Generating Precise Ad Targeting Hypotheses for Scalable Testing

Brian Goodchild (Columbia)

Abstract

1 Introduction

We are working towards designing a system to determine how features of users such as personal and behavioral information affect the targeted advertisements they are served online. More specifically, we want to use *observational data* collected from participating users to generate sound *targeting hypotheses* that attempt to explain the ads they receive. This differs from common approaches which rely on randomized controlled experiments, where user actions and information are assigned randomly to fake accounts, then ads are collected from these fake accounts. The main problem we face is dealing with confounding factors that are correlated with both user behavior and the recommendations they are served.

We have a few techniques in the works for generating targeting hypotheses. The details of these techniques are explored in section 3. At a high level, hypothesis generation is the step during which, for each recommendation, a, a targeting hypothesis, H_a, is generated, and is interpreted as a disjunctive set of features, any combination of which, can cause the recommendation a to be served to a user. This step is akin to variable selection (in fact, we use regression models with lasso penalties) and requires that we divide our users into training and testing sets for generating and testing hypotheses respectively.

Currently, we are evaluating the effectiveness of different hypothesis generating techniques using data from a semi-controlled experiment. We constructed a set of 300 fake Facebook profiles and assigned features to each in the form of demographic information and page "likes." Over time, we had the accounts perform the page likes and collect the advertisements displayed. These advertisements are the targeted recommendations and subsets of the user features are used as targeting hypotheses. The features are not randomly assigned, but rather

feature	conditioned on	feature values
ethnicity	assigned uniformly from one of four options, each with probability 1/4	Baseline, AsianAmerican, HispanicAmerican, AfricanAmerican
gender	based on Facebook's gender distribution (47% male, 53% female)	Male, Female
age bracket	*gender, ethnicity*	[18-24], [25-34], [35-44], [45-54], [55-64], [65+]
education	*gender, ethnicity, age bracket*	high school, college, graduate school
income	*gender, ethnicity, age bracket, education*	[30k-50k], [50k-100k], [100k-250k]
politics	*gender, ethnicity, age bracket, education, income*	conservative, liberal
relationship status	*gender, ethnicity, age bracket*	single, in relationship, married
parental status	*gender, ethnicity, age bracket, relationship status*	has children

assigned based on distributions gleaned from Facebook's own audience insights tool.

This report is structured as follows. In section 2, we explain our experimental design and data collection process. In section 3 we describe three hypothesis generation techniques for use in randomized, observational, or semi-randomized studies. In section 4 we test these hypotheses by calculating propensity-weighted average treatment effects and use bootstrapping to estimate population parameters for significance testing, and in section 5 we go over some initial empirical results.

2 Dataset and Methodology

Our dataset consists of 300 users, each assigned demographic features from the set {*gender, age, ethnicity, education, income, politics, relationship status, parental status*} and page likes from a set of 1633 distinct pages. The advertisements served to the users amounted to 580,780 over six months with 17,326 distinct ad campaigns. User features were assigned in the following manner:\

Let D_i be the set of demographic features for user i, then the set of pages to

like is assigned as follows:

$$P(\textit{user } i \textit{ likes page } j) = c \cdot \alpha_{D_i,j} \cdot \frac{\#\textit{ users who liked page } j}{\#\textit{ total users}}$$

Where c is a constant and $\alpha_{D_i,j}$ is the "affinity," or multiplicative increase in probability, for a user matching on demographic features D_i to like page j. Note that $\alpha_{D_i,j} \geq 1$ since we have no way of determining a "negative" affinity towards a page for any group of Facebook users who match on a set of demographic features.\

After generating user profiles, we conducted a controlled experiment in which a Facebook account was created for each user. Over 6 months, (March - August, 2017), batches of users logged into the accounts, liked a random subsets of their assigned pages, then scraped the displayed ads. This means that there is an associated timestamp for each page like and for each ad. We will use this fact in the following section.

3 Generating Targeting Hypotheses

The goal of Hypothesis generation is to form H_a, a set of user features that we suspect are targeted by an advertisement a. We consider only disjunctive hypotheses, in which any combination of the features in H_a may have individually, or together, caused the ad to be targeted to some users. Hypothesis generation is an application of the well-known technique of feature selection. Here we cover three techniques, beginning with a simple lasso-style logistic regression; then we extend the regression to utilize time; finally, we end with a group-lasso-style regression that utilizes time and attempts to control for unobserved confounding variables.

Lasso Any feature selection method could suffice to generate targeting hypotheses. However, we typically want to generate *small* hypotheses, since larger ones are difficult to test and interpret. We use variations of constrained logistic regression models to achieve this goal. The most basic of these models builds a lasso-style logistic regression using observed user features and responses.

A study consists of a set U of m users, each with n observed features (page likes and demographic variables in our case). We define the design matrix $X \in \{0,1\}^{m \times n}$ and response $Y^a \in \{0,1\}^m$ as:

$$X_{i,j} := \mathbb{1}[\text{user } \ i \text{ assigned feature } j] \tag{1}$$

$$Y_i^a := \mathbb{1}[\text{user } i \text{ served ad } a] \tag{2}$$

We model the log odds of being served an ad given observed features:

$$log\left(\frac{P(Y_i^a = 1|X_i)}{P(Y_i^a = 0|X_i)}\right) = b^T X_i \tag{3}$$

$$b = \underset{\beta \in \mathbb{R}^n}{\arg\min} -l(\beta|X) + \alpha||\beta||_1 \tag{4}$$

Where α is a hyperparameter governing the severity of the lasso penalty and $l(\beta|X)$ is the log-likelihood function:

$$l(\beta|X) = \left[\frac{1}{m} \sum_{i=1}^{m} Y_i^a \cdot (X_i^T \beta) - \log(1 + e^{X_i^T \beta}) \right] \tag{5}$$

Here, we disregard time and model both X and Y^a as they appear at the end of the experiment. In other words, we consider only whether a user will like some page or be served some ad at any time throughout the course of the experiment, and disregard the relative ordering of these events. A more optimal approach would consider the state of a user *at the time they were served ad a*.

Time-adjusted Lasso When using purely observational data, we expect that intervention will be required to test the hypotheses we generate. Intervention generally requires creating accounts, which drastically limits scalability. Thus we value the precision and accuracy of hypothesis generating techniques.

In an effort to generate hypotheses with higher precision and accuracy, we leverage time. In this approach, our response variable Y^a remains the same. However, rather than modeling $X_{i,j}$ as a simple dummy representing feature assignment, we now consider the actual actions performed by each user at the time they were first served ad a, or the actions performed by the last time *any* user was served a in the case of users who are not served the ad. In this setting, demographic variables can be thought of as actions performed before the start of the experiment, while page likes are actions performed at distinct times throughout.

Let U_a be the set of users served ad a, with $t_{i,j}, t_{i,a}, i \in U_a$ being the time at which user i performed action j or was served ad a respectively, and T_a the last time any user was served ad a. We construct X^a as follows:

$$X_{i,j}^a := \begin{cases} \mathbb{1}[t_{i,j} < t_{i,a}] & if\ i \in U_a \\ \mathbb{1}[t_{i,j} < T_a] & otherwise \end{cases} \tag{6}$$

And our logistic regression changes only in the design matrix:

$$log\left(\frac{P(Y_i^a = 1|X_i^a)}{P(Y_i^a = 0|X_i^a)}\right) = b^T X_i^a \tag{7}$$

$$b = \underset{\beta \in \mathbb{R}^n}{\arg\min} -l(\beta|X^a) + \alpha||\beta||_1 \tag{8}$$

Claim1: Time-adjusted lasso regression adequately controls for time.

Time-adjusted Group Lasso We claim that the previous example should work well when the confounding variables, such as demographics, are explicitly known. However, in the general case, we will not always have access to such information. In an attempt to control for unknown confounders, we again leverage time. Here we model user actions with two features – one signifying that a user has performed an action by a particular time, and one signifying that a user has been "assigned" the action to perform, regardless of whether it has yet been performed. Recall that assignment of pages to like was an explicit stage in our experimental design; however we do not consider this to be the general case. With data collected from observational studies, variable "assignment" can be modeled as the dummy variable that is equal to 1 when a user was witnessed to have performed the action associated with the variable (e.g. liked the page) at any time throughout the course of the experiment.

Let X and X^a be defined as in def. 1 and 6 respectively. We define our new design matrix $\mathbb{X} \in \{0,1\}^{m \times 2n}$:

$$\mathbb{X} := [X^a, X] \tag{9}$$

In order to maintain precision, we now restrict our lasso penalty to pertain to groupings of the elements in the coefficient vector β. Let $g(j) : \mathbb{Z}^+ \mapsto \mathbb{Z}^+ \times \mathbb{Z}^+$ be a function that maps the index of feature j to two column indices k, k' such that $\mathbb{X}_k = X^a_j$ and $\mathbb{X}_{k'} = X_j$. Our new group lasso regression now works as follows:

$$log\left(\frac{P(Y^a_i = 1|\mathbb{X}_i)}{P(Y^a_i = 0|\mathbb{X}_i)}\right) = b^T \mathbb{X}_i \tag{10}$$

$$b = \underset{\beta \in \mathbb{R}^n}{\arg\min} -l(\beta|\mathbb{X}) + \alpha \sum_{j=1}^{n} ||\beta_{g(j)}||_2 \tag{11}$$

For each group $g(j)$, we have two features $\mathbb{X}_k$, which captures the effect of performing an action, and $\mathbb{X}_{k'}$ which models the effect of a user's proclivity to like a page. Now, consider the nonzero coefficients of b. By enforcing the group lasso penalty, we ensure that either $b_k, b_{k'}$ are both zero or both nonzero. We now have method of estimating the effect of performing actions while controlling for unobserved confounders, and can interpret the coeffecients as:

Case I: $b_k > b_{k'} \Rightarrow$ The effect of some unobserved confounder is stronger than the effect of the performing the action j.\ **Case II:** $b_k < b_{k'} \Rightarrow$ The effect of performing the action j is stronger than the effect of any unobserved confounders.\ **Case III:** $b_k = b_{k'} \Rightarrow$ Inconclusive (and highly unlikely).

In **Case I**, we may omit action j from our hypothesis, H_a, or we may choose to investigate further what confounders might be at work (potentially using some kind of factor analysis). In **Case II:** we have strong evidence that action j is targeted, and thus we include it in H_a.

Claim2: Time-adjusted group lasso regression controls for time and does not overcorrect for unobserved confounders.

4 Estimating Treatment Effects

After generating H_a, our disjunctive hypothesis consisting of page and/or confounder variables, we gather the users from the testing set in an observational study in which we have two groups, *treatment* and *control*, and interpret H_a as a disjunctive treatment such that any user who matches at least one of the features in H_a is considered to "satisfy" the hypothesis, and thus is in the treatment group, and any user who does not match on any is considered a control. Let $Z_{i,a} := \mathbb{1}\{\textit{user } i \textit{ satisfies } H_a\}$. In all notation moving forward, when we omit the a, it is assumed that we are talking about an arbitrary but particular ad a.

Our goal is to estimate the Average Treatment Effect, which can be stated as the expected difference in the observed and counterfactual outcomes given the treatment assignment:

$$ATE = \mathbb{E}[Y_i^1 - Y_i^0 | Z_i] \tag{12}$$

Propensity Scoring Since we do not have randomization, we cannot substitute the control group for the counterfactual as is typically done in randomized control trials. However, we have access to the actual distributions of our confounded features, since we generated them ourselves.

Thus far we have been using the *Inverse Propensity Scoring* (a.k.a. *Inverse Probability of Treatment Weighting*) estimator:

$$\hat{ATE}_{IPS} = \frac{\sum_{i=1}^{n} (1/p_i) Y_i Z_i}{\sum_{i=1}^{n} Z_i (1/p_i)} - \frac{\sum_{i=1}^{n} [1/(1-p_i)] Y_i (1-Z_i)}{\sum_{i=1}^{n} (1-Z_i)[1/(1-p_i)]} \tag{13}$$

Where the *propensity score* $p_i := P(Z_i = 1 | C_i)$ and C_i is the set of potential confounding variables, which for us is the demographic features used to generate the page distributions, discussed in section 2.\

Significance Testing To test the significance of our estimators, we use the bootstrapping method, where we resample with replacement from the set of users, then calculate $\hat{ATE}_{IPS}$ for 1000 repeated iterations. From these estimates, we are able to estimate the sample standard deviation of the mean $\hat{s}$ which we can use to calculate our test statistic $z = \hat{ATE}/\hat{s}$, which we can compare to a normal distribution with mean 0 and variance 1.

5 Empirical Results

We ran our experiment from March - August 2017, when our fake user accounts were recognized and removed by Facebook. Throughout these six months we were able to collect 17,326 unique ads across 580,780 impressions served to our 300 fake accounts. In order to ensure we can test hypotheses, we prune ads that were displayed fewer than 5% or greater than 95% of users, leaving us with 1,314

unique ads. Our users were assigned demographic (feature, value) pairs derived from the eight demographic features described in section 2. The pages assigned to users based on these features were drawn from a set of 1633 unique Facebook pages.

Our initial evaluation compares regular lasso, time-adjusted lasso, and time-adjusted group lasso hypotheses. For time-adjusted lasso, we include all demographic features in the regression, and consider this to be an approximate "ground truth." We use the inverse propensity scoring methodology in section 4 to estimate the average causal effects of hypotheses generated by both methods. Recall that the goal of hypothesis generation is to find hypotheses that are likely to be true. Table 5 shows an abbreviated version of the end-to-end results on one ad for the three methods.

In total, we were able to uncover statistically significant ($p \leq 0.05$) targeting hypotheses for 54 ads using regular lasso (lasso), 151 with time-adjusted lasso with demographic confounders (TAlasso), and 332 with time-adjusted group lasso (TAGlasso). This is surprising, given that time-adjusted lasso was expected to be our "ground truth." This surprising result suggests that we might need to work out a more rigorous method for hypothesis testing, since we would expect the regression which includes explicit confounders to perform better than one that models them implicitly. Moving forward, we will consider the results of the TAlasso regression as ground truth in the presence of demographic-confounder-based targeting.

The intuition behind our time-adjusted group lasso approach is that *page_assigned* variables should control for unobserved confounders. Additionally, it would be helpful if, for a fixed page j, $page\_assigned_j > page\_liked_j$ directly implies the targeting of an unobserved confounder. To test this, we compared the hypotheses generated by TAGlasso to the statistically significant hypotheses generated by TAlasso in which demographic-confounder-based targeting was present. We found that, of the 60 ads containing hypotheses with demographic confounders, there is at least one j in the TAGlasso hypothesis such that $page\_assigned_j > page\_liked_j$ for 53 of them, leaving only 7 "false negatives" for which this is not the case. Additionally, of 51 statistically significant hypotheses generated by TAlasso that do not contain demographic confounders, TAGlasso generated hypotheses such that for all pages, j, in the hypothesis, $page\_liked_j > page\_assigned_j$, and there were only 7 "false positives" for which this is the case. This would suggest that TAGlasso is fairly effective at suggesting the presence of confounder-based targeting, even in the absence of explicit information about these confounders.

Finally, we would hope that the pages that appear in TAGlasso hypotheses would be associated with the confounders in TAlasso hypotheses for the same ad. Since we have access to the ground-truth distributions of pages across confounders, we were able to compare these hypotheses. Figure 1 shows demographic affinities for pages uncovered by TAGlasso, for which $page\_assigned > page\_liked$. The x axis shows all demographic values that appear across TAlasso hypotheses. For

method	**hypothesis**	$\hat{ATE}$	p	b
lasso	{facebook.com/thebeatles/}	0.3067738	0.0000126	[0.56734(page, 911)]
TAlasso	{facebook.com/thebeatles/}	0.3067738	0.0000075	[9.397(page, 911)]
TAGlasso	{facebook.com/thebeatles/}	0.6135477	0.0000044	[1.61438(page_assigned, 911), 5.02425(page_liked, 911)]
lasso	{facebook.com/DailyCaller/}	0.0298173	0.3548779	[0.01664(page, 1277)]
TAlasso	{age_braket: [65, -1]}	0.7783333	0.0000175	[0.05233(age_braket, [65, -1])]
TAGlasso	{facebook.com/numbersusa/, facebook.com/DailyCaller/}	0.0659713	0.3115039	[0.16870(page_assigned, 980), 0.027046(page_liked, 980), 0.78680(page_assigned, 1277), 0.03904(page_liked, 1277)

Table 1: Results of two regression analyses using our three methods. The text fields of each advertisement were: 1) *"Meet Julian Lennon stores.barnesandnoble.com For a reading/signing..."* and 2) *"NYSDOH - New York State Health Department If you're 50 or older get screened for colon cancer..."* In the first ad, all three methods agree that the page "thebeatles" was being targeted, with *page_liked* > *page_assigned* in the case of time-adjusted group lasso. In the second case, time-adjusted lasso found a targeted demographic confounder, which is reflected in the fact that *page_liked* < *page_assigned* for both groups in the time-adjusted group lasso regression.

all confounder values, c in a given TAlasso hypothesis, H_a^{TA}, we gathered all pages in the TAGlasso hypothesis for the same ad, H_a^{TAG}. Next for each page $j \in H_a^{TAG}$, if $page\_assigned_j > page\_liked_j$, we plot $affinity(c, j)$ which is defined as the multiplicative probability that a Facebook user with confounder value c will like page j compared to the baseline Facebook user. In most cases we find that the confounder values indeed have a higher affinity for these pages, with the [65+] age bracket being the most pronounced.

Does time-adjusted group lasso signify the presence of confounder-based targeting? %%% End document

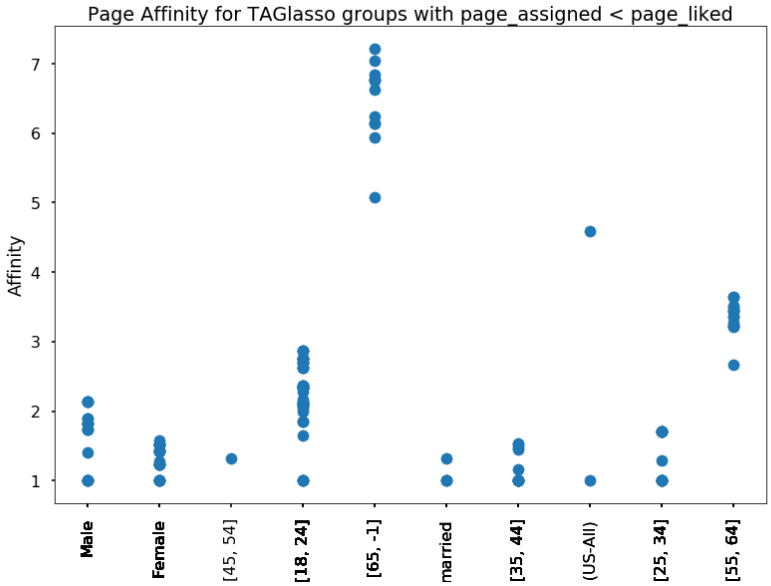

Figure 1: Page affinities for pages that appear in TAGlasso hypotheses for each of the demographic values targeted by the same ad. The values on the x axis are demographic values uncovered by TAlasso run with demographic confounders explicitly added to the design matrix. Affinity is defined as the multiplicative probability that a user matching the demographic value will like the page compared to the baseline Facebook user.