The Annals of Applied Statistics
2025, Vol. 19, No. 2, 866–888
https://doi.org/10.1214/24-AOAS1994

ESTIMATING HETEROGENEOUS CAUSAL EFFECTS OF HIGH-DIMENSIONAL TREATMENTS: APPLICATION TO CONJOINT ANALYSIS

BY MAX GOPLERUD[1,a], KOSUKE IMAI[2,b] AND NICOLE E. PASHLEY[3,c]

[1]*Department of Government, University of Texas at Austin,* [a]*mgoplerud@austin.utexas.edu*
[2]*Department of Government and Department of Statistics, Harvard University,* [b]*imai@harvard.edu*
[3]*Department of Statistics, Rutgers University,* [c]*nicole.pashley@rutgers.edu*

Estimation of heterogeneous treatment effects is an active area of research. Most of the existing methods, however, focus on estimating the conditional average treatment effects of a single, binary treatment given a set of pretreatment covariates. In this paper we propose a method to estimate the heterogeneous causal effects of high-dimensional treatments, which poses unique challenges in terms of estimation and interpretation. The proposed approach finds maximally heterogeneous groups and uses a Bayesian mixture of regularized logistic regressions to identify groups of units who exhibit similar patterns of treatment effects. By directly modeling group membership with covariates, the proposed methodology allows one to explore the unit characteristics that are associated with different patterns of treatment effects. Our motivating application is conjoint analysis, which is a popular type of survey experiment in social science and marketing research and is based on a high-dimensional factorial design. We apply the proposed methodology to the conjoint data, where survey respondents are asked to select one of two immigrant profiles with randomly selected attributes. We find that a group of respondents with a relatively high degree of prejudice appears to discriminate against immigrants from non-European countries, like Iraq. An open-source software package is available for implementing the proposed methodology.

1. Introduction. Over the past decade, a number of researchers have exploited modern machine learning algorithms and proposed new methods to estimate heterogeneous treatment effects using experimental data. They include tree-based methods (e.g., Imai and Strauss (2011), Athey and Imbens (2016), Wager and Athey (2018), Hahn, Murray and Carvalho (2020)), regularized regressions (e.g., Imai and Ratkovic (2013), Tian et al. (2014), Künzel et al. (2019)), ensemble methods (e.g., van der Laan and Rose (2011b), Grimmer, Messing and Westwood (2017)), and frameworks that allow for the use of generic machine learning methods (e.g., Chernozhukov et al. (2023), Imai and Li (2025)). This methodological development, however, has largely been confined to settings with a single, binary treatment variable; some exceptions include a time-varying treatment (e.g., Almirall et al. (2014)) and a relatively small number of treatments (e.g., Imai and Ratkovic (2013)).

In this paper we estimate the heterogeneous effects of a *high-dimensional* treatment by analyzing the data from conjoint experiments in which the number of possible treatment combinations exceeds the sample size. While the high dimensionality in treatment effect heterogeneity problems typically comes from the number of covariates or moderators, conjoint experiments provide an additional difficulty due to high dimensionality of treatment. We address the methodological challenge of effectively summarizing the complex patterns of heterogeneous treatment effects that are induced by the interactions among the treatments themselves as well as the interactions between the treatments and unit characteristics.

Received June 2024; revised November 2024.
Key words and phrases. Causal inference, factorial design, mixture model, randomized experiment, regularized regression.

Methodological contributions. We consider a common setting where researchers wish to use a small number of groups to summarize heterogeneous treatment effects and characterize these groups using several pretreatment covariates (e.g., Chernozhukov et al. (2023), Imai and Li (2025)). We show that once researchers select the number of groups to be used for summarizing heterogeneous treatment effects, finding the maximally heterogeneous groups in terms of potential outcomes is equivalent to maximizing the likelihood function based on the latent group membership. Furthermore, modeling the conditional probability of an individual's latent group membership using the moderators of interest yields maximally heterogeneous groups that are predicted well by these moderators.

A primary methodological challenge with high-dimensional treatments is characterizing both the interactions among a large number of treatment variables and their relationships with moderating covariates. Our methodology addresses this by finding maximally heterogeneous groups while characterizing the relationship between group membership and unit characteristics. Thus, it is possible to understand the types of units that are likely to exhibit similar treatment effect patterns.

Since optimizing over the latent group membership is difficult, we marginalize it out, leading to a mixture of experts model (e.g., Gormley and Frühwirth-Schnatter (2019), Gupta and Chintagunta (1994)). We also develop estimation strategies by bringing together two previously disconnected literatures, one on mixture models and the other on sparsity-inducing penalties to fuse factor levels.

Empirical application. Conjoint analysis is a popular survey experimental methodology in social sciences and marketing research (e.g., Hainmueller, Hopkins and Yamamoto (2014), Rao (2014)). Conjoint analysis is a variant of factorial designs (Dasgupta, Pillai and Rubin (2015)) with a large number of factorial treatments—so large that typically not all possible treatments are observed. Under the most commonly used "forced-choice" design, respondents are asked to evaluate a pair of profiles whose attributes are randomly selected based on factorial variables with several levels.

In the specific experiment we reanalyze, the original authors used a conjoint analysis to measure immigration preferences by presenting each survey respondent with several pairs of immigrant profiles with varying attributes including education, country of origin, and job experience (Hainmueller and Hopkins (2015)). For each pair the respondent was asked to choose one profile they prefer. The authors then analyzed the resulting response patterns to understand which immigrant characteristics play a critical role in forming the immigration preferences of American citizens.

In the methodological literature on factorial designs and conjoint analysis, researchers have focused on average marginal effects, which represent the average effect of one factor level relative to another level of the same factor averaging over the randomization distribution of the remaining factors (Hainmueller, Hopkins and Yamamoto (2014), Dasgupta, Pillai and Rubin (2015)). Many empirical researchers use subgroup analysis to explore how these marginal effects depend on a small number of moderating covariates (e.g., Hainmueller and Hopkins (2015), Newman and Malhotra (2019)).

Unfortunately, such an approach often results in low statistical power and may suffer from multiple testing problems (Liu and Shiraito (2023)). More fundamentally, by marginalizing other treatments, researchers may miss important interactions among treatments. Although some have explored the estimation of interaction effects (e.g., Dasgupta, Pillai and Rubin (2015), Egami and Imai (2019), De la Cuesta, Egami and Imai (2022)), few have investigated how to estimate heterogeneous treatment effects of high-dimensional treatments.

Moreover, there is even less prior research that models how the effects of high-dimensional treatments vary as a function of moderators. One exception is Robinson and Duch (2024), which uses a BART-based approach for conjoint experiments, but their heterogeneous effects of interest are different from ours (see Section 5.4 for comparison).

Related models. To overcome this challenge, we develop a mixture of regularized logistic regression models under our general methodological framework of treatment effect heterogeneity with high-dimensional treatments. We combine and extend two distinct strands of methodological research. First, a growing literature explores regularization with high-dimensional factors, and their interactions, by fusing or grouping levels of factors together (e.g., Bondell and Reich (2009), Post and Bondell (2013), Stokell, Shah and Tibshirani (2021)). This methodology is well suited to factorial experiments because it provides a natural way of interpreting empirical findings by identifying a set of factor levels that characterize distinct treatment effects (e.g., Egami and Imai (2019)).

However, since our goal is to identify groups of individuals with heterogeneous effects, we use a mixture model that finds the maximally heterogeneous groups (see Section 3.2). Although the marketing literature has long applied mixture models to analyzing heterogeneity in conjoint experiments (e.g., Gupta and Chintagunta (1994), Andrews, Ainslie and Currim (2002)), they focused on settings with low-dimensional treatments. In the high-dimensional setting, some combine mixture models with sparsity constraints (e.g., Khalili and Chen (2007), Städler, Bühlmann and Van De Geer (2010), Khalili (2010)), but these constraints are not designed to induce the fusion of factor levels that is essential in conjoint analysis.

Our model, therefore, synthesizes both of these approaches by using a finite mixture model with a prior that encourages fusing levels, while respecting the hierarchical structure—fusing main effects of factors only if their interactions are also fused (Yan and Bien (2017)). For efficient computation we develop an expectation–maximization (EM) algorithm (Dempster, Laird and Rubin (1977)) by exploiting the representation of ℓ_1 and ℓ_2 penalties as a mixture of Gaussians (e.g., Figueiredo (2003), Polson and Scott (2011), Ratkovic and Tingley (2017), Goplerud (2021)). We derive a tractable algorithm that adapts the latent overlapping group LASSO developed in sparse modeling to fusion required in factorial experiments.

The rest of the paper is organized as follows. In Section 2 we discuss the motivating application, which is a conjoint analysis of American citizens' preferences regarding immigrant features. We also briefly describe a methodological challenge to be addressed. In Section 3 we present our proposed methodology. In Section 4 we show our method performs well in a realistic numerical simulation. In Section 5 we apply this methodology and reanalyze the data from the motivating conjoint analysis. Section 6 concludes with a discussion. The R package `FactorHet` (Goplerud, Pashley and Imai (2025)) can be used to implement our methodology, and Goplerud, Imai and Pashley (2025a) provide replication code for our application and simulations.

2. Motivating application: Conjoint analysis of immigration preferences.

Our motivating application is a conjoint analysis of American immigration preferences. In this section we introduce the experimental design and discuss the results of previous analyses that motivate our methodology for estimating heterogeneous treatment effects.

2.1. *The experimental design*.

In an influential study, Hainmueller and Hopkins (2015) use conjoint analysis to estimate the effect of immigrant attributes on preferences for admission to the United States (data are available at the AJPS Dataverse https://doi.org/10.7910/DVN/25505). The authors conduct an online survey experiment using a sample of 1407 American adults. Each survey respondent assessed five pairs of immigrant profiles with randomly selected attributes. For each pair a respondent was asked to choose which of the two immigrant profiles they preferred to admit to the United States.

The attributes of immigrant profiles used in this factorial experiment, with number of levels provided in parentheses, are gender (2), education (7), employment plans (4), job experience (4), profession (11), language skills (4), country of origin (10), reasons for applying (3), and

prior trips to the United States (5). For completeness these factors and their levels are reproduced as Table A1 of the Supplementary Material (Goplerud, Imai and Pashley (2025b)). In total, there exist over 1.4 million possible profiles, implying more than 2×10^{12} possible comparisons of two profiles that are possible in the experiment. It is clear that with 1407 respondents, even though each respondent performs five comparisons, not all possible profiles can be included. Thus, exploring treatment effect heterogeneity requires a methodological development that goes beyond the models used previously in the causal inference literature for binary treatments.

The levels of each factor variable were independently randomized to yield one immigrant profile. Randomization was subject to some restrictions such that profession and education factors result in sensible pairings (e.g., ruling out doctors with less than two-years of college education) and immigrants whose reason for applying is persecution must come from Iraq, Sudan, Somalia, or China. The ordering of attributes was also randomized for each respondent. The experiment additionally collected data on the respondents, including demographic information, partisanship, attitudes toward immigration, and ethnocentrism. A rating for each immigrant profile was also recorded, but that metric is not the focus of our analysis.

2.2. *Heterogeneous treatment effects*. Hainmueller and Hopkins (2015) conducted their primary analysis based on a linear regression model where the unit of analysis is an immigrant profile (rather than a pair) and the outcome variable is an indicator for whether a given profile was chosen. The predictors of the model include the indicator variable for each immigrant attribute. The model also includes the interactions between education and profession as well as between country of origin and reasons for applying to account for the restricted randomization scheme mentioned above. Finally, the standard errors are clustered by respondent.

As formalized in Hainmueller, Hopkins and Yamamoto (2014), the regression coefficient represents the average marginal component effect (AMCE) of each attribute averaging over all the other attributes including those of the other profile in a given pair. Figure 1 reproduces the estimated overall AMCEs of country of origin where the baseline category is Germany. There is little country effect with the exception of Iraq, which negatively affects the likelihood of being preferred by a respondent.

Beyond the AMCEs these authors and others, including Newman and Malhotra (2019), have explored the heterogeneous treatment effects among respondents by conducting many subgroup analyses based on a number of respondent characteristics including partisanship

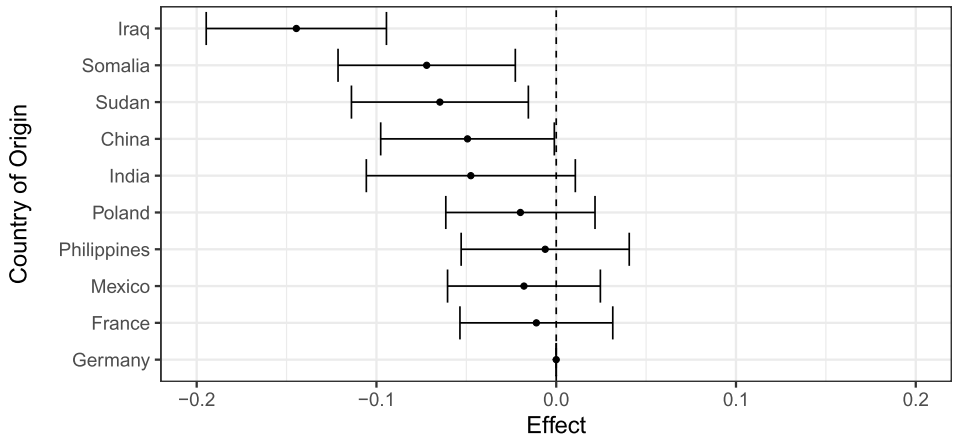

FIG. 1. *Estimated average marginal component effects of country of origin where the baseline is Germany, with effect estimates as given in Hainmueller and Hopkins (2015).*

TABLE 1
List of subset analyses performed in Hainmueller and Hopkins (2015), listed by moderator and how it was split to form subgroups

Moderator	Split
Education	Any college education or no college education
Ethnocentrism	Median ethnocentrism measure
Political party	Republican or Democrat
Percent of foreign born workers in respondent's industry	High or low
Household income	More or less than $50,000
Fiscal exposure to immigration	High or low
ZIP code demographics	$< 5\%$ immigrants, $> 5\%$ immigrants (primarily from Latin America), or $> 5\%$ immigrants (primarily not from Latin America)
Race/ethnicity	White or non-white
Hispanic ethnicity	Hispanic or non-hispanic
Ideology	Liberal or conservative
Immigration attitudes	Supports or does not support reducing immigration
Gender	Male or female
Age	Young or old

and level of education. Table 1 shows all of the subgroup analyses performed by Hainmueller and Hopkins (2015) and how the respondents were broken up into groups. We find that 13 subgroup analyses were performed (excluding those used for robustness checks), with results from the first three (education, ethnocentrism, and political party) presented in the main paper. Of those three analyses, the authors find some evidence of heterogeneous effects of country of origin between subsets that differ on ethnocentrism but little evidence of heterogeneity beyond this. The other 10 analyses can be found in their appendix, and the authors conclude that participants responded similarly, in general, across those subgroups.

Our goal is to build a methodology that enables one to more systematically explore heterogeneous treatment effects in conjoint experiments. Subgroup analyses, like those conducted in the original analysis, can be problematic for several reasons. First, the analyst must conduct a separate analysis for each moderator of interest, leading to multiple testing problems. Second, typically the moderators are dichotomized (or broken up into a small number of groups), requiring the analyst to decide how to split the data. Third, they are not amenable to exploration of how multiple moderators might work together to change outcomes.

To address these issues, one could include the moderators as covariates within the regression. However, if the goal is to provide estimated heterogeneous effects with straightforward interpretations, regressions with possibly complex interactions are not ideal. To estimate heterogeneous effects, we need to not only interact a large number of treatments but we will have to further interact all main and interaction effects of treatments with the moderators. It is unclear how to best reduce the dimensionality of both the moderator and treatment space in a classic regression setup. It is also challenging to interpret the interactions from these models to understand the characteristics of units that lead to different treatment effect patterns.

In sum, researchers must parsimoniously characterize how a large number of possible treatment combinations interact with several key moderators of interest. The goal is to obtain estimates of heterogeneous effects and understand how the covariate distributions of units with different treatment effects differ. We now turn to our methodology which is designed to address these challenges and result in interpretable estimates.

3. Modeling heterogeneous effects of high-dimensional treatments. We now describe the proposed methodology. To simplify the exposition, we focus on a general factorial design.

This design corresponds to conjoint analysis with a single task per person, where there is only one profile assessed rather than a comparison of profiles, and complete randomization of all combinations of factor levels. Extensions to independent factor randomization and realistic conjoint analyses are immediate and will be discussed and applied in Section 5.

3.1. *Setup.* Suppose that we have a simple random sample of N units. Consider a factorial design with J factors where each factor $j \in \{1, \ldots, J\}$ has $L_j \geq 2$ levels. The treatment variable for unit i, denoted by $\boldsymbol{T}_i$, is a J-dimensional vector of random variables, each of which represents the assigned level of the corresponding factor variable. For example, the jth element of this random vector $T_{ij} \in \{0, 1, 2, \ldots, L_j - 1\}$ represents the level of factor j which is assigned to unit i.

Following Dasgupta, Pillai and Rubin (2015), we define the potential outcome for unit i as $Y_i(\boldsymbol{t})$, where $\boldsymbol{t} \in \mathcal{T}$ represents the realized treatment with $\mathcal{T}$ representing the support of the randomization distribution for $\boldsymbol{T}_i$. Then the observed outcome is given by $Y_i = Y_i(\boldsymbol{T}_i)$. The notation implicitly assumes no interference between units (Rubin (1980)). In this paper, for the sake of concreteness, we focus on the binary outcome $Y_i \in \{0, 1\}$. Extensions to nonbinary outcomes are straightforward. Lastly, we observe a vector of p_x pretreatment covariates for each unit i and denote it by $\boldsymbol{X}_i$. All together, we observe $(Y_i, \boldsymbol{T}_i, \boldsymbol{X}_i)$ for each unit i.

To illustrate the notation, consider a simplified version of our motivating example where each respondent i observes a single immigrant profile and must decide whether to support that immigrant's admission or not. Then $\boldsymbol{T}_i$ is a vector indicating the level respondent i sees for each of the nine immigrant attributes. The outcome variable Y_i is an indicator for whether respondent i chooses to support admission for the immigrant with whom they are presented. Lastly, $\boldsymbol{X}_i$ denotes a vector of covariates for respondent i that we hypothesize might moderate the treatment effect. In our application, $\boldsymbol{X}_i$ included political party, education, demographics of their ZIP code, ethnicity, and Hispanic prejudice score (see Section 5.1 for details).

The randomness in our data, $(Y_i, \boldsymbol{T}_i, \boldsymbol{X}_i)$ comes from two sources: random sampling of units into the study and random assignment of units to treatments. For simplicity, we assume units are sampled via simple random sampling (though our method can incorporate sampling weights). The randomization of treatment assignment implies $\{Y_i(\boldsymbol{t})\}_{\boldsymbol{t} \in \mathcal{T}} \perp\!\!\!\perp \boldsymbol{T}_i$ for each i where the exact mode of randomization will determine the distribution of $\boldsymbol{T}_i$. In many conjoint experiments, researchers independently and uniformly randomize each factor. However, in some cases, including our application, researchers may exclude certain unrealistic combinations of factor levels (e.g., doctor without a college degree), leading to the dependence between factors. In all cases, researchers have complete knowledge of the randomization distribution of the factorial treatment variables.

Based on random sampling and random treatment assignment alone, we can conduct valid inference for marginal treatment effects of interest using simple regression or difference-in-means estimator (see Hainmueller, Hopkins and Yamamoto (2014)). If we wish to explore treatment effect heterogeneity across treatments and covariates, however, a model-based approach is useful. We next introduce our model, which will allow us to explore heterogeneous effects in a principled manner while also handling the high-dimensional nature of the data.

3.2. *General framework.* The most basic causal quantity of interest is the AMCE, which is defined for any given factor j as

$$\delta_j(l, l') = \mathbb{E}\big[Y_i(T_{ij} = l, \boldsymbol{T}_{i,-j}) - Y_i\big(T_{ij} = l', \boldsymbol{T}_{i,-j}\big)\big], \tag{3.1}$$

where $l \neq l' \in \mathcal{T}_j$ with $\mathcal{T}_j$ representing the support of the randomization distribution for T_j. The expectation in equation (3.1) is taken over the distribution of the other factors $\boldsymbol{T}_{i,-j}$ as well as the random sampling of units from the population. Thus, the AMCE averages

over two sources of causal heterogeneity—heterogeneity across treatment combinations and across units. Different treatment combinations may have distinct impacts on units with varying characteristics. Our goal is to model these potentially complex heterogeneous treatment effects using an interpretable model.

We propose to model heterogeneous treatment effects based on K distinct treatment effect patterns where $K \geq 2$ is chosen by a researcher, based on their desired granularity of heterogeneity. This approach, which is based on a fixed number of subgroups to characterize treatment effect heterogeneity, is commonly used by empirical researchers. Others have studied various methodological aspects of this approach albeit in the context of binary treatment (Chernozhukov et al. (2023), Imai and Li (2025)).

Our goal is to summarize the treatment effect heterogeneity by dividing the population into K subpopulations and characterizing these groups based on a set of pretreatment covariates, or "moderators," denoted by $\boldsymbol{X}_i$. In particular, we would like to construct K groups such that across-group treatment effect heterogeneity is maximized while minimizing the within-group heterogeneity. Since the treatments of interest are high dimensional, we focus on finding maximally heterogeneous groups in terms of average potential outcomes rather than their contrasts. We can then estimate any treatment effects of interest within each group.

Let $Z_i \in \{1, \ldots, K\}$ denote the latent group membership of unit i and $\mathcal{Z} = \{Z_i\}_{i=1}^n$. We use $\zeta_k(\boldsymbol{t}) = \mathbb{E}[Y_i(\boldsymbol{t}) \mid Z_i = k]$ to represent the average potential outcome under treatment $\boldsymbol{t}$ for group k. Under the randomization of T_i, define the estimated within-group average outcome under treatment t for group k and the estimated overall average outcome as $\hat{\zeta}_k(\boldsymbol{t}; \mathcal{Z}) = \sum_{i=1}^N I\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\} Y_i / \sum_{i=1}^N I\{Z_i = k, \boldsymbol{T}_i = \boldsymbol{t}\}$ and $\widehat{\bar{Y}}(\boldsymbol{t}) = \sum_{i=1}^N I\{\boldsymbol{T}_i = \boldsymbol{t}\} Y_i / \sum_{i=1}^N I\{\boldsymbol{T}_i = \boldsymbol{t}\}$, respectively.

Given the number of groups K selected by researchers, we show how to find maximally heterogeneous groups in terms of potential outcomes. The following proposition establishes that maximizing the Kullback–Leibler (KL) divergence of potential outcomes between groups is equivalent to maximizing the log-likelihood over groups and their centroids. We emphasize that this equivalence result does not assume the existence of a "correct" number of groups.

PROPOSITION 1 (Finding maximally heterogeneous groups). *Maximally heterogeneous groups in the terms of the KL divergence of potential outcomes can be found by maximizing the log-likelihood function over the group membership and the centroids of groups*,

$$\begin{aligned} &\operatorname*{argmax}_{\mathcal{Z}} \left\{ \sum_{k=1}^K \sum_{i=1}^N \mathbf{1}\{Z_i = k\} \mathrm{KL}\big(\hat{\zeta}_k(\boldsymbol{T}_i; \mathcal{Z}) \| \widehat{\bar{Y}}(\boldsymbol{T}_i)\big) \right\} \\ &\quad = \operatorname*{argmax}_{\mathcal{Z}} \sum_{k=1}^K \sup_{\zeta_k} \sum_{i=1}^N \mathbf{1}\{Z_i = k\} \big[Y_i \log \zeta_k(\boldsymbol{T}_i) + (1 - Y_i) \log\big\{1 - \zeta_k(\boldsymbol{T}_i)\big\} \big], \end{aligned} \tag{3.2}$$

where Y_i is binary, the KL divergence of two Bernoulli distributions with means μ_1 and μ_2 is given by $\mathrm{KL}(\mu_1 \| \mu_2) = \mu_1 \log \mu_1/\mu_2 + (1 - \mu_1) \log(1 - \mu_1)/(1 - \mu_2)$, *and $\{\hat{\zeta}_k(\boldsymbol{t}; \mathcal{Z})\}_{k=1}^K$ denotes the maximizers of the right-hand side of equation* (3.2) *given $\mathcal{Z}$.*

Section C of the Supplementary Material provides a proof of a more general result for the natural exponential family distributions (see Chi, Chi and Baraniuk (2016), for a similar result in the Gaussian case). The log-likelihood formulation is equivalent to the classification maximum likelihood approach in mixture modeling (McLachlan (1982)).

We now extend the above equivalence result to the settings in which we further model the group membership Z_i using a set of moderators $\boldsymbol{X}_i$, that is, $\pi_k(\boldsymbol{x}) = \Pr(Z_i = k \mid \boldsymbol{X}_i = \boldsymbol{x})$

for $k = 1, 2, \ldots, K$. Such a model helps characterize and understand the types of units that comprise each group. The next proposition shows that maximizing the log-likelihood function of this extended model is equivalent to finding K maximally heterogeneous groups such that the group memberships are predicted well by the moderators.

PROPOSITION 2 (Finding maximally heterogeneous groups with moderators). *Suppose that we extend the setting of Proposition 1 and additionally model the conditional probability of each individual's group membership, given categorical moderators* $\{\pi_k(\boldsymbol{X}_i)\}_{k=1}^K$. *Then maximally heterogeneous groups in terms of the KL divergence of potential outcomes with the entropy of group membership probabilities as a penalty term can be found by maximizing the log-likelihood function of the extended model,*

$$\begin{aligned}
&\operatorname*{argmax}_{\mathcal{Z}} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{1}\{Z_i = k\} \mathrm{KL}\big(\hat{\zeta}_k(\boldsymbol{T}_i; \mathcal{Z}) \| \widehat{\overline{Y}}(\boldsymbol{T}_i)\big) - \sum_{i=1}^{N} H\big(\{\hat{\pi}_k(\boldsymbol{X}_i; \mathcal{Z})\}_{k=1}^K\big) \right\} \\
&\quad = \operatorname*{argmax}_{\mathcal{Z}} \sum_{k=1}^{K} \sup_{\zeta_k, \pi_k} \sum_{i=1}^{N} \mathbf{1}\{Z_i = k\} \big[Y_i \log \zeta_k(\boldsymbol{T}_i) + (1 - Y_i) \log\{1 - \zeta_k(\boldsymbol{T}_i)\} \\
&\qquad + \log \pi_k(\boldsymbol{X}_i) \big],
\end{aligned} \tag{3.3}$$

where $H(\{p_k\}_{k=1}^K) = -\sum_{k=1}^K p_k \log p_k$ *(by convention, if* $p_k = 0$, *then* $p_k \log p_k = 0$*) is the entropy, and* $\hat{\pi}_k(\boldsymbol{x}; \mathcal{Z}) = \sum_{i=1}^N \mathbf{1}\{Z_i = k, \boldsymbol{X}_i = \boldsymbol{x}\} / \sum_{i=1}^N \mathbf{1}\{\boldsymbol{X}_i = \boldsymbol{x}\}$ *and* $\hat{\zeta}_k(\boldsymbol{t}; \mathcal{Z})$ *are the maximizers of the log-likelihood function of the right-hand side of equation* (3.3), *given* $\mathcal{Z}$.

Proof is given in Section D of the Supplementary Material. Since the entropy $H(\{\hat{\pi}_k(\boldsymbol{x}; \mathcal{Z})\}_{k=1}^K)$ is maximized when $\hat{\pi}_k(\boldsymbol{x}) = 1/K$, Proposition 2 shows that adding a group membership model based on moderators encourages finding groups whose memberships are well predicted by the moderators.

Direct optimization of equations (3.2) and (3.3) over $\mathcal{Z}$ has been studied under the name of "classification maximum likelihood" in the literature on mixture models (McLachlan (1982)). For completeness Section G.3 of the Supplementary Material provides an estimation algorithm for this approach, which modifies the proposed algorithm described in Section 3.5. Unfortunately, the classification maximum likelihood approach suffers from the incidental parameter problem because the cardinality of $\mathcal{Z}$ increases with the sample size N, leading to an asymptotic bias and inconsistency (Bryant and Williamson (1978)).

To address this problem, a dominant approach in the literature is Bayesian, treating the right-hand side of equation (3.3) as a log-posterior that consists of a log-likelihood and a log-prior over $\mathcal{Z}$, that is, $\Pr(Z_i = k \mid \boldsymbol{X}_i) = \pi_k(\boldsymbol{X}_i)$. By marginalizing out $\mathcal{Z}$, we avoid the incidental parameter problem, yielding the objective function known as a mixture maximum likelihood (McLachlan (1982)).

The model is called "mixture-of-experts" when π_k depends on $\boldsymbol{X}_i$ (Gormley and Frühwirth-Schnatter (2019)) with the following objective function:

$$\{\hat{\zeta}_k, \hat{\pi}_k\}_{k=1}^K = \operatorname*{argmax}_{\{\zeta_k, \pi_k\}_{k=1}^K} \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k(\boldsymbol{X}_i) \zeta_k(\boldsymbol{T}_i)^{Y_i} \{1 - \zeta_k(\boldsymbol{T}_i)\}^{1 - Y_i} \right]. \tag{3.4}$$

While this setup no longer appears to provide a direct characterization of the optimal groups, Proposition 3 shows that a mixture-of-experts model finds maximally heterogeneous groups as in Proposition 2 but with an additional penalty that encourages less extreme posterior probabilities of group memberships.

PROPOSITION 3 (Finding maximally heterogeneous groups with a mixture of experts). *Maximizing the likelihood function under a mixture-of-experts model is equivalent to finding maximally heterogeneous groups, as in Proposition 2, with an additional penalty. That is, the following equality holds for any* $\mathcal{Z}$:

$$\begin{aligned}
&\operatorname*{argmax}_{\zeta,\pi} \sum_{i=1}^{N} \log\left[\sum_{k=1}^{K} \pi_k(\boldsymbol{X}_i)\zeta_k(\boldsymbol{T}_i)^{Y_i}\{1-\zeta_k(\boldsymbol{T}_i)\}^{1-Y_i}\right] \\
&\quad= \operatorname*{argmax}_{\zeta,\pi} \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}\{Z_i=k\}\big[Y_i \log \zeta_k(\boldsymbol{T}_i) + (1-Y_i)\log\{1-\zeta_k(\boldsymbol{T}_i)\} \\
&\qquad + \log \pi_k(\boldsymbol{X}_i) - \log \tilde{\pi}_k\big(\boldsymbol{X}_i, Y_i, \boldsymbol{T}_i; \{\zeta_{k'}, \pi_{k'}\}_{k'=1}^{K}\big)\big],
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\pi}_k\big(\boldsymbol{X}_i, Y_i, \boldsymbol{T}_i; \{\zeta_{k'}, \pi_{k'}\}_{k'=1}^{K}\big) &= \Pr\big(Z_i = k \mid Y_i, \boldsymbol{T}_i, \boldsymbol{X}_i, \{\zeta_{k'}, \pi_{k'}\}_{k'=1}^{K}\big) \\
&= \frac{\pi_k(\boldsymbol{X}_i)\zeta_k(\boldsymbol{T}_i)^{Y_i}\{1-\zeta_k\}^{1-Y_i}}{\sum_{k'=1}^{K} \pi_{k'}(\boldsymbol{X}_i)\zeta_{k'}(\boldsymbol{T}_i)^{Y_i}\{1-\zeta_{k'}\}^{1-Y_i}}
\end{aligned}$$

is the posterior membership probability for group k.

Proof of the proposition directly follows from a well-known identity (e.g., Celeux, Frühwirth-Schnatter and Robert (2019)) and hence is omitted. The equality in Proposition 3 holds for any group membership $\mathcal{Z}$, including its maximum-a-posteriori (MAP) estimate, that is, $\hat{Z}_i = \operatorname{argmax}_k \tilde{\pi}_k(\boldsymbol{X}_i, Y_i, \boldsymbol{T}_i; \{\hat{\zeta}_{k'}, \hat{\pi}_{k'}\}_{k'=1}^{K})$. Thus, our proposed model can be seen as finding maximally heterogeneous groups while imposing a penalty that encourages finding groups that are well predicted by the moderators $\boldsymbol{X}_i$ but with less extreme group membership probabilities based on the data.

All together, our results provide a justification for using a mixture-of-experts model for heterogeneous effect estimation under the settings with high-dimensional treatments. We emphasize that a primary motivation for the use of Bayesian approach is to resolve the incidental parameter problem with classification maximum likelihood. Importantly, the results above do not assume a specific data generating process. Instead, we have shown that, given the number of groups and appropriate prior distributions, researchers can find maximally heterogeneous groups by fitting a mixture-of-experts model.

3.3. *Model specification.* Since $\mathcal{T}$ is high dimensional, many treatment combinations are unobserved with a typical sample size. Thus, nonparametric estimation is not applicable. We, therefore, model $\zeta_k(\boldsymbol{t})$ using a regularized logistic regression where an ANOVA-style sum-to-zero constraint is imposed separately for each factor to facilitate merging of different levels within each factor. This modeling strategy identifies a relatively small number of treatment combinations while avoiding the specification of a baseline level for each factor (Egami and Imai (2019)). The interpretation of $\zeta_k(\boldsymbol{t})$ under this model is still the average of potential outcome under treatment $\boldsymbol{t}$ in group k. Note that we do not assume homogeneity of outcomes or effects within each group.

We use a multinomial logistic regression for $\pi_k(\boldsymbol{x})$,

$$\zeta_k(\boldsymbol{T}_i) = \frac{\exp(\psi_k(\boldsymbol{T}_i))}{1+\exp(\psi_k(\boldsymbol{T}_i))}, \quad \text{and} \quad \pi_k(\boldsymbol{X}_i) = \frac{\exp(\boldsymbol{X}_i^{\top}\boldsymbol{\phi}_k)}{\sum_{k'=1}^{K}\exp(\boldsymbol{X}_i^{\top}\boldsymbol{\phi}_{k'})}, \tag{3.5}$$

where $\boldsymbol{\phi}_1 = \mathbf{0}$ for identification. For $\psi_k(\boldsymbol{T}_i)$, we assume an additive model and include both main effects and two-way interaction effects with a common intercept μ shared across all

groups,

$$\psi_k(\boldsymbol{T}_i) = \mu + \sum_{j=1}^{J}\sum_{l=0}^{L_j-1} \mathbf{1}\{T_{ij}=l\}\beta_{kl}^{j} + \sum_{j=1}^{J-1}\sum_{j'>j}\sum_{l=0}^{L_j-1}\sum_{l'=0}^{L_{j'}-1} \mathbf{1}\{T_{ij}=l, T_{ij'}=l'\}\beta_{kll'}^{jj'}$$
$$= \mu + \tilde{\boldsymbol{T}}_i^{\top}\boldsymbol{\beta}_k,$$

for each $k = 1, 2, \ldots, K$ where $\tilde{\boldsymbol{T}}_i$ is the vector of indicators, $\mathbf{1}\{T_{ij}=l\}$ and $\mathbf{1}\{T_{ij}=l, T_{ij'}=l'\}$, and $\boldsymbol{\beta}_k$ is a stacked column vector containing all coefficients for group k. Inclusion of higher-order interactions is straightforward (see Section E of the Supplementary Material) and hence is omitted in the main paper for notational simplicity.

For identification, we use the following ANOVA-type sum-to-zero constraints:

$$\sum_{l=0}^{L_j-1}\beta_{kl}^{j} = 0, \quad \text{and} \quad \sum_{l=0}^{L_j-1}\beta_{kll'}^{jj'} = \sum_{l'=0}^{L_{j'}-1}\beta_{kll'}^{jj'} = 0, \tag{3.6}$$

for $j, j' = 1, 2, \ldots, J$ with $j' > j$. We write them compactly as

$$\boldsymbol{C}^{\top}\boldsymbol{\beta}_k = \mathbf{0}, \tag{3.7}$$

where each row of $\boldsymbol{C}^{\top}\boldsymbol{\beta}_k$ corresponds to one of the constraints given in equation (3.6).

3.4. *Sparsity-inducing prior.* Given the high dimensionality of this model, we use a sparsity-inducing prior. In our application we have a total of 315 β coefficients for each group. In factorial experiments it is desirable to regularize the model such that certain levels of each factor are fused together when their main effects and all interactions are similar (Post and Bondell (2013), Egami and Imai (2019)). For example, we would like to fuse levels l_1 and l_2 of factor j if $\beta_{l_1}^{j} \approx \beta_{l_2}^{j}$ and $\beta_{l_1 l'}^{jj'} \approx \beta_{l_2 l'}^{jj'}$ for all other factors j' and all of its levels l'.

We encourage such fusion by applying a structured sparsity approach of Goplerud (2021) that generalizes the group and overlapping group LASSO (e.g., Yuan and Lin (2006), Yan and Bien (2017)) while allowing positive semidefinite penalty matrices. For computational tractability we use ℓ_2 norm instead of the ℓ_∞ norm, which is used in GASH-ANOVA (Post and Bondell (2013)). An additional benefit of the use of regularization is that it gives us some protection against finding spurious relations (see Gelman, Hill and Yajima (2012)).

For illustration, consider a simple example with one group and two factors—factor one has three levels, and factor two has two levels. In this case, our penalty contains four terms,

$$\sqrt{(\beta_0^1-\beta_1^1)^2 + (\beta_{00}^{12}-\beta_{10}^{12})^2 + (\beta_{01}^{12}-\beta_{11}^{12})^2}$$
$$+ \sqrt{(\beta_0^1-\beta_2^1)^2 + (\beta_{00}^{12}-\beta_{20}^{12})^2 + (\beta_{01}^{12}-\beta_{21}^{12})^2}$$
$$+ \sqrt{(\beta_1^1-\beta_2^1)^2 + (\beta_{10}^{12}-\beta_{20}^{12})^2 + (\beta_{11}^{12}-\beta_{21}^{12})^2}$$
$$+ \sqrt{(\beta_0^2-\beta_1^2)^2 + (\beta_{00}^{12}-\beta_{01}^{12})^2 + (\beta_{10}^{12}-\beta_{11}^{12})^2 + (\beta_{20}^{12}-\beta_{21}^{12})^2}.$$

The first three terms encourage the pairwise fusion of the levels of factor one whereas the fourth encourages the fusion of the two levels of factor two. For compact notation the penalty can also be written using the sum of Euclidean norms of quadratic forms,

$$\|\boldsymbol{\beta}^{\top}\boldsymbol{F}_1\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta}^{\top}\boldsymbol{F}_2\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta}^{\top}\boldsymbol{F}_3\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta}^{\top}\boldsymbol{F}_4\boldsymbol{\beta}\|_2,$$

where $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{F}_3$ are appropriate positive semidefinite matrices to encourage the fusion of the pairs of levels in factor one and $\boldsymbol{F}_4$ encourages the fusion of the two levels in factor two,

and $\boldsymbol{\beta} = [\beta_0^1\ \beta_1^1\ \beta_2^1\ \beta_0^2\ \beta_1^2\ \beta_{00}^{12}\ \beta_{10}^{12}\ \beta_{20}^{12}\ \beta_{01}^{12}\ \beta_{11}^{12}\ \beta_{21}^{12}]^\top$. Note that the sum-to-zero constraints make this type of fusion of factors together sensible for sparsity.

We generalize this formulation to an arbitrary number of factors and factor levels. For each factor that contains L_j levels, we have $\binom{L_j}{2}$ penalty matrices to encourage pairwise fusion. Imposing additional constraints is a simple extension; for example, for ordered factors, one might use penalties that penalize the differences between adjacent levels (e.g. l and $l+1$). Let $G = \sum_{j=1}^{J} \binom{L_j}{2}$ represent the total number of penalty matrices. For $g = 1, 2, \ldots, G$, we use $\boldsymbol{F}_g$ to denote a penalty matrix such that $\sqrt{\boldsymbol{\beta}^\top \boldsymbol{F}_g \boldsymbol{\beta}}$ is equivalent to the ℓ_2 norm on the vector of differences between all main effects and interactions containing a main effect. We note that $\{\boldsymbol{F}_g\}_{g=1}^G$ is not directly chosen but rather are determined by factors in the experiment (J, L_j, whether j is ordered or unordered) and the included interactions (as well as the use of "latent overlapping groups"); see Section H.4 of the Supplementary Material.

We interpret this penalty as a prior under our Bayesian framework described in Section 3.2,

$$p(\boldsymbol{\beta}_k \mid \{\boldsymbol{\phi}_k\}_{k=2}^K) \propto (\lambda \bar{\pi}_k^\gamma)^m \exp\left(-\lambda \bar{\pi}_k^\gamma \sum_{g=1}^{G} \sqrt{\boldsymbol{\beta}_k^\top \boldsymbol{F}_g \boldsymbol{\beta}_k}\right), \tag{3.8}$$

where $\bar{\pi}_k = \sum_{i=1}^N \pi_k(\boldsymbol{X}_i)/N$ and $m = \operatorname{rank}([\boldsymbol{F}_1, \ldots, \boldsymbol{F}_G])$. We follow existing work in allowing the penalty on the treatment effects $\boldsymbol{\beta}_k$ to be scaled by the group-membership size $\bar{\pi}_k$ when $\gamma = 1$ (Khalili and Chen (2007), Städler, Bühlmann and Van De Geer (2010)). On the other hand, when $\gamma = 0$, the $\bar{\pi}_k$ disappears, implying no use of the $\boldsymbol{X}_i$ in the prior. We note that the prior on $p(\boldsymbol{\beta} \mid \{\boldsymbol{\phi}_k\}_{k=2}^K)$ is guaranteed to be proper when all pairwise fusions are encouraged by $\{\boldsymbol{F}_g\}_{g=1}^G$, although in other circumstances it may be improper (Goplerud (2021)). Section F of the Supplementary Material provides additional details. Following Zahid and Tutz (2013), we use a normal prior distribution for the coefficients for the moderators.

The resulting regularization is invariant to the choice of baseline group $\boldsymbol{\phi}_1 = \mathbf{0}$, which is the first row of the $K \times p_x$ coefficient matrix $\boldsymbol{\phi}$. The prior distribution is given by

$$p(\{\boldsymbol{\phi}_k\}_{k=2}^K) \propto \exp\left(-\frac{\sigma_\phi^2}{2} \sum_{l=1}^{p_x} [\boldsymbol{\phi}_{2l}, \ldots, \boldsymbol{\phi}_{Kl}]^\top \boldsymbol{\Sigma}_\phi [\boldsymbol{\phi}_{2l}, \ldots, \boldsymbol{\phi}_{Kl}]\right), \tag{3.9}$$

where $\boldsymbol{\Sigma}_\phi$ is a $(K-1) \times (K-1)$ matrix with $[\boldsymbol{\Sigma}_\phi]_{kk'} = (K-1)/K$ if $k = k'$ and $[\boldsymbol{\Sigma}_\phi]_{kk'} = -1/K$ otherwise. We set σ_ϕ^2 to $1/4$ for a relatively diffuse prior.

As noted in a recent survey, "ensuring generic identifiability for general [mixture of expert] models remains a challenging issue" (Gormley and Frühwirth-Schnatter ((2019), p. 294)). Although mixtures with a Bernoulli outcome variable are generally unidentifiable, several aspects of our methodology are expected to alleviate the identifiability problem. First, a typical conjoint analysis has repeated observations per unit i (Grün and Leisch (2008)). Second, our model is a mixture of experts rather than a mixture model (Jiang and Tanner (1999)). Third, our treatment variables, which act as covariates in a mixture ofF experts, are randomized and hence uncorrelated with one another. Lastly, our model regularizes the coefficients through an informative prior. While a formal identifiability analysis of our model is beyond the scope of this paper, the simulation analysis (Section 4) shows that our model can accurately recover the coefficients in a realistic setting. It is also possible to use a bootstrap-based procedure to examine the identifiability issue in a specific setting (Grün and Leisch (2008)).

3.5. *Estimation and inference*. We fit our model by finding a maximum of the log-posterior using an extension of the expectation–maximization (EM; Dempster, Laird and Rubin (1977)) algorithm known as the Alternating Expectation–Conditional Maximization

(AECM; Meng and van Dyk (1997)) algorithm. Equation (3.10) defines our (observed) log-posterior using the terms defined in equations (3.3), (3.8), and (3.9), where we collect all model parameters as $\boldsymbol{\theta}$,

$$\log p(\boldsymbol{\theta} \mid \{Y_i, \boldsymbol{X}_i, \boldsymbol{T}_i\}_{i=1}^N) = \sum_{i=1}^N \log\left[\sum_{k=1}^K \pi_k(\boldsymbol{X}_i)\zeta_k(\boldsymbol{T}_i)^{Y_i}\{1-\zeta_k(\boldsymbol{T}_i)\}^{1-Y_i}\right] + \sum_{k=1}^K \log p(\boldsymbol{\beta}_k \mid \{\boldsymbol{\phi}\}_{k=2}^K) + \log p(\{\boldsymbol{\phi}_k\}_{k=2}^K) + \text{const.} \tag{3.10}$$

For now, we assume the value of regularization parameter λ is fixed, although we discuss this issue in Section 3.6. The linear constraints on $\boldsymbol{\beta}_k$ given in equation (3.7) still hold but are suppressed for notational simplicity.

Section G of the Supplementary Material provides a full derivation of our AECM algorithm; each iteration involves two cycles where the data augmentation scheme enables iterative updating of the treatment effect parameters $\boldsymbol{\beta}$ and moderators $\boldsymbol{\phi}$. After augmenting with missing data, the update for $\boldsymbol{\beta}$ can be done using ridge regression; Section G.1 addresses the linear constraints imposed by $\boldsymbol{C}^T\boldsymbol{\beta}_k = \boldsymbol{0}$. The update for $\boldsymbol{\phi}$ can be performed using a modified version of a multinomial logistic regression based on a standard optimizer (e.g., L-BFGS) (see Section G.2).

3.6. *Additional considerations.* Since fitting the proposed model is computationally expensive, we use the Bayesian Information Criteria (BIC), rather than cross-validation, to select the value of the regularization parameter λ (Khalili and Chen (2007), Khalili (2010), Chamroukhi and Huynh (2019)). Section G.4 of the Supplementary Material presents our degrees-of-freedom estimator and explains how we tune λ using Bayesian model-based optimization. Section G.5 discusses additional details of our EM algorithm including initialization and techniques to accelerate convergence.

We extend the above model and estimation algorithm to accommodate common features of conjoint analysis: (1) repeated observations for each individual respondent (Section H.1 of the Supplementary Material), (2) a forced choice conjoint design (Section H.2), and (3) standardization weights for factors with different numbers of levels L_j (Section H.3). Lastly, our experience suggests that the proposed penalty function, which consists of overlapping groups, often finds highly sparse solutions. Section H.4 details the integration of the latent overlapping group formulation of Yan and Bien (2017) into our framework to address this issue.

Once the model parameters are estimated, we can compute quantities of interest, such as the AMCEs, defined in equation (3.1). We do this separately for each group such that $\delta_{jk}(l, l')$ is the AMCE for factor j, changing from level l' to l in group k. Our estimator is the average of the estimated difference in predicted responses when changing from level l' to l of factor j, where the average is taken over the empirical distribution of the assignment on the other factors. This estimation is described in more detail in Section I of the Supplementary Material under various settings. We can use the empirical distribution here because treatment is randomly assigned.

To quantify the uncertainty of the parameter estimates, we rely on a quadratic approximation to the log-posterior distribution. To ensure its differentiability, we follow a standard approach in the regularized regression literature (e.g., Fan and Li (2001)) and fuse pairwise factor levels that are sufficiently close together. Section J of the Supplementary Material describes this process, deriving the Hessian of the log-posterior using Louis (1982)'s method and then using the delta method for inference on other of quantities of interest, for example, the AMCE.

Finally, in principle, our framework does not assume a "correct" data generating process. The choice of number of groups K should depend on the desired granularity of discovered heterogeneity, with more groups leading to finer levels of heterogeneity. Similarly, the choice of moderators should reflect the researcher's substantive interests. Section K.3 of the Supplementary Material shows performance of our method across different values of K and different specifications of the moderators when the true data generating process is a mixture model. As expected, the bias of AMCE is not affected by changing the specification of these parameters. However, there are some impacts on the estimation of conditional effects in terms of precision.

Common data-driven approaches for choosing K include use of an information criterion such as the BIC; however, while we find that these approaches work well under simulation settings (see Section K.3.1 for demonstration), they can perform poorly in practice (see Section L), especially when the component densities are misspecified or not especially well separated (Celeux, Frühwirth-Schnatter and Robert (2019)). Thus, even if a data-driven heuristic is used as a guide for choosing K, we suggest comparing different K as illustrated in Section 5.

4. Simulations. We explore the performance of our method using a simple but realistic simulation study. Specifically, we consider the case of a conjoint experiment with 10 factors ($J = 10$), each with three levels ($L_j = 3$). To evaluate the performance of the proposed method, we consider two different settings; in the first, we assume there are 1000 respondents who each perform five comparison tasks. In the second, we assume a larger experiment with 2,000 respondents who each perform ten tasks.

In all cases we assume that the data generating process follows a mixture of experts' models with three groups ($K = 3$). We calibrate the true $\boldsymbol{\beta}_k$ such that the implied average marginal component effects (AMCE) are comparable in magnitude to the empirical effects presented in Section 5. We use a set of five correlated continuous moderators and an intercept to again mimic a realistic empirical setting and choose $\{\boldsymbol{\phi}_k\}_{k=2}^3$ to relatively clearly separate respondents into different groups. Section K of the Supplementary Material presents complete description of the simulation settings and the true parameter values used for the $\boldsymbol{\beta}_k$ and marginal effects.

For each sample size, we independently generate 1000 simulated data sets by drawing N observations of moderators, randomly assigning a group membership to each observation based on the implied probabilities, given their moderators, and generating the observed treatment profiles completely at random. We fit our model to the data with $K = 3$ and examine the average marginal component effects in each group with respect to the first baseline level.

Figure 2 summarizes our results (see Section K.2 for the results regarding the estimated coefficients $\boldsymbol{\beta}_k$). The left panel illustrates a high correlation between the estimated effects and their true values ($\rho = 0.995$ for smaller sample size; $\rho = 0.999$ for larger sample size). While the performance overall is reasonably strong, we see that, even when the dataset is large, there is some degree of attenuation bias due to shrinkage.

The right panel shows the frequentist evaluation of our Bayesian posterior standard deviations. We compare the average posterior standard deviation against the standard deviation of the estimated effects across the 1000 Monte Carlo simulations. The average posterior standard deviations are noticeably smaller than the standard deviation of the estimates when the sample size is small. For the large sample size, however, our approximate Bayesian posterior standard deviations in this simulated example are roughly the same magnitude of the standard deviation of the sampling distribution of the estimator.

Even though our method's frequentist coverage is somewhat below the nominal level in small samples, this undercoverage appears to be primarily attributable to the shrinkage bias

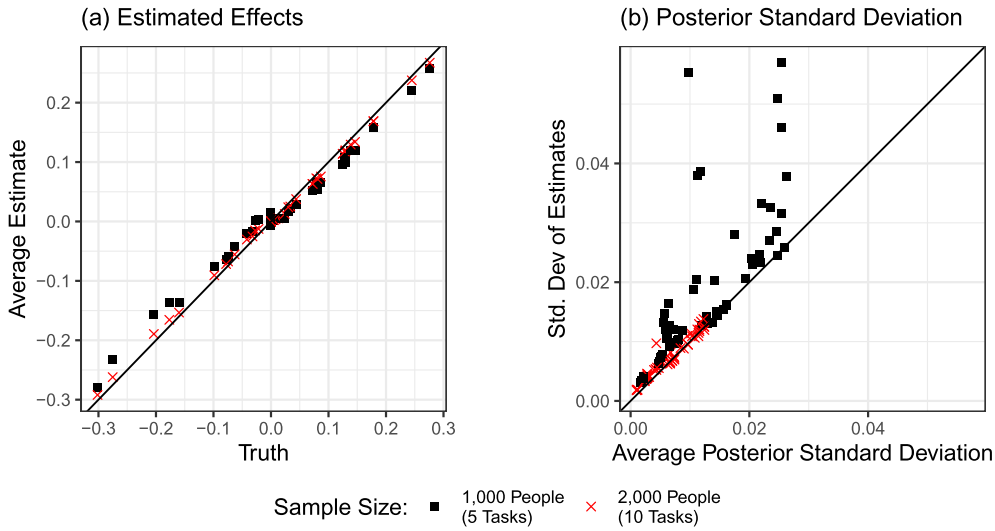

FIG. 2. *The empirical performance of the proposed estimator on simulated data. The black squares indicate the effects estimated for each group with the smaller sample size* (1000 *people completing five tasks*); *the red crosses indicate effects estimated with the larger sample size* (2000 *people completing* 10 *tasks*).

in our regularized estimation rather than the large sample discrepancy between our posterior standard deviations and the corresponding standard deviation of sampling distribution.

Section K.2 explores one way to address the limitations of the default estimator by exploring sample splitting and refitting the model, given the estimated sparsity pattern (i.e., which levels are fused together) and moderator effects ($\{\boldsymbol{\phi}_k\}_{k=2}^K$) on half of the data. This results in smaller bias and improved coverage at both sample sizes.

Section K.3 explores how, when the true data generating process is a mixture model, the "wrong" choice of K, for example, $K \in \{1, 2, 4\}$ as well as not using moderators (i.e., $\boldsymbol{X}_i = 1$) or using moderators in a different specification than the true model impacts our results. In both settings there is limited impact in terms of bias in terms of estimating the AMCE, although both types of misspecification incur a penalty in terms of root mean-squared error.

5. Empirical analysis. In this section we apply our methodology to the immigration conjoint data introduced in Section 2. We find evidence of effect heterogeneity for immigrant choice based on respondent characteristics. In particular, the immigrant's country of origin plays a greater role in forming the immigration preference of respondents with increased prejudice, as measured by a Hispanic prejudice score. Outside of this group, which accounts for about one third of the respondents, the country of origin factor plays a much smaller role.

5.1. *Data and model.* Following the original analysis, our model includes indicator variables for each factor and interactions between country and reason of application factors as well as those between education and job factors in order to account for the restricted randomization. We additionally include interactions between country and job as well as those between country and education, in accordance with the skill premium theory of Newman and Malhotra (2019). This theory hypothesizes that prejudiced individuals prefer highly skilled immigrants only for certain immigrant countries. This results in a total of 41 AMCEs and 222 average marginal interaction effects (AMIEs) for each group.

For modeling group membership, we include the respondents' political party, education, demographics of their ZIP code (we follow the original analysis and include the variables indicating whether respondents' ZIP codes had few immigrants, meaning $< 5\%$, and for

those from ZIPs with more than 5% foreign born, whether the majority were from Latin America), ethnicity, and Hispanic prejudice score. The Hispanic prejudice score was used by Newman and Malhotra (2019), though we negate it to make lower values correspond to lower prejudice for easier interpretation. The score is based on a standardized (and negated) feeling thermometer for Hispanics. The score ranges from -1.61 to 2.11 for our sample, where higher scores indicate higher levels of prejudice.

We remove respondents who are themselves Hispanic since the Hispanic prejudice score was not measured for these respondents. After removing entries with missing data, we have a sample of 1069 respondents. Most respondents evaluated five pairs of profiles, though five respondents have fewer responses in the data set used. The total number of observations is 5337 pairs of profiles. We do not incorporate the survey weights into our analysis to better demonstrate our methods, though it is possible to include them.

The original experiment was conducted using the forced choice design in which a respondent chooses one profile out of a pair of immigrant profiles. We follow Egami and Imai (2019) and model the choice as a function of differences in treatments as follows:

$$\begin{aligned}
&\psi_k(\boldsymbol{T}_i^L, \boldsymbol{T}_i^R) \\
&\quad = \mu + \sum_{j=1}^{J} \sum_{l \in L_j} \beta_{kl}^{j} (\mathbf{1}\{T_{ij}^L = l\} - \mathbf{1}\{T_{ij}^R = l\}) \\
&\qquad + \sum_{j=1}^{J-1} \sum_{j'>j} \mathbf{1}\{\mathcal{I}(j, j')\} \sum_{l \in L_j} \sum_{l' \in L_{j'}} \beta_{kll'}^{jj'} (\mathbf{1}\{T_{ij}^L = l, T_{ij'}^L = l'\} - \mathbf{1}\{T_{ij}^R = l, T_{ij'}^R = l'\}),
\end{aligned}$$

where $\boldsymbol{T}_i^L$ and $\boldsymbol{T}_i^R$ represent the factors for the left and right profiles and $\mathcal{I}(j, j') = 1$ if an interaction between j and j' is include in the model. The outcome variable Y_i is equal to 1 if the left profile is selected and is equal to 0 if the right profile is chosen.

To account for randomization restrictions, we include interactions between country of origin and reason for applying as well as between job and education. To test relevant theories, we include additional interactions between country of origin and job as well as country of origin and education. These interaction effects proved to be very small in magnitude (see Section L of the Supplementary Material). Thus, we do not explore higher order interactions, given the commonly adopted principles of hierarchy and sparsity (Wu and Hamada (2021)), which implies that lower-order effects are expected to be more significant than higher-order effects and we should expect an even smaller number of nonzero higher-order effects. With this linear predictor formulation, the estimation and inference proceed as explained in Section 3.

We conduct two analyses, one with two groups and the other with three groups. These two models perform equally well in terms of out-of-sample classification, a data-driven measure that can be used to choose the number of groups. Using more than three groups does not give improved substantive insights and provides little improvement in model performance. As noted previously, each analysis optimizes the BIC to calibrate the amount of regularization and employs standardization weights to account for factors with different number of levels (see Sections G.4 and H.3 of the Supplementary Material, respectively, for details). We treat education and job experience as ordered factors and only penalize the differences between adjacent levels.

We report our findings using only the full data estimates, that is, without the sample splitting explored in Section K.2. Initial experiments found that the results were somewhat sensitive to specific folds chosen, and thus we report only the full data results in the main text. Section L illustrates the distribution of estimates across 20 different sample splits.

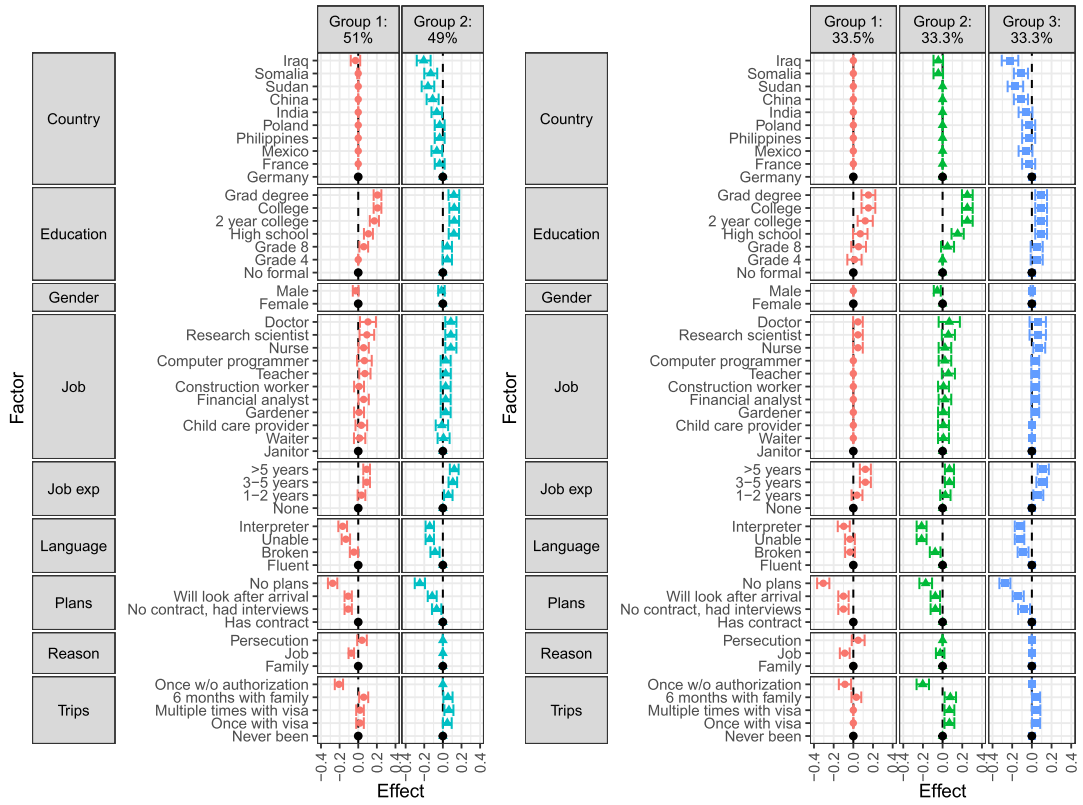

FIG. 3. *Estimated average marginal component effects using a two-group (left) and three-group (right) analysis. The point estimates and 95% Bayesian credible intervals are shown. A solid circle represents the baseline level of each factor. Numbers after colons give average posterior predictive probabilities for each group.*

5.2. *Estimated heterogeneity.* We focus on the AMCE for each factor as the quantity of interest and separately estimate it for each group. Under our model for the forced choice design, the AMCE of level l vs. level l' of factor j within group k can be written as

$$\begin{aligned}\delta_{jk}(l,l') = \frac{1}{2}\mathbb{E}[\{&\Pr(Y_i = 1 \mid Z_i = k, T_{ij}^L = l, \boldsymbol{T}_{i,-j}^L, \boldsymbol{T}_i^R) \\ &- \Pr(Y_i = 1 \mid Z_i = k, T_{ij}^L = l', \boldsymbol{T}_{i,-j}^L, \boldsymbol{T}_i^R)\} \\ &+ \{\Pr(Y_i = 0 \mid Z_i = k, T_{ij}^R = l, \boldsymbol{T}_{i,-j}^R, \boldsymbol{T}_i^L) \\ &- \Pr(Y_i = 0 \mid Z_i = k, T_{ij}^R = l', \boldsymbol{T}_{i,-j}^R, \boldsymbol{T}_i^L)\}].\end{aligned}$$

The expectation is over the population of respondents and the distribution of the factors not involved in this AMCE. That is, we compute the AMCE separately for the left and right profiles and then average them to obtain the overall AMCE. We estimate this quantity using the fitted model and averaging over the empirical distribution of the factorial treatments.

Figure 3 presents the estimated AMCEs and their 95% Bayesian credible intervals for the two-group and three-group analyses in the left and right panels, respectively. Group 2 in the two-group analysis and Group 3 in the three-group analysis display stronger impacts of country of origin than the other groups. The respondents in these groups give the most preference to immigrants from Germany and the least preference to immigrants from Iraq (followed by Sudan). The significant negative effects of Iraq in Group 2 of the two-group analysis and Group 3 of the three-group analysis are consistent with the significant negative

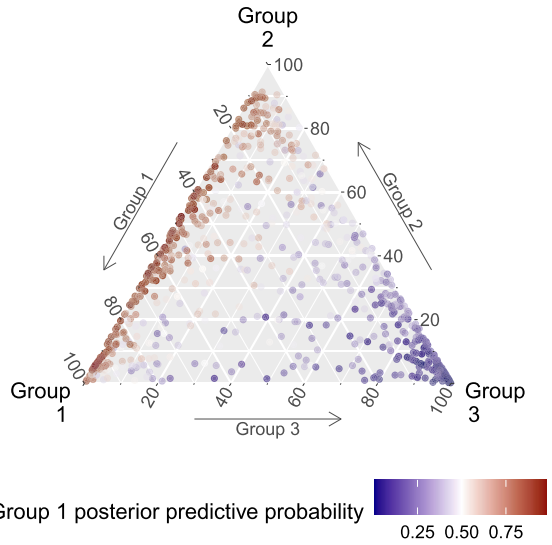

FIG. 4. *Ternary plot of the joint posterior predictive probability of belonging to each group in the three-group analysis (three axes) where the color of each dot represents the posterior predictive probability of belonging to Group* 1 *under the two-group analysis.*

effect for Iraq found by Hainmueller and Hopkins (2015). The patterns we observe for the other factors are also similar for these two groups in the two analyses.

Across all groups, respondents prefer educated and experienced immigrants who already have contracts (over those who have no contracts or plans). Respondents also prefer immigrants who have better language skills, although this feature matters less for respondents in Group 1 of the three group analyses.

For both analyses, the respondents in Group 1 do not care much about immigrant's country of origin. Instead, they place a greater emphasis on education and reason for immigration when compared to those in the other groups. While the differences between Groups 1 and 2 in the three-group analysis are generally substantively small, those in Group 2 appear to place more emphasis on education and prior entry without legal authorization. Those in Group 1, on the other hand, give a slight benefit to immigrants whose reason for immigration is persecution.

Indeed, for the three-group analysis, Groups 1 and 2 together correspond roughly to Group 1 of the two-group analysis. In fact, about 81% of the respondents who belong to Group 1 of the two-group analysis are the members of either Group 1 or 2 in the three-group analysis, using a weighted average of their estimated group membership posterior predictive probabilities.

Figure 4 visualizes these posterior predictive probabilities of group membership under the three-group analysis with each dot colored by the posterior predictive probability of belonging to Group 1 under the two-group analysis. According to this ternary plot, those observations that are likely to be part of Group 1 under the two-group analysis (i.e., red dots) are likely to be split between Groups 1 and 2 under the three group analysis. In contrast, those who have a high probability of belonging to Group 2 under the two-group analysis (i.e., blue dots) tend to be part of Group 3.

Figure 3 shows fusion of various factor levels due to regularization. The levels being fused appear sensible. For example, "doctor" and "research scientist," both occupations requiring high levels of education, are consistently fused together. For education, use of the ordinal structure ensures only adjacent levels can be fused. We see sensible cut points for fusion; in the two group analysis, Group 1 differentiates individuals who have at least a college degree, and Group 2 differentiates individuals who have at least a high school degree.

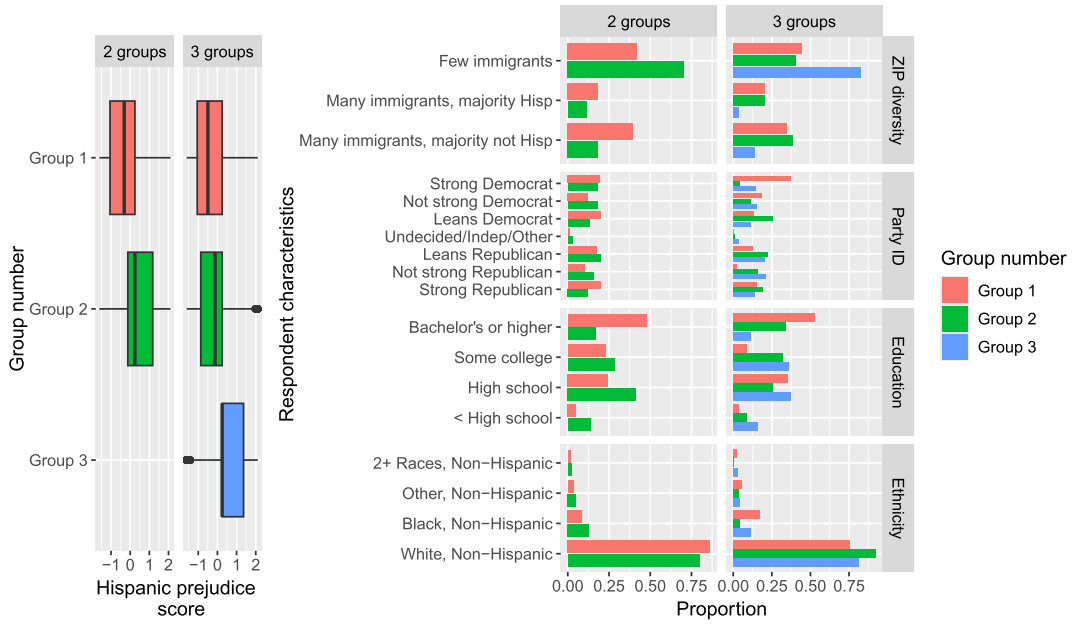

FIG. 5. *Distribution of respondent characteristics for each group. Left set of plots shows weighted box plots of the Hispanic prejudice moderator within each group over the posterior predictive distribution using a two-group (left) and three-group (right) analysis. Right set of plots shows the distribution of categorical moderators within each group over the posterior predictive distribution using a two-group (left) and three-group (right) analysis.*

The comparison of AMCEs across subgroups can be misleading, as they depend on the choice of baseline category (Leeper, Hobolt and Tilley (2020)). Section L of the Supplementary Material presents an alternative quantity that avoids issues of baseline dependency (marginal means; Leeper, Hobolt and Tilley (2020)). The results are generally similar to AMCEs shown above.

5.3. *Group membership.* Who belongs to each group? The left panel of Figure 5 shows the distribution of Hispanic prejudice score for each group weighted by the corresponding posterior predictive group membership probability for each individual respondent. The plot shows that for the two-group analysis, those with high prejudice score are more likely to be part of Group 2. For the three-group analyses, those with high prejudice are more likely to be in Group 3. This is consistent with the finding above that the respondents in those groups put more emphasis on immigrant's country of origin.

The right panel of the figure shows the distribution of other respondent characteristics. In general, Group 2 in the two-group analysis and Group 3 in the three-group analysis consist of those who live in ZIP codes with few immigrants and have lower educational achievements. For the three-group analysis, those in Group 2 tend to be Republicans, whereas those in Group 1 are more likely to be Democrats. This is consistent with the finding of a larger penalty for entry without legal authorization in Group 2. Group 3 contains a mix of political ideologies, though it has more respondents who identify as Undecided/Independent/Other or not strong Republican than the other two groups.

Which respondent characteristics are predictive of the group membership? In addition to the covariate distribution for each group shown in Figure 5, we can also find how important each moderator is in predicting group membership, conditional on all other moderators. We examine how the predicted probabilities of group memberships change across respondents

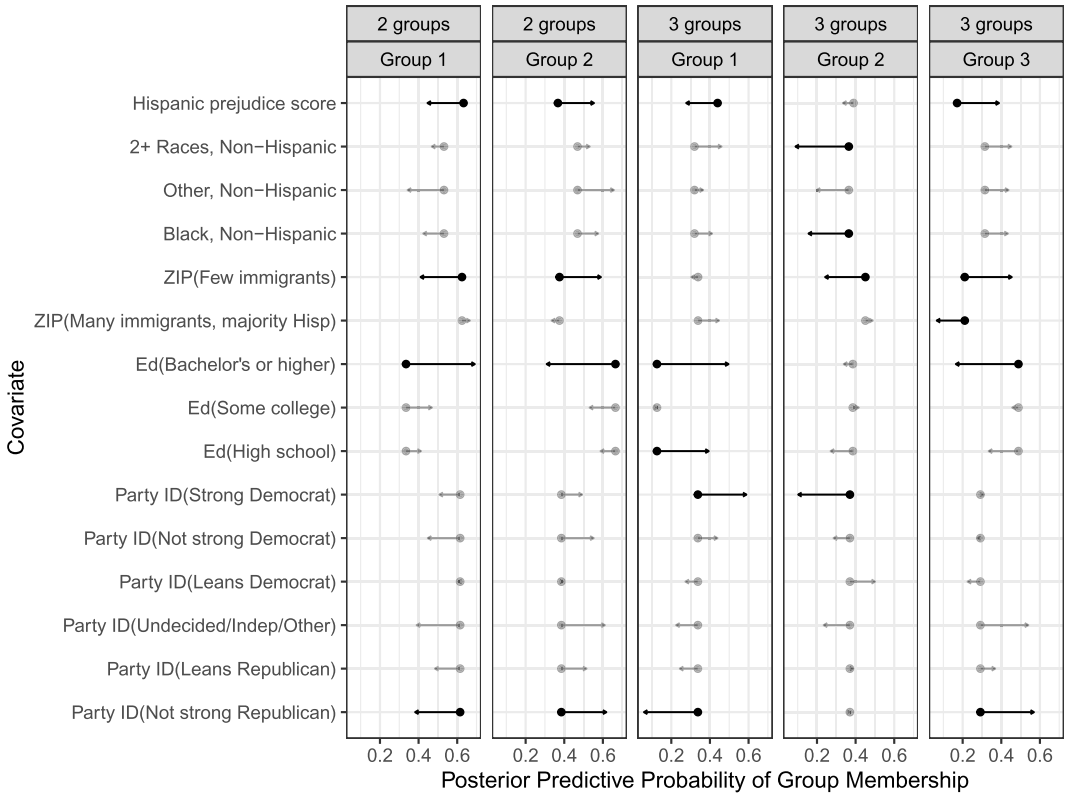

FIG. 6. *The impact of moderator values on likelihood of being assigned to groups, for two-group (left two plots) and three-group (right three plots) analysis. Dark arrows indicate that there is a significant effect of the moderator on group membership, that is, that the corresponding quantity defined in equation (5.1) is statistically significant.*

with different characteristics. Specifically, we estimate

$$\mathbb{E}\big[\pi_k(X_{ij} = x_1, \boldsymbol{X}_{i,-j}) - \pi_k(X_{ij} = x_0, \boldsymbol{X}_{i,-j})\big], \tag{5.1}$$

where x_0 and x_1 are different values of covariate of interest X_{ij}. If X_{ij} is a categorical variable, we set x_0 to the baseline level and x_1 to the level indicated on the vertical axis. If X_{ij} is a continuous variable, as in the case of the Hispanic prejudice score, then x_0 and x_1 represent the 25th and 75th percentile values. The solid arrows represent whether the corresponding 95% Bayesian credible interval covers zero or not. Section L of the Supplementary Material shows the effect of changing a moderator on the absolute value of the changes in predicted probabilities of group membership. In some cases, changing a moderator shows a small average change but a larger average of absolute changes.

Consistent with the earlier findings, Figure 6 shows that those with high Hispanic prejudice scores tend to be part of Group 2 in the two-group analysis and Group 3 in the three-group analysis, even after controlling for other moderators. These respondents are also less likely to be members of Group 1 in both analyses. Party ID also plays a statistically significant role (indicated by dark arrow). Controlling for other factors, in the three-group analysis, not strong Republicans tend to be part of Group 3 and more strong Democrats belonging to Group 1. On average, respondents in Group 1 have higher education in both analyses.

Finally, we estimate the average marginal interaction effects (AMIEs) between two factors (Egami and Imai (2019)), which can be computed by subtracting the two AMCEs from the average effect of changing the two factors of interest at the same time. Thus, the AMIE represents the additional effect of changing the two factors beyond the sum of the average

effects of changing one of the factors alone. Formally, we can define the AMIE of changing factors j and j' from levels l_j and $l_{j'}$ to levels l'_j and $l'_{j'}$, respectively, as follows:

$$\mathbb{E}\big[Y_i(T_{ij} = l_j, T_{ij'} = l_{j'}, \boldsymbol{T}_{i,-j,-j'}) - Y_i\big(T_{ij} = l'_j, T_{ij'} = l'_{j'}, \boldsymbol{T}_{i,-j,-j'}\big)\big] - \delta_j(l_j, l'_j) - \delta_{j'}(l_{j'}, l'_{j'}).$$

All of the AMIE effects found are quite small, so we do not present those results here. According to the skill-premium theory of Newman and Malhotra (2019), we expect to find an interaction between job and country or education and country, in at least some groups. Unfortunately, our analysis does not find support for this hypothesis.

5.4. *Comparison to an alternative method.* While there exist few methods to estimate heterogeneous effects of high-dimensional treatments, an exception is Robinson and Duch (2024), who develop a BART-based method for analyzing heterogeneity in conjoint experiments. The primary goal of their method is the estimation of the conditional average marginal effects (CAMCE) for each individual given their covariate values.

While our method is motivated by a different goal—finding an interpretable set of groups with distinctive treatment effects—our method can also produce estimates of the CAMCE for any set of covariates. The two methods can be compared in this task by examining CAMCE. Formally, under our model the CAMCE for factor j, comparing levels l and l' for covariates $\boldsymbol{X}_i$, is a weighted average of the group-specific AMCEs, denoted by $\delta_{jk}(l, l')$,

$$\text{CAMCE}_j(l, l'; \boldsymbol{X}_i) = \sum_{k=1}^{K} \delta_{jk}(l, l')\pi_k(\boldsymbol{X}_i). \tag{5.2}$$

By plugging in our estimates $\hat{\pi}_k(\boldsymbol{X}_i)$ and $\hat{\delta}_{jk}(l, l')$, we can estimate the CAMCE.

Section B of the Supplementary Material compares the estimated CAMCE obtained from our method and Robinson and Duch's (2024) (`cjbart`) using the same moderators and treatments. Our method discovers a considerable degree of heterogeneity in the CAMCEs, whereas `cjbart` shows limited treatment effect variation for most countries. Under our model the estimated heterogeneous effects are more strongly associated with predictors than `cjbart`; for example, our method finds a clear association, on average, between the estimated CAMCE and prejudice or party identification whereas `cjbart` does not.

6. Concluding remarks. We have shown that a Bayesian mixture of regularized logistic regressions can be effectively used to estimate heterogeneous treatment effects of high-dimensional treatments. The proposed approach finds maximally heterogeneous groups and yields interpretable results, illuminating how different sets of treatments have heterogeneous impacts on distinct groups of units. We apply our methodology to conjoint analysis, which is a popular survey experiment. Our analysis shows that individuals with high prejudice score tend to discriminate against immigrants from certain non-European countries. These individuals tend to be less educated and live in areas with few immigrants. Future research should consider the derivation of optimal treatment rules in this setting as well as the empirical evaluation of such rules. Another important research agenda is the estimation of heterogeneous effects of high-dimensional treatments in observational studies.

Acknowledgments. We thank Jelena Bradic, Ray Duch, Tom Robinson, Teppei Yamamoto, and participants at the 2021 Joint Statistical Meetings, the University of North Carolina Chapel Hill Methods and Design Workshop, the Bocconi Institute for Data Science and Analytics Seminar, and the 2022 American Political Science Association Annual Meeting for helpful feedback on this paper. We also thank two anonymous reviewers from the Maguro Peer Pre-Review Program at Harvard's Institute for Quantitative Social Science.

Funding. This research was done using services provided by the OSG Consortium (https://doi.org/10.21231/906P-4D78), which is supported by the National Science Foundation awards #2030508 and #2323298. Imai thanks the Alfred P. Sloan Foundation (2020–13946) for partial support. Pashley was partially supported by the NSF Graduate Research Fellowship (DGE1745303). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material for "Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis" (DOI: 10.1214/24-AOAS1994SUPPA; .pdf). It contains mathematical proofs, additional simulation and empirical analyses, and relevant discussions (Goplerud, Imai and Pashley (2025b)).

Replication data for: Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis (DOI: 10.1214/24-AOAS1994SUPPB; .zip). It contains the replication code and data (Goplerud, Imai and Pashley (2025a)).

REFERENCES

ALMIRALL, D., GRIFFIN, B. A., MCCAFFREY, D. F., RAMCHAND, R., YUEN, R. A. and MURPHY, S. A. (2014). Time-varying effect moderation using the structural nested mean model: Estimation using inverse-weighted regression with residuals. *Stat. Med.* **33** 3466–3487.

ANDREWS, R. L., AINSLIE, A. and CURRIM, I. S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *J. Mark. Res.* **39** 479–487.

ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci.* **113** 7353–7360.

BONDELL, H. D. and REICH, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics* **65** 169–177. https://doi.org/10.1111/j.1541-0420.2008.01061.x

BRYANT, P. and WILLIAMSON, J. A. (1978). Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* **65** 273–281.

CELEUX, G., FRÜHWIRTH-SCHNATTER, S. and ROBERT, C. P. (2019). Model selection for mixture models – perspectives and strategies. In *Handbook of Mixture Analysis* (S. Frühwirth-Schnatter, G. Celeux and C. P. Robert, eds.) 118–154. CRC Press/CRC, Boca Raton.

CHAMROUKHI, F. and HUYNH, B.-T. (2019). Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. *J. Soc. Fr. Stat.* **160** 57–85.

CHERNOZHUKOV, V., DEMIRER, M., DUFLO, E. and FERNANDEZ-VAL, I. (2023). Fisher-Schultz Lecture: Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in India. Technical Report. Available at arXiv:1712.04802.

CHI, J. T., CHI, E. C. and BARANIUK, R. G. (2016). k-POD: A method for k-means clustering of missing data. *Amer. Statist.* **70** 91–99.

DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 727–753.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–22.

DE LA CUESTA, B., EGAMI, N. and IMAI, K. (2022). Improving the external validity of conjoint analysis: The essential role of profile distribution. *Polit. Anal.* **30** 19–45.

EGAMI, N. and IMAI, K. (2019). Causal interaction in factorial experiments: Application to conjoint analysis. *J. Amer. Statist. Assoc.* **114** 529–540.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

FIGUEIREDO, M. A. T. (2003). Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **25** 1150–1159.

GELMAN, A., HILL, J. and YAJIMA, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* **5** 189–211.

GOPLERUD, M. (2021). Modelling heterogeneity using Bayesian structured sparsity. Technical Report. Available at arXiv:2103.15919.

GOPLERUD, M., IMAI, K. and PASHLEY, N. E. (2025a). Replication data for "Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis." https://doi.org/10.7910/DVN/YAHPEH

GOPLERUD, M., IMAI, K. and PASHLEY, N. E (2025b). Supplement to "Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis." https://doi.org/10.1214/24-AOAS1994SUPPA, https://doi.org/10.1214/24-AOAS1994SUPPB

GOPLERUD, M., PASHLEY, N. E. and IMAI, K. (2025). FactorHet: Estimate heterogeneous effects in factorial experiments using grouping and sparsity. R package version 1.0.0. https://doi.org/10.32614/CRAN.package.FactorHet

GORMLEY, I. C. and FRÜHWIRTH-SCHNATTER, S. (2019). Mixture of experts models. In *Handbook of Mixture Analysis* (S. Frühwirth-Schnatter, G. Celeux and C. P. Robert, eds.) 271–307. Chapman and Hall/CRC.

GRIMMER, J., MESSING, S. and WESTWOOD, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Polit. Anal.* **25** 413–434.

GRÜN, B. and LEISCH, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *J. Classification* **25** 225–247.

GUPTA, S. and CHINTAGUNTA, P. K. (1994). On using demographic variables to determine segment membership in logit mixture models. *J. Mark. Res.* **31** 128–136.

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Anal.* **15** 965–1056.

HAINMUELLER, J. and HOPKINS, D. J. (2015). The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants. *Amer. J. Polit. Sci.* **59** 529–548.

HAINMUELLER, J., HOPKINS, D. J. and YAMAMOTO, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Polit. Anal.* **22** 1–30.

IMAI, K. and LI, M. L. (2025). Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments. *J. Bus. Econom. Statist.* **43** 256–268.

IMAI, K. and RATKOVIC, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* **7** 443–470.

IMAI, K. and STRAUSS, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Polit. Anal.* **19** 1–19.

JIANG, W. and TANNER, M. A. (1999). On the identifiability of mixtures-of-experts. *Neural Netw.* **12** 1253–1258.

KHALILI, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *Canad. J. Statist.* **38** 519–539.

KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038.

KÜNZEL, S. R., SEKHON, J. S., BICKEL, P. J. and YU, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. USA* **116** 4156–4165.

LEEPER, T. J., HOBOLT, S. B. and TILLEY, J. (2020). Measuring subgroup preferences in conjoint experiments. *Polit. Anal.* **28** 207–221.

LIU, G. and SHIRAITO, Y. (2023). Multiple hypothesis testing in conjoint analysis. *Polit. Anal.* **31** 380–395.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233.

MCLACHLAN, G. J. (1982). The classification and mixture maximum likelihood approaches to cluster analysis. In *Classification, Pattern Recognition and Reduction of Dimensionality*, (P. R. Krishnaiah and L. N. Kanal, eds.) **2** 199-208. North-Holland.

MENG, X.-L. and VAN DYK, D. A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 511–567.

NEWMAN, B. J. and MALHOTRA, N. (2019). Economic reasoning with racial hue: Is the immigration consensus purely race neutral? *J. Polit.* **81** 153–166.

POLSON, N. G. and SCOTT, S. L. (2011). Data augmentation for support vector machines. *Bayesian Anal.* **6** 1–24.

POST, J. B. and BONDELL, H. D. (2013). Factor selection and structural identification in the interaction ANOVA model. *Biometrics* **69** 70–79.

RAO, V. R. (2014). *Applied Conjoint Analysis*. Springer, Berlin Heidelberg.

RATKOVIC, M. and TINGLEY, D. (2017). Sparse estimation and uncertainty with application to subgroup analysis. *Polit. Anal.* **25** 1–40.

ROBINSON, T. S. and DUCH, R. M. (2024). How to detect heterogeneity in conjoint experiments. *J. Polit.* **86** 412–427.

RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ-1-penalization for mixture regression models. *TEST* **19** 209–256.

STOKELL, B. G., SHAH, R. D. and TIBSHIRANI, R. J. (2021). Modelling high-dimensional categorical data using nonconvex fusion penalties. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 579–611.

TIAN, L., SLIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532.

VAN DER LAAN, M. J. and ROSE, S. (2011b). *Targeted Learning*: *Causal Inference for Observational and Experimental Data*. Springer, Berlin.

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242.

WU, C. J. and HAMADA, M. S. (2021). *Experiments*: *Planning*, *Analysis*, *and Optimization*, 3rd ed. Wiley, New York.

YAN, X. and BIEN, J. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statist. Sci.* **32** 531–560.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67.

ZAHID, F. M. and TUTZ, G. (2013). Ridge estimation for multinomial logit models with symmetric side constraints. *Comput. Statist.* **28** 1017–1034.