailer often stratified experiments and limited distribution of direct mail campaigns to certain zip-codes. Moreover, we will demonstrate in the next subsection that zip-code segmentation captures sufficient customer heterogeneity to enable targeting.

Our outcome measure is purchase incidence in the four weeks after the in-home date of the direct mail campaign. We construct the outcome variable using the retailer's transaction data, which includes customer identifiers and combines both online and in-store channels.

[6]Recall that our proposed method does not use these characteristics to transfer information across campaigns. However, we will refer to these characteristics in our post-analysis of the embeddings.

5.2 Targeting Performance

We assume that the firm wants to rank eligible customer segments according to their expected incremental purchase incidence in response to the focal campaign. We evaluate the performance of our proposed model using three steps.

Step 1: We randomly split the focal data into two subsets. Within each iteration, we randomly select 30% of customers in the focal data as focal training data and the remaining 70% as focal validation data.

Step 2: We use the 30% focal training data and the data from the source campaigns to estimate the Combined Model. The Combined Model uses the difference-in-means estimates and the corresponding precisions for all customer segments and campaigns (source and focal).

Step 3: We use the 70% focal validation data to compare the performance of the Focal Only and Combined Model. We rank customers in the focal validation data based on the predicted treatment effects and assign the focal treatment to the top-φ customers. For each value of φ we calculate the incremental purchase incidence acquired from mailing to these customers.[7]

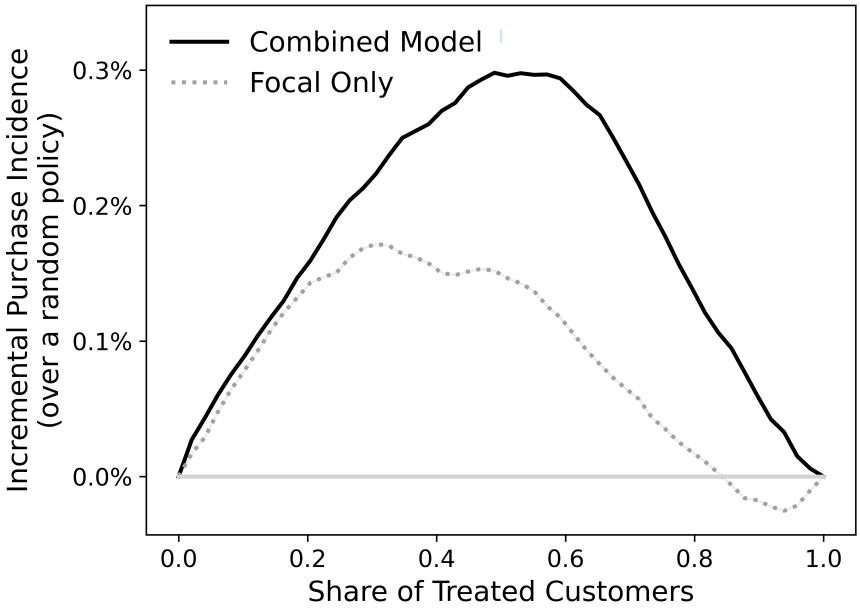
As a benchmark, we estimate the Focal Only model. The Focal Only model uses the 30% focal training data and a differences-in-means estimator to predict the treatment effect for each of the eligible segments in the focal campaign. The Focal Only policy does not use any data from the source campaigns.

We compare the performance of the Focal Only and Combined Model in Figure 7. To aid interpretation, we report performance improvements for each model over a "Random" benchmark. The Random policy ranks customer segments randomly and assigns the promotional treatment to the top-φ segments, where $\varphi \in [0, 1]$. In Figure 7, the x-axis measures

[7]We provide additional details about both the construction of the policies and the evaluation of performance in Appendix E.

the share of treated customers (φ). The y-axis measures the performance improvement as the purchase incidence gain over the Random policy.

Figure 7: Targeting Performance in the Empirical Application



The figure reports the incremental purchase incidence over a random policy for the Focal Only and Combined Model. The x-axis measures the share of treated customers (φ). For a given φ, the y-axis measures the difference in purchase incidence between a targeting model and the Random policy in percentage points.

We highlight two observations in Figure 7. First, the Focal Only models improves upon the Random policy for almost all values of $\varphi \in [0, 1]$. The performance of the Focal Only model confirms that the zip-code segment definitions capture meaningful heterogeneity in the treatment effects. Furthermore, the performance improvements are larger on the left side of the plot, indicating that it is easier to identify a small number of customer segments that are most responsive to the focal campaign. The Focal Only model struggles to rank-order the least-responsive segments.

Second, Figure 7 demonstrates that information from source campaigns can improve targeting policies for a focal campaign. The Combined Model yields higher expected purchase

incidence than the Focal Only model for essentially all values of φ. The only exceptions are small values of φ, where the Focal Only model is also able to identify the segments that are most responsive to the focal campaign. The Combined Model and Focal Only model use the same (zip-code) segments, and in the absence of source data, the two models yield equivalent expected performance. We conclude that the performance improvement for the Combined Model is attributable to the transfer of information about treatment effects from the source data.

In Table 1, we report the areas under the performance curves for different targeting models (AUTOC; Yadlowsky et al. (2025)). Each row corresponds to a different targeting model, and for each model, we report the AUTOC score together with the standard errors from bootstrap iterations. On average over $\varphi \in [0, 1]$, the Focal Only model outperforms the Random Policy by %0.084, and the Combined model more than doubles the gains from targeting (%0.173). Using average revenue per purchase incidence and a number of direct mail pieces distributed each year, a %0.1 increase in purchase incidence yields over $65 million in additional revenue annually. For a large retailer, improved targeting yields sizable improvements in marketing effectiveness.

We can also compare the performance of the proposed model to the Standard PMF. In Table 1, the Standard PMF does not yield an improvement over the Random Policy. The challenge is that the Standard PMF assumes all treatment effects measurements for the source campaigns are ground truth. However, as we previously noted, there is substantial variation in sample sizes across source campaigns. Treatment effects for some segment x campaign combinations could be measured with as few as 30 experimental samples, while other segment x campaigns could contain over 2,000 samples. By treating these data as equally-informative, the Standard PMF overfits noisy treatment effect measures, and the embeddings poorly predict the true treatment effects. In contrast, our proposed approach (Combined Model) accounts for the precision of the treatment effect measures in the source

Table 1: Incremental Purchase Incidence: AUTOC

Targeting Method	**AUTOC (%)**
Focal Only	0.084 (0.042)
Embeddings Only	0.172 (0.035)
Combined Model	0.173 (0.035)
Standard PMF	0.003 (0.050)

The table reports the incremental performance of the Combined Model over the Focal Only model when using different campaigns as Focal. Performance is measured as revenue per customer, and incremental performance is measured as the area between the $R(\varphi)$ curves for the Combined Model and the Focal Only model (see Appendix E). Standard errors are in parentheses.

campaigns and substantially improves performance over both the Random Policy and the Focal Only model.

5.3 Embedding Structure

The embeddings estimated by our proposed model capture meaningful relationships between customer segments and marketing campaigns that were unobserved by the model. To investigate the campaign embeddings, we identified source campaigns that the model located close to the focal campaign. Recall that the focal campaign in our empirical analysis is the Q3 Beauty Postcard. The three closest campaigns in the embeddings space are Q1 Beauty Postcard, Clinique Brochure, and Lancome Brochure. Notably, all of these three campaigns promoted beauty products. This was not initially obvious to our research team from the campaign descriptions. Clinique and Lancome campaigns were sponsored by vendors rather than the retailer's marketing team, and they had no word "beauty" in the campaign description. We related these campaigns to the beauty products *after* observing the estimation results.

In our empirical setting, the campaigns most-distant from the focal campaign focused on lapsed customers. Reactivation campaigns tend to promote discounts rather than product categories. They are also distributed to customers with no spending in the last year, while the Q3 Beauty Postcard was mailed to customers with recent purchases in the beauty category. The differences in content and audiences can explain why these campaigns were placed in different parts of the embedding space.

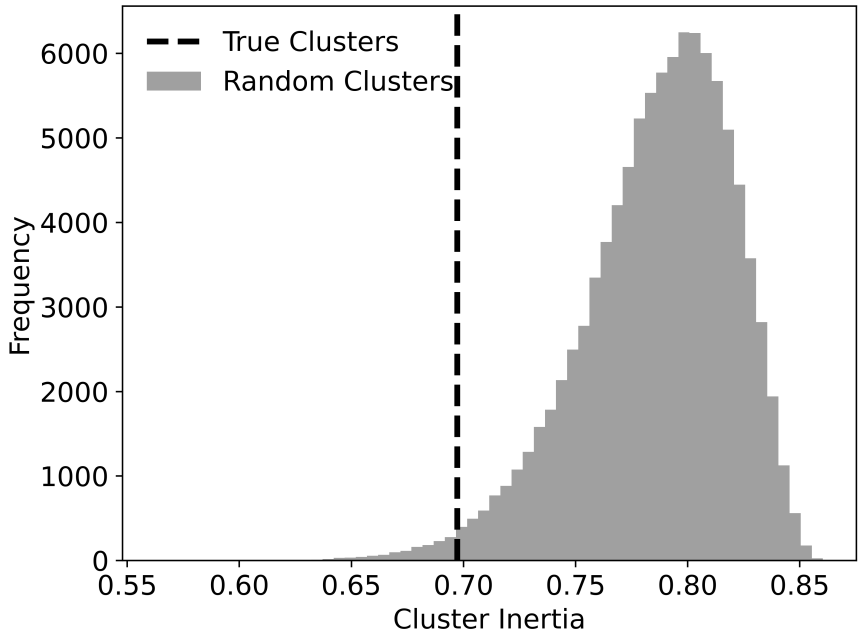
To investigate the segment embeddings, recall that the segments in our empirical application are geographic divisions, representing US Postal Service 2-digit zip-codes. We would expect that segments that are geographic neighbors would respond more similarly to firms' marketing campaigns than segments that are not located close to each other.[8] We also expect that segment embeddings produced by our model will position segments that respond more similarly to firms' marketing actions closer to each other in the embedding space. We next evaluate whether segments that are geographic neighbors are also positioned as neighbors in the embedding space.

We start by identifying the number of segments (2-digit zip codes) within each 1-digit zip code. We then use segment embeddings to calculate *inertia*. *Inertia* summarizes the Euclidean distances from each segment to the centroid of the cluster (1-digit zip code). It calculates a within-cluster sum-of-squared distances to measure the dispersion of the 2-digit zip codes in each 1-digit zip code, and then aggregates these sum-of-squares across all 2-digit zip codes in the dataset.

To interpret this measure for the estimated segment embeddings, we compare it with a random benchmark. In the random benchmark we randomly reassign segments to 1-digit zip codes, while preserving the number of segments in each 1-digit zip code. We repeat the random reassignments 100,000 times and present the random benchmark as a distribution.

[8]This expectation is commonly used as an identifying assumption in research comparing the behavior of customers on either side of state borders (see for example, Shapiro et al., 2021; Anderson et al., 2010).

Figure 8: Segment Embeddings Capture the Zip-Code Assignments



The figure reports the histogram of *Inertia* measures for 100,000 random assignments of 2-digit zip codes to 1-digit zip codes. *Inertia* is measured using Euclidean distances in the segment embedding space. The dashed vertical line represents the *inertia* for the true allocation of 2-digit zip codes to 1-digit zip codes.

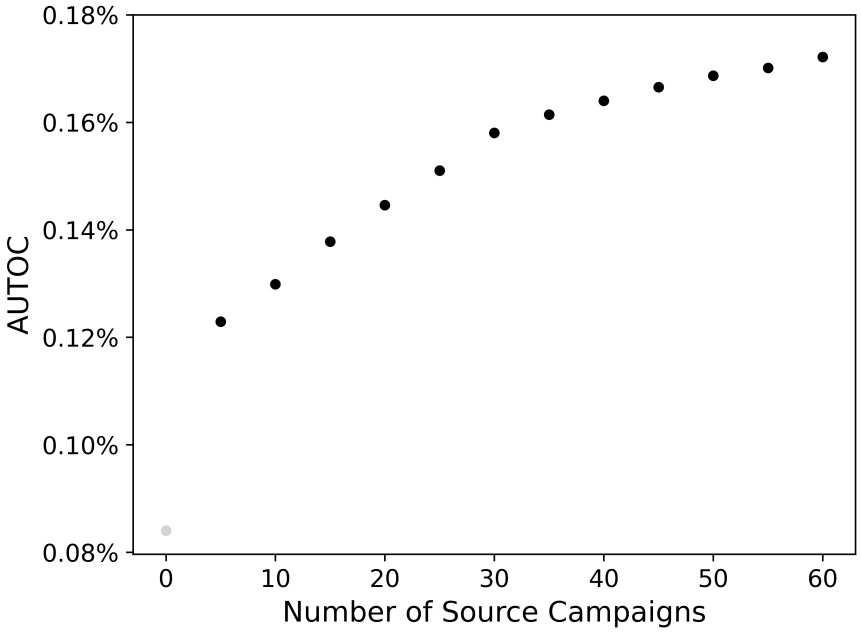
Figure 8 reports both the *inertia* for the true zip code assignments (dashed vertical line) and the distribution of *inertia* for the random assignment. The *inertia* of the true assignments is lower than the *inertia* for all but 1.4% of the 100,000 random assignments. We emphasize that the Combined Model did not have access to the zip code information during training; the model infers embeddings from the treatment effect measures for the source and focal campaigns. The evidence that neighboring zip codes are also neighbors in the embedding space indicates that the model predicts that neighboring zip codes will have a similar pattern of response across the marketing promotions.

5.4 Information Value of Source Campaigns

We finish the empirical analysis by investigating how much different source campaigns contribute to the performance of the Combined Model. We first evaluate the targeting perfor-

mance of the Combined Model with random subsets of the source campaigns in Figure 9. The x-axis indicates how many random source campaigns are used to estimate the Combined Model. The y-axis reports the AUTOC score. The left and right-most points corresponds to the performance of the Focal Only and the Combined Model in Table 1. We observe that the performance of the Combined Model improves with additional source campaigns. However, there are diminishing returns: the incremental performance improvements are larger when there are fewer source campaigns, and the curve flattens out when there are many source campaigns. In our empirical setting, incorporating 50% of the available source campaigns (30 out of 60), yields 75% of the improvement compared to the full sample.

Figure 9: Incremental Purchase Incidence with Random Source Campaigns

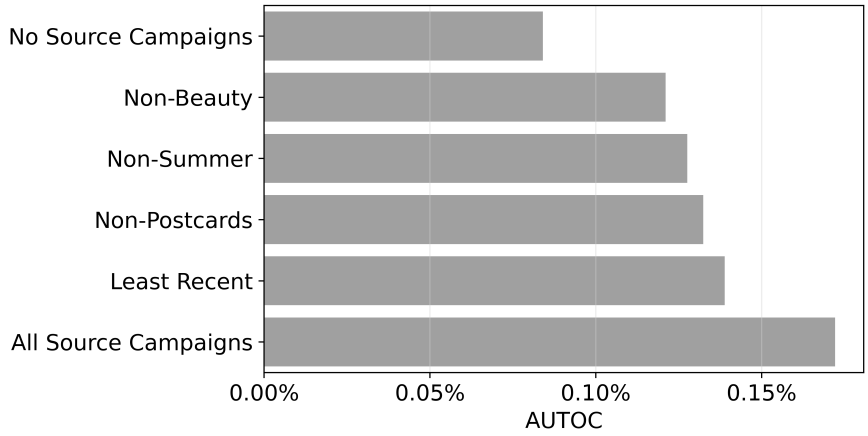


The figure reports the targeting performance of the Combined Model when estimated with random subsets of source campaigns. For each number of source campaigns, we report the average AUTOC score over 100 bootstrap iterations. The left-most point corresponds to the Focal Only policy.

Combining information across marketing campaigns improves targeting performance of the focal policy even when source campaigns are intuitively less relevant. In Figure 10, we estimate the performance of the Combined Model with subsets of source campaigns. The focal

campaign in our analysis is a Q3 Beauty Postcard. The Combined Model outperforms the Focal Only baseline even when source campaigns only include (1) non-beauty campaigns, (2) campaigns in different seasons (not summer), (3) non-postcards (catalogs and brochures), and (4) campaigns conducted over 6 months ago (least recent). The performance with seemingly less-relevant campaigns tends to be lower than when randomly choosing source campaigns (see Figure 9), but this information is still valuable for improving the focal targeting policy, and our proposed method provides a way to incorporate this source data into predictions of treatment effects.

Figure 10: Incremental Purchase Incidence with Seemingly Less Relevant Source Campaigns



The figure reports the targeting performance (AUTOC) of the Combined Model when estimated with source campaigns that are seemingly less relevant. For each evaluation, we randomly selected 10 source campaigns that satisfy the requirement.

6 Conclusion

Targeting marketing promotions is an important application for machine learning in marketing. The performance of a targeting policy depends upon how much training data is available to reliably estimate treatment effects for different customers. Traditionally, firms have used information from either the same campaign conducted in a prior period, or a pilot experiment. We propose a method that augments these data by transferring information between

marketing campaigns.

An important challenge when transferring information between campaigns is that campaigns often vary on many dimensions. This includes variation in marketing actions, variation in timing (seasonality), and variation in the customers eligible for different campaigns. These differences are often poorly documented, and it is unclear *a priori* how they contribute to variation in treatment effects.

Our proposed method addresses this challenge by summarizing information about customers and campaigns in separate embeddings. This solution offers many advantages. It does not require that the design characteristics of the campaigns are well documented. The model easily scales to handle many customers segments and many source campaigns. We explicitly account for uncertainty in the treatment effect measurements using a Bayesian framework. The method learns from source campaigns even if there is variation in which customers are eligible to receive each campaign.

We recognize both limitations to the proposed method and opportunities for future research. First, the model assumes that the customer segments are known in advance. In our empirical application the segment structure was obtained from the geographic allocation of households to zip codes. In other applications, the segment structure may not be known in advance. Future extensions to the model could focus on learning the customer segmentation along with predicting treatment effects (Kim et al., 2023; Zhang and Misra, 2024).

Second, the model assumes that there is a sample of focal data available. This data helps to locate the focal campaign in the campaign embedding space. Other approaches rely on the observable design characteristics of the campaigns and characteristics of customer segments to combine information across campaigns (see for example Huang et al., 2024). In Equation (21), we discussed incorporating segment and campaign characteristics into our approach using a linear specification, and future research could extend the proposed methods to accommodate more flexible functional forms.

With more applications, the boundaries of the proposed model can be explored more completely. In our empirical application the source and focal campaign all distributed promotions through direct mail in a luxury retail setting. Future studies could investigate the use of alternative marketing channels and larger variation in both value propositions and marketing goals. This will contribute to our understanding of the limitations of the model and identify additional opportunities for future research.

References

Adams, Ryan Prescott, George E Dahl, and Iain Murray (2010), "Incorporating side information in probabilistic matrix factorization with gaussian processes." *arXiv preprint arXiv:1003.4944*.

Anderson, Eric T, Nathan M Fong, Duncan I Simester, and Catherine E Tucker (2010), "How sales taxes affect customer and firm behavior: The role of search on the internet." *Journal of Marketing Research*, 47, 229–239.

Ascarza, Eva and Ayelet Israeli (2022), "Eliminating unintended bias in personalized policies using bias-eliminating adapted trees (beat)." *Proceedings of the National Academy of Sciences*, 119, e2115293119.

Dubé, Jean-Pierre and Sanjog Misra (2023), "Personalized pricing and consumer welfare." *Journal of Political Economy*, 131, 131–189.

Ellickson, Paul B, Wreetabrata Kar, and James C Reeder III (2023), "Estimating marketing component effects: Double machine learning from targeted digital promotions." *Marketing Science*, 42, 704–728.

Hitsch, Günter J, Sanjog Misra, and Walter W Zhang (2024), "Heterogeneous treatment effects and optimal targeting policy evaluation." *Quantitative Marketing and Economics*, 22, 115–168.

Hoffman, Matthew D, Andrew Gelman, et al. (2014), "The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo." *J. Mach. Learn. Res.*, 15, 1593–1623.

Huang, Ta-Wei, Eva Ascarza, and Ayelet Israeli (2024), "Incrementality representation learning: Synergizing past experiments for intervention personalization." *Available at SSRN 4859809*.

Kim, Mingyung, Eric Bradlow, and Raghuram Iyengar (2023), "A bayesian dual-network clustering approach for selecting data and parameter granularities." *Available at SSRN 4497834*.

Lakshminarayanan, Balaji, Guillaume Bouchard, and Cedric Archambeau (2011), "Robust bayesian matrix factorisation." In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 425–433, JMLR Workshop and Conference Proceedings.

Nesterov, Yurii (2009), "Primal-dual subgradient methods for convex problems." *Mathematical programming*, 120, 221–259.

Salakhutdinov, Ruslan and Andriy Mnih (2008), "Bayesian probabilistic matrix factorization using markov chain monte carlo." In *Proceedings of the 25th international conference on Machine learning*, 880–887.

Shapiro, Bradley T, Günter J Hitsch, and Anna E Tuchman (2021), "Tv advertising effectiveness and profitability: Generalizable results from 288 brands." *Econometrica*, 89, 1855–1879.

Simester, Duncan, Artem Timoshenko, and Spyros I Zoumpoulis (2020), "Efficiently evaluating targeting policies: Improving on champion vs. challenger experiments." *Management Science*, 66, 3412–3424.

Simester, Duncan, Artem Timoshenko, and Spyros I Zoumpoulis (2025), "A sample size calculation for training and certifying targeting policies." *Management Science*.

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager (2025), "Evaluating treatment prioritization rules via rank-weighted average treatment effects." *Journal of the American Statistical Association*, 120, 38–51.

Yang, Hongxia (2017), "Bayesian heteroscedastic matrix factorization for conversion rate prediction." In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2407–2410.

Zhang, Walter and Sanjog Misra (2024), "Coarse personalization." In *Proceedings of the 25th ACM Conference on Economics and Computation*, 1206–1208.

Zhang, Walter Wang (2024), *Optimal comprehensible targeting*. Ph.D. thesis, The University of Chicago.

Online Appendix to "Transfer Learning for Targeted Marketing: A Bayesian Matrix Factorization Approach"

A Table of Notations

Model Structure	
S	Number of customer segments.
C	Number of marketing campaigns.
m_{cs}	True treatment effect for campaign c in segment s. Matrix version of m_{cs} is M.
$\widehat{m}_{cs}$	Observed treatment effect for campaign c in segment s. Matrix version of $\widehat{m}_{cs}$ is $\widehat{\mathrm{M}}$.
λ_{cs}	Observed precision of $\widehat{m}_{cs}$. Matrix version of λ_{cs} is Λ.
q_{cs}	Binary variable identifying whether segment s is eligible for campaign c.
U_s	Segment embeddings.
V_c	Campaign embeddings.
K	Dimensionality of U_s and V_c.
λ_m	Variance of normal distribution of $p\left(m_{cs}\middle\vert U_s, V_c\right)$ (Equation 2).
X_s, X_c	Observable segment and campaign covariates (respectively).
β_U, β_V	Weights for covariates in segment and campaign embeddings (see Equation 21).
Prior Parameters	
$\Omega_U \equiv (\mu_U, \Lambda_U)$	Prior parameters for U.
$\Omega_V \equiv (\mu_V, \Lambda_V)$	Prior parameters for V.
$p(\Omega_U)$	Gaussian-Wishart priors for (μ_U, Λ_U).
$p(\Omega_V)$	Gaussian-Wishart priors for (μ_V, Λ_V).
$p(\lambda_m)$	Gamma prior for λ_m.
Model Inference	
$\widetilde{\mu}_{U_s}, \widetilde{\Lambda}_{U_s}$	Parameters of normal distribution used to estimate U.
$\widetilde{\mu}_{V_c}, \widetilde{\Lambda}_{V_c}$	Parameters of normal distribution used to estimate V.
$\widetilde{\eta}_U, \widetilde{\kappa}_U, \Lambda_U$	Parameters of normal distribution used to estimate Ω_U.
$\widetilde{W}_U, \widetilde{\nu}_U$	Parameters of Wishart distribution used to estimate Ω_U.
$\widetilde{\eta}_V, \widetilde{\kappa}_V, \Lambda_V$	Parameters of normal distribution used to estimate Ω_V.
$\widetilde{W}_V, \widetilde{\nu}_V$	Parameters of Wishart distribution used to estimate Ω_V.
α_0, β_0	Parameters of Gamma distribution used to estimate λ_m.
Simulation Analysis DGP	
$\mathcal{U}_s$, $\mathcal{V}_c$	Latent characteristics of segments and campaigns.
m_{cs}^{DGP}	True treatment effects from DGP.
λ_f	Gaussian noise component of true treatment effects (unexplained by embeddings).
λ_ϵ	Variation in individual responses.

B Closed Form Solution for Treatment Effects

To find the closed-form solution for $p(m_{cs}|U, V, \lambda_m, \widehat{M})$ in Equation (6) we use Bayes' theorem to decompose the original probability:

$$p(m_{cs}|U, V, \lambda_m, \widehat{M}) = \frac{p(m_{cs}, U, V, \lambda_m, \widehat{M})}{p(U, V, \lambda_m, \widehat{M})} = \frac{p(\widehat{M}|m_{cs}, U, V, \lambda_m)p(m_{cs}|U_s, V_c, \lambda_m)}{p(\widehat{M}|U, V, \lambda_m)} \tag{25}$$

Equation (25) depends on whether $\widehat{m}_{cs}$ is available ($q_{cs} = 1$) or unobservable ($q_{cs} = 0$).

Case 1: $q_{cs} = 0$. There is no available estimate of the treatment effect for campaign c and segment s. In this case $\widehat{M}$ does not contain $\widehat{m}_{cs}$, thus $p(\widehat{M}|m_{cs}, U, V, \lambda_m) = p(\widehat{M}|U, V, \lambda_m)$ and

$$p(m_{cs}|U, V, \lambda_m, \widehat{M}) = p(m_{cs}|U_s, V_c, \lambda_m) = \mathcal{N}\left(m_{cs}\middle|U_s'V_c, \lambda_m^{-1}\right) \tag{26}$$

Case 2: $q_{cs} = 1$. Only $\widehat{m}_{cs}$ would be impacted by conditioning on m_{cs}. Thus,

$$\begin{aligned} p(m_{cs}|U, V, \lambda_m, \widehat{M}) &= \frac{p(m_{cs}, U, V, \lambda_m, \widehat{M})}{p(U, V, \lambda_m, \widehat{M})} = \frac{p(\widehat{m}_{cs}|m_{cs})p(m_{cs}|U_s, V_c, \lambda_m)}{p(\widehat{m}_{cs}|U_s, V_c, \lambda_m)} \\ = \frac{\mathcal{N}\left(\widehat{m}_{cs}|m_{cs}, \lambda_{cs}^{-1}\right)\mathcal{N}\left(m_{cs}|U_s'V_c, \lambda_m^{-1}\right)}{\mathcal{N}\left(\widehat{m}_{cs}|U_s'V_c, (\lambda_{cs}+\lambda_m)^{-1}\right)} &= \mathcal{N}\left(m_{cs}\middle|\frac{U_s'V_c\lambda_m + \widehat{m}_{cs}\lambda_{cs}}{\lambda_{cs}+\lambda_m}, \frac{1}{\lambda_{cs}+\lambda_m}\right) \end{aligned} \tag{27}$$

After combining Equations (26) and (27) we obtain Equation (6).

C Hamiltonian Monte Carlo Algorithm

In Section 3.3, we proposed using the Hamiltonian Monte Carlo (HMC) algorithm to sample λ_m. The performance of the HMC depends on two key tuning parameters: the step size ϵ and the number of steps L.

Researchers have proposed various extensions to the standard HMC algorithm to identify appropriate values for these hyperparameters. We implemented the HMC algorithm with adaptive step-size tuning, as discussed in Hoffman et al. (2014). The core idea behind this method is to adaptively adjust the step size during the warm-up period to reach a desired

acceptance rate α^* of the sampling (Nesterov, 2009). This process requires maintaining the average "past" deviations δ^t of the realized acceptance probability α^t from the target acceptance rate α^*, as well as the average of the past step sizes. For stability, we use a weighted average of the logarithms of the step sizes. We describe our implementation of the HMC sampling in Algorithm 2.

Algorithm 2 Hamiltonian Monte Carlo Sampling

Inputs: $\lambda_m^{t-1}, \tau, \overline{\delta}, \overline{\epsilon}$
Initialize: $\epsilon^0 = 1, t_0 = 10, \kappa_0 = 0.75, \gamma_0 = 0.05, \alpha^* = 0.8$
if $t = 1$ **then** $\epsilon = \epsilon^0$
Set $L^t \leftarrow \lceil 1/\epsilon^{t-1} \rceil$
Set $\lambda \leftarrow \lambda_m^{t-1}$
Sample $\omega \sim \mathcal{N}(\omega \mid 0, 1)$
Set $\widetilde{\lambda} \leftarrow \lambda$, $\widetilde{\omega} \leftarrow \omega$
for $l = 1, \ldots, L$ **do**
    Update $\widetilde{\omega} \leftarrow \widetilde{\omega} + \frac{\epsilon}{2} \cdot \nabla \pi(\widetilde{\lambda})$.
    Update $\widetilde{\lambda} \leftarrow \widetilde{\lambda} + \epsilon \cdot \widetilde{\omega}$.
    Update $\widetilde{\omega} \leftarrow \widetilde{\omega} + \frac{\epsilon}{2} \cdot \nabla \pi(\widetilde{\lambda})$.
Set $\alpha^t \leftarrow \min\left\{1, \frac{\pi(\widetilde{\lambda})}{\pi(\lambda)} \exp\left[\frac{\omega^2 - \widetilde{\omega}^2}{2}\right]\right\}$
Sample a uniformly distributed random variable $u \sim \mathcal{U}(0, 1)$.
Set next sampled λ_m as

$$\lambda_m^t \leftarrow \begin{cases} \tilde{\lambda} & \text{if } u \leq \alpha^t, \\ \lambda & \text{otherwise.} \end{cases}$$

if $t \leq \tau$ **then**
    Set $\overline{\delta} \leftarrow \left(1 - \frac{1}{t+t_0}\right)\overline{\delta} + \frac{1}{t+t_0}\left(\alpha^* - \alpha^t\right)$
    Set $\epsilon \leftarrow 10\epsilon^0 \exp\left(-\frac{\sqrt{t}}{\gamma_0}\overline{\delta}\right)$
    Set $\log \overline{\epsilon} \leftarrow t^{-\kappa_0} \log \epsilon + (1 - t^{-\kappa_0}) \log \overline{\epsilon}$
else
    Set $\epsilon \leftarrow \overline{\epsilon}$

D Data Generating Process in Section 4

We provide a complete specification for sampling the synthetic data in Algorithm 3. In the default setting, we simulated 1,000 campaigns and 1,000 segments with two-dimensional

latent campaign and segment characteristics with the following input parameters:

- Number of segments and campaigns: $S = 1,000$ and $C = 1,000$
- Treatment effects: $\mathcal{F}(x, y) = x'y$ and $\lambda_f = 0.1$
- Measurement error: $\lambda_\epsilon = 0.001$, $N_{focal} = 10^5$, and $\overline{N}_{source} = 2,000$

Algorithm 3 Generating Simulated Data

Inputs: S, C, λ_f, $\mathcal{F}$, λ_ϵ, N_{focal}, $\overline{N}_{source}$
Initialize: $\Sigma_{\mathcal{U}} \leftarrow \left(\begin{smallmatrix} 4 & 0 \\ 0 & 4 \end{smallmatrix}\right)$, $\Sigma_{\mathcal{V}} \leftarrow \left(\begin{smallmatrix} 4 & 0 \\ 0 & 4 \end{smallmatrix}\right)$, $w_s \leftarrow 1/S$
for $s = 1$ to S **do**
  Sample $\mu_s^{\mathcal{U}} \sim$ Uniform $\left\{\left(\begin{smallmatrix} 3 \\ 3 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ 3 \end{smallmatrix}\right), \left(\begin{smallmatrix} 3 \\ 1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)\right\}$
  Sample $\mathcal{U}_s \sim \mathcal{N}\left(\mu_s^{\mathcal{U}}, \Sigma_{\mathcal{U}}\right)$
for $c = 1$ to C **do**
  Sample $\mu_c^{\mathcal{V}} \sim$ Uniform $\left\{\left(\begin{smallmatrix} 3 \\ 3 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ 3 \end{smallmatrix}\right), \left(\begin{smallmatrix} 3 \\ 1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)\right\}$
  Sample $\mathcal{V}_c \sim \mathcal{N}\left(\mu_c^{\mathcal{V}}, \Sigma_{\mathcal{V}}\right)$
for $s = 1$ to S, $c = 1$ to C **do**
  Set $\mu_{cs} = \mathcal{F}\left(\mathcal{U}_s, \mathcal{V}_c\right)$
  Sample $m_{cs}^{DGP} \sim \mathcal{N}\left(\mu_{cs}, \lambda_f^{-1}\right)$
Set $N_C \leftarrow N_{focal}$
for $c = 1$ to $C - 1$ **do**
  Sample $N_c \sim$ Exponential $\left(\overline{N}_{source}\right)$
for $s = 1$ to S, $c = 1$ to C **do**
  $\lambda_{cs} = w_s N_c \lambda_\epsilon$
for $s = 1$ to S, $c = 1$ to C **do**
  Sample $\hat{m}_{cs} \sim \mathcal{N}\left(m_{cs}^{DGP}, \lambda_{cs}^{-1}\right)$

E Policy Evaluation in the Empirical Illustration

We consider a targeting policy $\mathcal{P}_\varphi$ that allocates a marketing promotions to top-φ% of customers with the highest predicted treatment effects:

$$\mathcal{P}_{\varphi,s} = \begin{cases} 1 & \text{if } \widehat{m}_{focal,s} \geq Q_\varphi\left(\widehat{m}_{focal,1:S}\right) \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

where $\mathcal{P}_{\varphi,s}$ indicates a recommended action in segment s for the focal campaign, $\widehat{m}_{focal,s}$ is a predicted treatment effect in segment s for the focal campaign, and $Q_\varphi(x)$ is a quantile function for the input array x.

For each model and different values of $\varphi \in [0, 1]$, we compute the expected value per customer of the policy in Equation (28) using the Horvitz-Thompson estimator:

$$\mathcal{V}_\varphi = \frac{1}{N} \sum_i Y_i \cdot \frac{\mathbb{I}\left(W_i = \mathcal{P}_{\varphi,s_i}\right)}{\mathbb{P}\left(W_i = \mathcal{P}_{\varphi,s_i}\right)} \tag{29}$$

where N is a number of customer in the focal campaign, s_i indicates the segment to which customer i belongs to, and W_i is a treatment indicator for customer i in the focal campaign.

In Figure 7, we compare the performance of the targeting policies to a Random policy. For the random policy, because the ranking of the segments is randomized, each customer has φ probability of being treated regardless of segment s, and we can rewrite:

$$\mathcal{V}_\varphi^R = \varphi V_0 + (1 - \varphi) V_1 \tag{30}$$

where V_1 and V_0 represent the expected reward per customer in the Treatment and Control conditions respectively. Figure 7 reports the difference between Equations 29 and 30:

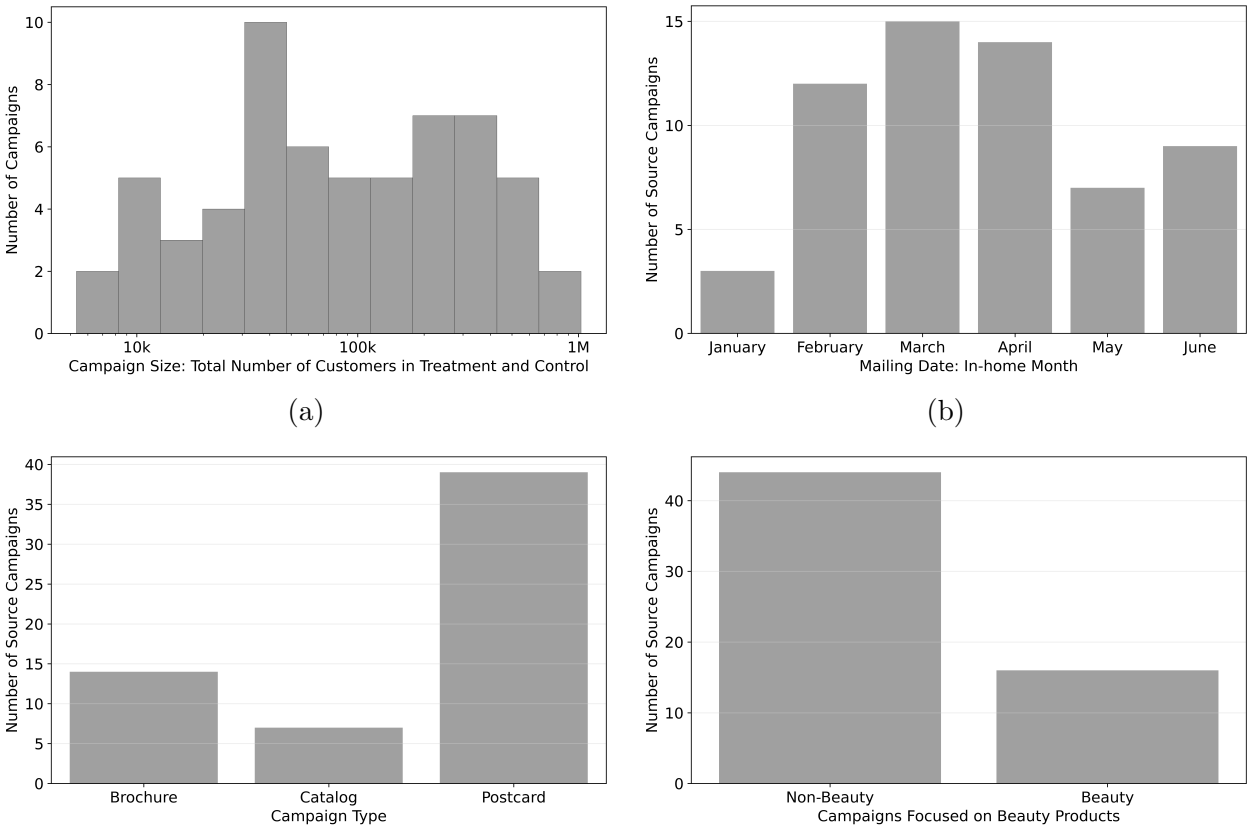
$$\mathcal{R}(\varphi) = \mathcal{V}_\varphi - \mathcal{V}_\varphi^R = \frac{1}{N} \sum_i Y_i \cdot \frac{\mathbb{I}\left(W_i = \mathcal{P}_{\varphi,s_i}\right)}{\mathbb{P}\left(W_i = \mathcal{P}_{\varphi,s_i}\right)} - \varphi V_0 - (1 - \varphi) V_1 \tag{31}$$

When $\mathcal{R}(\varphi) \geq 0$ for $\varphi \in [0, 1]$ it indicates that the ranking of customers by the policy is aligned with the ranking of treatment effects for these customers out-of-sample. We notice that for $\varphi = 1$ or $\varphi = 0$, the targeting policies (model-based and random) are identical to uniform treatment or control, so $\mathcal{R}(1) = \mathcal{R}(0) = 0$. To compare two targeting policies we can compare their respective $\mathcal{R}(\varphi)$ values for a given φ or integrate the area between the $\mathcal{R}(\varphi)$ curves for two models over the range $\varphi \in [0, 1]$.

F Campaign Summary Statistics and Randomization Check+

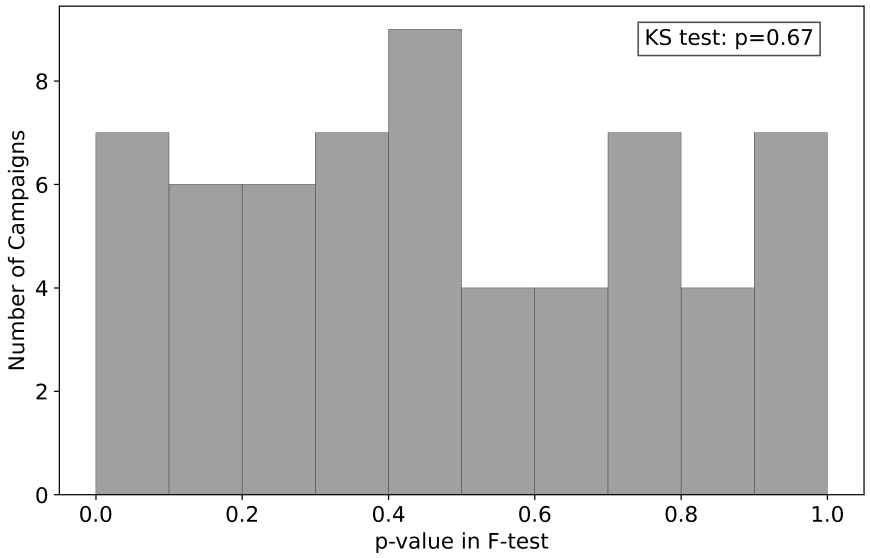
Figure 11 summarizes the characteristics of 60 source marketing campaigns in our empirical analysis.

Figure 11: Summary Statistics of Source Campaigns



We conduct randomization checks using an F-test approach. Specifically, for each marketing campaign, we estimate a linear regression where the binary treatment indicator (*mail* or *no mail*) is regressed on five pretreatment covariates. The covariates include spending per customer in the past three months, number of store visits in the past three months, number of weeks since the last purchase, share of purchases made through the online channel, and share of products returned (all measured before the in-home date for the respective campaign). We estimate a separate regression for each campaign, record the p-value of the F-test in each regression and then summarize the distribution in Figure 12. A Kolmogorov–Smirnov (KS) test does not reject the null hypothesis that the p-values are uniformly distributed; the p-value of the KS test is 0.67. This suggests that the treatment assignment is independent of pretreatment variables, supporting the validity of the randomization.

Figure 12: Distribution of p-values in the Randomization Test



The figure reports histograms of the distribution of p-values in the randomization test for each of 61 campaigns (60 source campaigns and 1 focal campaign). The y-axis represents the count of campaigns, and the x-axis represents p-values from the F-test with five pretreatment variables.