Hierarchical Clustering As a Novel Solution to the Notorious Multicollinearity Problem in Observational Causal Inference

Yufei Wu
Airbnb, Inc.
yufei.wu@airbnb.com

Zhiying Gu
Airbnb, Inc.
zhiying.gu@airbnb.com

Alex Deng
Airbnb, Inc.
alex.deng@airbnb.com

Jacob Zhu
Airbnb, Inc.
jacob.zhu@airbnb.com

Linsha Chen
Airbnb, Inc.
linsha.chen@airbnb.com

ABSTRACT

Multicollinearity is a long lasting challenge in observational causal inference, especially under regressional settings — highly correlated independent variables make it difficult to isolate their individual impacts on outcomes of interest. While common solutions such as shrinkage estimators, principal component regressions, and partial linear regression are helpful in prediction problems, a crucial limitation hinders their applicability to causal inference problems — they cannot provide the original causal relationships. To fill the gap, we present an innovative and intuitive solution, by employing hierarchical clustering to aggregate data in a way that effectively alleviates collinearity. This method is generally applicable to causal problems featuring multicollinearity. We use a marketing application to demonstrate how and why it works.

Expenditures on different advertising channels often exhibit correlations, making it exceedingly difficult to separately measure their impact. Many previous studies proposed to leverage granular cross-sectional data for better identification but, to our knowledge, none explicitly addressed multicollinearity, which undermines causal identification even with granular data. We propose to hierarchically cluster geographic units based on marketing spend correlation to reduce collinearity, and to implement a Bayesian Marketing Mix Model with cluster-level data. Such clustering happens in two steps — we first normalize and demean geo-level data to establish a common scale and to eliminate the common trends; we then calculate pairwise distance to summarize marketing spend correlation between geos and cluster the ones with moderate to strong correlation. Both descriptive evidence and regression analysis affirm that such hierarchical clustering effectively mitigates collinearity and facilitates the separate identification of the impact of different marketing channels.

KEYWORDS

Marketing Mix Model, Hierarchical Bayesian, Hierarchical Clustering, Multicollinearity

1 INTRODUCTION

Everyday business inquiries frequently revolve around causal inference, specifically seeking to understand the impact of particular business decisions. To address this, three common approaches are typically employed: A/B testing, quasi-experimentation, and observational causal inference methods. While A/B testing and quasi-experimentation are often preferred due to their ability to provide exogenous variation for identification, their implementation can be prohibitively costly or subject to biases resulting from business or technical constraints. Observational causal inference methods, such as matching methods, synthetic control, and double machine learning are designed to mitigate biases, but are not applicable to some questions we aim to address given the data properties. Furthermore, the aforementioned approaches are more effective in measuring the causal impact of single interventions rather than attributing causal impact holistically across multiple interconnected factors that may contribute to the final outcome. In such scenarios, regression approaches provide a more suitable alternative. Regression methods only necessitate aggregated panel data to concurrently identify the causal impact of multiple factors. However, two challenges undermine our ability to confidently affirm that the estimated parameters from regression methods represent the true causal impact. These challenges are the existence of confounding factors and multicollinearity among covariates. While common solutions such as shrinkage estimators, principal component regressions, and partial linear regression are helpful in prediction problems, a crucial limitation hinders their applicability to causal inference problems — they cannot provide the original causal relationships.

This paper introduces a novel approach that specifically addresses the second challenge, namely multicollinearity. To illustrate the practical application and effectivenss of this approach, we demonstrate its implementation in a marketing measurement context (Marketing Mix Marketing) at Airbnb. We also conclude this paper with a discussion on the broad applicability of this approach.

In marketing, one question of paramount importance is to causally attribute sales to spend across channels - such as Google Search, YouTube, Display, etc. However, advertisers often allocate their expenditures across ad channels in a correlated manner, particularly during peak seasons. When attempting to estimate a regression model, highly correlated variables result in larger estimate variances and imprecise attribution of channel contributions to sales. It is not uncommon to observe regression coefficients switching signs when highly correlated inputs are introduced, consequently undermining the confidence of business stakeholders in the model results.

In this marketing application, we have access to panel data consisting of ad impressions categorized by channel and geographic location (Designated Market Area, or DMA) over a specific time period. When we analyze the data by pooling all geographic locations together, we observe a high level of cross-channel correlation. However, it is worth noting that certain geographic locations exhibit higher cross-channel correlations compared to others. To address

the issue of multicollinearity, we propose a novel approach that leverages the variations in correlation patterns across different geographic locations. The objective is to restructure the data in a way that significantly reduces the multicollinearity problem. Our proposed method involves utilizing hierarchical clustering to group geographic locations based on their correlation patterns. The key aspect of this approach lies in defining the distance metric used in the clustering algorithm.

In our methodology, we define the distance between two DMAs as the sum of channel-specific distances. Each channel-specific distance measures the similarity in the cross-channel correlation between the two geographic locations. By incorporating this distance metric into the hierarchical clustering process, we can effectively group the DMAs in a manner that minimizes multicollinearity across channels. This innovative approach allows us to transform the data structure, mitigating the challenges posed by multicollinearity and providing a more robust foundation for further analysis. By adopting this methodology, we can improve our understanding of the causal relationships between channels and accurately attribute their impact on sales or other relevant outcomes. We will demonstrate this improvement with both data descriptive evidence and regression results.

The remainder of this paper is organized as follows. In section 2, we describe the Marketing Mix Modelling (MMM) problem formulation and the related work. In section 3, we present the data properties that motivate our approach to reduce multicollinearity. In section 4, we introduce Hierarchical Clustering and the distance metric designed specifically to address the multicollinearity problem. In section 5, we show how this novel method improves results, and in section 6, we briefly discuss other applications that can utilize this methodology.

2 BAYESIAN STRUCTURAL MODEL FORMULATION

This paper focuses on providing an innovative and intuitive solution to the notorious multicollinearity problem. To demonstrate the effectiveness of their approach, we apply it to a specific application called Bayesian Marketing Mix Modeling (MMM), built upon Jin et al. [10]. MMM is a widely applied method in the industry for estimating the performance of various marketing channels in a holistic manner. It takes into account factors such as seasonality, trend (representing organic demand), and mix of different marketing channels when forecasting sales. [1]

2.1 Marketing Mix Model Setup

We model sales as a non-linear function of seasonality, and advertisement impressions of each channel with a Bayesian Model. Let g denote a DMA and $t = 1, ..., T$ denote time (we use weekly data).

$$y_{g,t} = \mu t^{\lambda} + seasonality_{g,t} + \alpha Z_{g,t} + \sum_{k=1}^{K} \beta_k AdStock(x_{k,g,t}) + \epsilon_{g,t}$$

There are K media channels, and G DMAs. $x_{g,t}$ is the impression of channel k at week t. Let $y_{g,t}$ be the response variable at week t, which could be sales or log transformed sales. We include μt^k, $seasonality_{g,t}$, and contemporaneous correlation with covariates $Z_{g,t}$ to capture the evolution of organic sales over time. $AdStock(x_{g,t})$ is the transformed impressions that captures: (1) diminishing return; (2) lag of the effect; (3) carryover effect of the impressions. Ng et al. [13] uses a different formulation which only estimates the saturation effect. In this paper, we adapt Google's proposed shape formulation of marketing effects that is more flexible.[10] The $AdStock$ function can be defined as:

$$AdStock_{k,g} = \left(\frac{\sum_{l=0}^{L} \tau_k^{(l-\theta_k)^2} x_{t-l,m}}{\sum_{l=0}^{L} \tau_k^{(l-\theta_k)^2}} \right)^{\rho}$$

$\rho \in (0, 1]$ captures (potentially diminishing) returns to scale; $\tau \in (0, 1)$ governs the carryover rate over time; and θ indicates the lagged peak effect.

2.2 Account for Confounding factors

While it is not the main focus of this paper, we take into account confounding factors when modeling trend and seasonality in order to properly capture organic demand. When modeling trend and seasonality, it is crucial to strike the right balance between flexibility and strictness – excessive flexibility may lead to overfitting, whereas overly rigid parametric formulations can result in a poor fit of the model. In this marketing use case, we can easily overfit a model that performs poorly out of sample because we have high dimensional parameters space - we have to estimate 4 parameters per channel. Keeping this tradeoff in mind, in addition to including exponential trend and sinusoidal seasonality following Jin et al. [10], we also include as an additional covariate an index of Google Search query volume for travel and accommodation brands excluding Airbnb ($Z_{g,t}$ in the above equation), to capture confounding factors that affect organic demand contemporaneously.

3 DATA PROPERTIES

3.1 Pre-Process Data

We take two important steps to pre-process data in preparation for descriptive analysis and modeling. First, we normalize bookings, channel impressions, and the covariate to establish a common scale across DMAs of different sizes. This will make it easier (A) to interpret the impact of a certain level of marketing activity; and (B) to model the common trend and seasonality later. Second, we decompose channel impressions into the common trend and seasonality and residual variation across DMAs, so we can focus on correlation in the residual variation next.

3.2 Characteristics of DMA-Level Data

There is a decent level of correlation between marketing channels and DMAs, even after eliminating the common trend and seasonality across DMAs for each channel. As Figure 1 shows, the variation in residual impressions is moderately to strongly correlated across the five channels.[2]

Such correlation is more pronounced across some DMAs than other DMAs. Figure 2 compares two sets of DMAs – DMAs in the

[1] While experimentation can be used to measure the performance of some channels, it is not always feasible due to practical constraints.

[2] We anonymize the channels as A, B, C, D, and E.

Figure 1: Correlation of Residual Channel Impressions

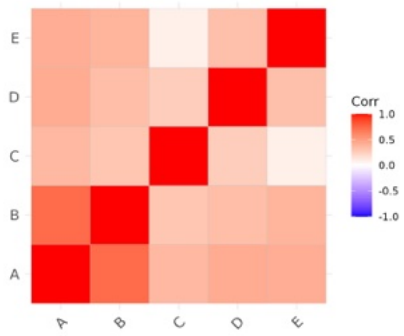
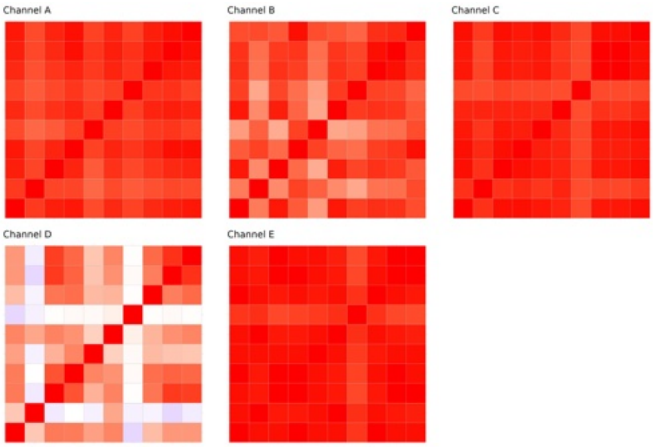
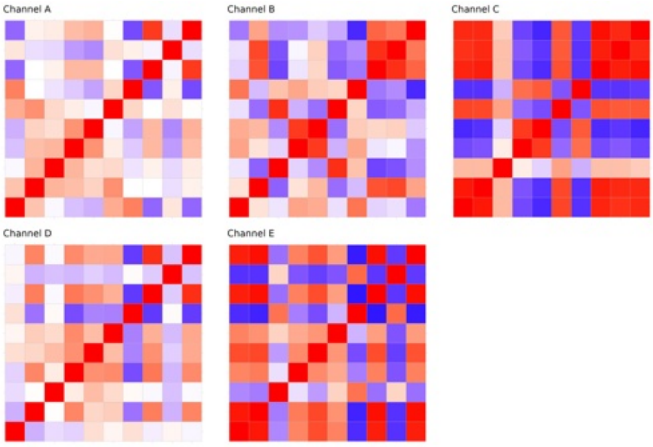


Figure 2: Correlation of Residual Impressions Across DMAs

(a) Correlation Across DMAs in the Xth Ventile of Size: By Channel



(b) Correlation Across DMAs in the Yth Ventile of Size: By Channel



Xth ventile of baseline sales exhibit extremely high correlation for four out of the five marketing channels, while DMAs in the Yth ventile exhibit relatively little correlation for all channels.[3]

[3]Throughout this paper, DMA IDs and channel names have been anonymized, while channel impressions have been indexed.

4 HIERARCHICAL CLUSTERING AS A NOVEL SOLUTION TO MULTICOLLINEARITY

As overviewed in Section 1 and illustrated in Section 3, multicollinearity poses a fundamental challenge in separately measuring the impact of different marketing channels. While common solutions such as shrinkage estimators, principal component regressions, and partial linear regression are helpful in prediction problems, a crucial limitation hinders their applicability to causal inference problems — they cannot provide the original causal relationships for business interpretability.

To overcome this limitation, we propose a novel and intuitive approach that defines distances and hierarchically clusters geographic areas in a way that effectively mitigates cross-channel multicollinearity. We first calculate a pairwise distance or dissimilarity metric to summarize marketing spend correlation between geos and then use that metric to cluster the geos with moderate to strong correlation. For each channel k, we calculate the distance between two DMAs i and j as follows, where X_{ik} denotes the time series of residual impressions, after eliminating the common trend and seasonality, for channel k in DMA i.

$$Distance_{ijk} = 1 - Correlation(X_{ik}, X_{jk})$$

We then calculate an overall distance across all channels, which is the square-root of the sum of squared distances across the channels.

$$Distance_{ij} = \sqrt{\sum_{k=1}^{K} Distance_{ijk}^2}$$

This distance measure reflects correlation between DMAs across multiple channels and is used to hierarchically cluster DMAs. We adopt a complete-linkage hierarchical clustering algorithm which works as follows:[11][14]

(1) Start with assigning each DMA to its own cluster;
(2) Then proceed iteratively, joining the two most similar clusters at each step, continuing until there is just a single cluster. Distance or dissimilarity between two clusters is based on the farthest pair.

This algorithm produces a dendrogram in Figure 3, which illustrates how DMAs are clustered at each step. The horizontal axis lays out the DMAs while the vertical axis shows the distance. This algorithm offers a lot of flexibility in how aggressively we want to cluster DMAs or how many clusters we want to have – we can pick any cutoff distance between 0 (i.e., each DMA in a separate cluster) and 4 (i.e., all DMAs in one cluster).

We use a cutoff distance of 1.5, which corresponds to correlation of at least 0.33 on average for each channel and produces 42 clusters, but also consider alternative clustering strategies for sensitivity. Intuitively speaking, we group DMAs that feature moderate to strong correlation into the same cluster.

5 RESULTS

5.1 Hierarchical Clustering of DMAs

The hierarchical clustering approach produces intuitive results. We visualize channel impressions over time across DMAs within each cluster, confirming that DMAs within the same cluster tend to have highly correlated impressions over time for at least some channels.

Figure 3: Dendrogram Illustrating Hierarchical Clustering of DMAs

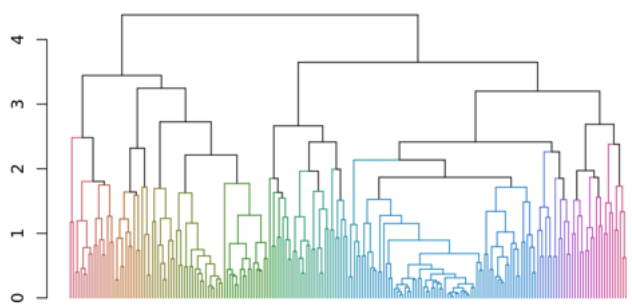


Figure 4 exemplifies such patterns using one small cluster (Cluster 1) and one larger cluster (Cluster 4). Each chart illustrates the variation in channel residual impression over time (the horizontal axis) and across DMAs (the vertical axis). Within each cluster, the color patterns over time are quite similar across DMAs for most channels, reflecting moderate to high correlation.

5.2 Descriptive Evidence Shows Clustering Mitigates Multicollinearity

Descriptive evidence confirms our intuition that hierarchical clustering can effectively mitigate collinearity. By grouping moderately to highly correlated DMAs into the same cluster, we have significantly reduced correlation in residual channel impressions. As Figure 5 visualizes, correlation decreased generally across channels, by 8% to 43%.

Further, as Figure 6 demonstrates, clustering preserves variation in channel impressions both (A) within clusters over time and (B) across clusters within the same time period. This is promising for separately identifying the impact of different channels using panel data at the cluster-week level. When testing alternative clustering strategies, we consider the reduction in correlation and the preservation of variation as two important criteria.

5.3 Regression Results Confirm Clustering Alleviates Multicollinearity

Panel linear regression analysis affirms the effectiveness of this hierarchical clustering method in mitigating collinearity and facilitating the separate identification of the impact of different marketing channels. After clustering, channel coefficient estimates are no longer subject to the problem of flipped signs when included together with other channels, and instead produce mostly intuitive results.

Table 1 summarizes panel linear regression results before and after clustering, with geo (DMA or cluster) fixed effects and week fixed effects included throughout the different specifications.[4] As Column 1 summarizes, most channel coefficients are negative if we use DMA-level data, while they would be positive if included individually. After clustering, in Column 2, the results become mostly intuitive – all four lower-funnel channels have positive estimated impact on sales (three of which are significant at 0.001

[4]All coefficient estimates have been scaled by a constant.

Figure 4: Heat Maps of Channel Residual Impressions Across DMAs Over Time

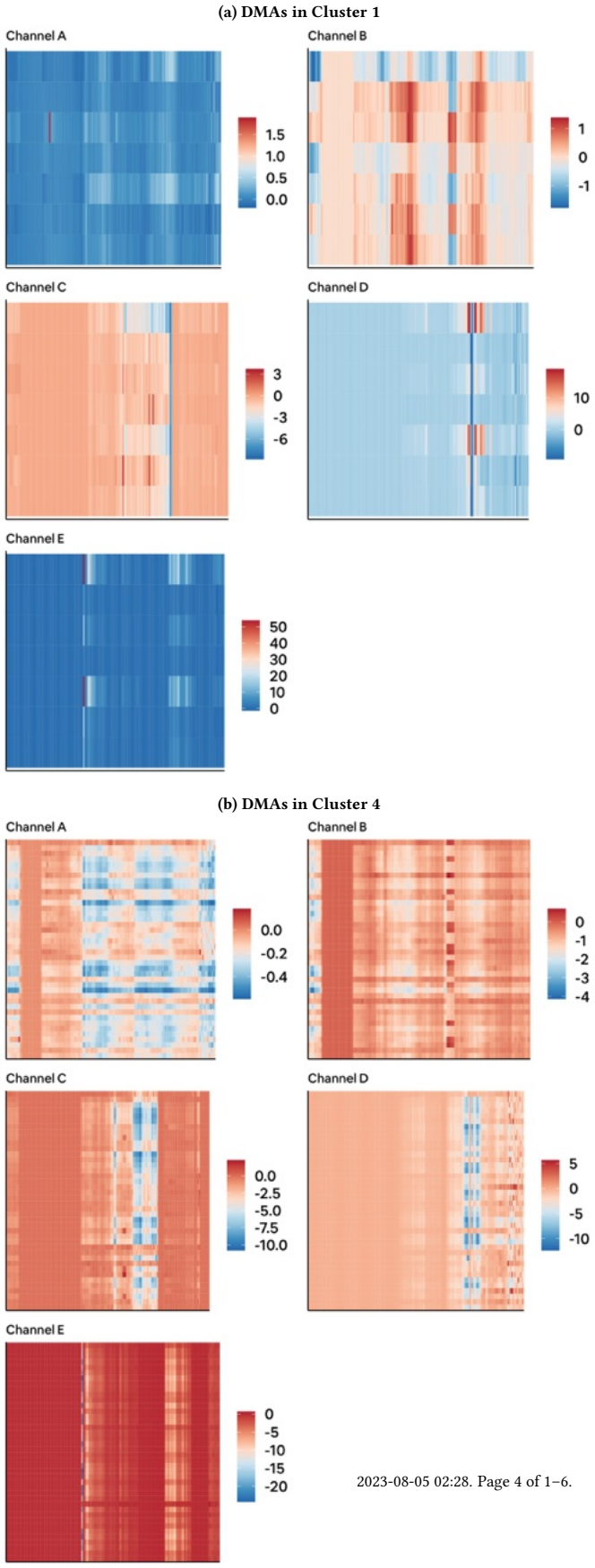


Figure 5: Clustering Reduces Cross-Channel Correlation

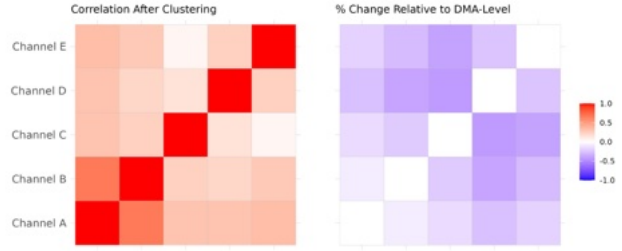
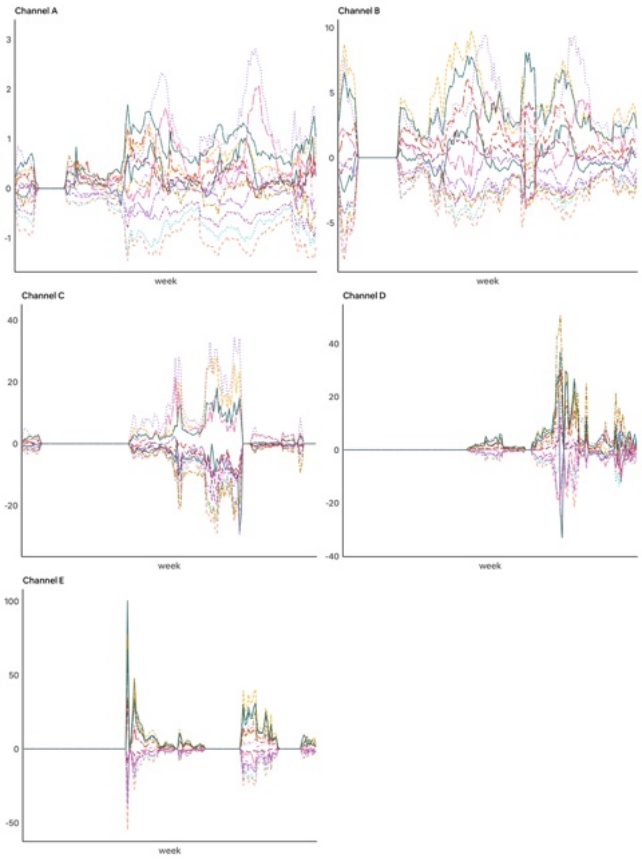


Figure 6: Variation in Residual Channel Impressions Across Clusters Over Time



level). The remaining channel with a negative coefficient is upper-funnel, where we expect extreme difficulty in detecting a lower funnel impact. Finally, these findings are robust to weighting the cluster based on the natural log of their baseline size.

5.4 Bayesian Model Results Using Cluster-Level Data

Now that both descriptive evidence and frequentist regression analysis affirm our hierarchical clustering approach effectively mitigates collinearity, we move forward to estimate the Bayesian model described in Section 2 using data at the cluster level. Similar to the

Table 1: Panel Linear Regression Summary

	DMA Level	Cluster Level	Cluster Level, Weighted
channel_a	5.570***	5.041***	5.227***
	(0.075)	(0.149)	(0.146)
channel_b	−0.055***	0.128***	0.097***
	(0.014)	(0.028)	(0.027)
channel_c	−0.003	0.043***	0.038***
	(0.004)	(0.008)	(0.007)
channel_d	0.007	0.007	0.009
	(0.005)	(0.010)	(0.010)
channel_e	−0.011***	−0.023**	−0.025***
	(0.003)	(0.008)	(0.007)
Num.Obs.	35 700	7140	7140
R2	0.862	0.946	0.946
R2 Adj.	0.861	0.944	0.944
AIC	113 008.3	34 358.7	48 912.9
BIC	114 492.8	35 561.6	50 115.8
RMSE	1.17	2.62	7.27

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

frequentist regressions, the Bayesian model also produces intuitive results, even with uninformative priors. Figure 7 visualizes the posterior distribution of channel-specific impact parameters (β) and carryover rate parameters (τ). Consistently with frequentist regression results earlier, we estimate a higher impact for channels A and B than the other channels.[5] Furthermore, the carryover estimates are also intuitive and consistent with our previous learning and knowledge about the different channels. For example, we expect some carryover for Channels C, D, and E, but not for Channels A and B, and it is reassuring that the estimates confirm that understanding even though we are using uninformative priors.[6]

6 DISCUSSION OF BROADER APPLICATIONS

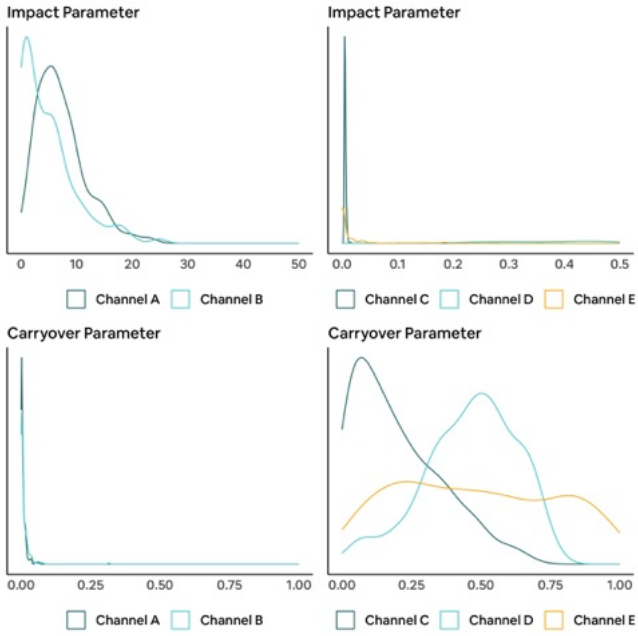
We have focused on an application in marketing mix modeling to demonstrate how and why hierarchical clustering can mitigate multicollinearity. However, this method is not constrained to the setting of marketing and instead is generally applicable to observational causal inference problems featuring multicollinearity. In our example, marketing data properties motivated us to cluster geographic units based on correlation in marketing activities. In other settings, one can decide which dimensions and criteria to use for clustering based on relevant data properties. The dimension to cluster data does not always have to be geographic. Futher more, in some settings, natural clusters might exist for one to consider.

For example, in the context of Customer Service, we would like to understand how customers' each interaction with our support agents contribute to the long term retention. However, oftentimes, these interaction experience metrics are highly correlated, such

[5]Note that the impact parameter estimates here should be interpreted differently from the frequentist panel linear regressions above, because the Bayesian structural model also estimates parameters that transforms the impressions for each channel into adstock based on the lag, carryover, and shape parameters. But we can still make broad comparisons of the impact parameter across channels, taking into account the other parameters.

[6]It is expected that the variance is high for the estimates of Channel E, such as the carryover parameter. Unlike the other channels, Channel E is upper funnel, so it is especially difficult to estimate its impact on lower funnel conversions.

Figure 7: Posterior Distributions of Parameters



as wait time and abandon rate, etc. In this case, we can leverage clustering to segment customer issue types into groups that have different degrees of correlation between wait and abandon rate.

7 CONCLUSION

In this paper, we propose to employ hierarchical clustering as an innovative and effective approach to address multicollinearity in regressional causal inference studies. It has several advantages. Firstly, hierarchical clustering provides a systematic and comprehensive method for identifying clusters that exhibit varying levels of multicollinearity, thus reducing the correlation of covariates across clusters. Furthermore, clustering circumvents the need to transform data into non-interpretable entities, as required by techniques such as Principal Component Analysis or Partial Linear Regressions. This ensures that the interpretability and meaningfulness of the variables are preserved throughout the analysis. In addition to its effectiveness, the proposed methodology is characterized by its ease of implementation. It can be readily applied to diverse applications facing similar challenges related to multicollinearity. The key lies in understanding the inherent properties of the data to define an appropriate distance metric for clustering that effectively reduces multicollinearity. This research contributes to enhancing the robustness and reliability of regressional causal inference studies.

8 ACKNOWLEDGEMENT

The authors would like to thank Carolina Barcenas (Airbnb, Inc.), Mike Anderson (Google LLC), Fei Cen (Google LLC) for their help and guidance on this project.

REFERENCES

[1] Ron Berman. 2018. Beyond the last touch: Attribution in online advertising. *Marketing Science* 37, 5 (2018), 771–792.
[2] Thomas Blake, Chris Nosko, and Steven Tadelis. 2015. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* 83, 1 (2015), 155–174.
[3] David Chan and Mike Perry. 2017. Challenges and opportunities in media mix modeling. (2017).
[4] Hao Chen, Minguang Zhang, Lanshan Han, and Alvin Lim. 2021. Hierarchical marketing mix models with sign constraints. *Journal of Applied Statistics* 48, 13-15 (2021), 2944–2960.
[5] Jamal I Daoud. 2017. Multicollinearity and regression analysis. In *Journal of Physics: Conference Series*, Vol. 949. IOP Publishing, 012009.
[6] Ruihuan Du, Yu Zhong, Harikesh Nair, Bo Cui, and Ruyang Shou. 2019. Causally driven incremental multi touch attribution using a recurrent neural network. *arXiv preprint arXiv:1902.00215* (2019).
[7] Wolfgang Härdle, Hua Liang, and Jiti Gao. 2000. *Partially linear models*. Springer Science & Business Media.
[8] Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: applications to nonorthogonal problems. *Technometrics* 12, 1 (1970), 69–82.
[9] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press. https://doi.org/10.1017/CBO9781139025751
[10] Yuxue Jin, Yueqing Wang, Yunting Sun, David Chan, and Jim Koehler. 2017. Bayesian methods for media mix modeling with carryover and shape effects. (2017).
[11] Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, 1 (2012), 86–97.
[12] Sridhar Narayanan and Kirthi Kalyanam. 2014. *Position effects in search advertising: A regression discontinuity approach.* Technical Report. Working paper.
[13] Edwin Ng, Zhishi Wang, and Athena Dai. 2021. Bayesian Time Varying Coefficient Model with Applications to Marketing Mix Modeling. *arXiv preprint arXiv:2106.03322* (2021).
[14] Chandan K Reddy and Bhanukiran Vinzamuri. 2018. A survey of partitional and hierarchical clustering algorithms. In *Data clustering*. Chapman and Hall/CRC, 87–110.
[15] Michael Thomas. 2020. Spillovers from mass advertising: An identification strategy. *Marketing Science* 39, 4 (2020), 807–826.
[16] Jon Vaver and Stephanie Shin-Hui Zhang. 2017. Introduction to the Aggregate Marketing System Simulator. (2017).
[17] Yueqing Wang, Yuxue Jin, Yunting Sun, David Chan, and Jim Koehler. 2017. A hierarchical Bayesian approach to improve media mix models using category data. (2017).
[18] Michael J Wolfe Sr and John C Crotts. 2011. Marketing mix modeling for the tourism industry: A best practices approach. *International Journal of Tourism Sciences* 11, 1 (2011), 1–15.