A Systematic Paradigm for Detecting, Surfacing, and Characterizing Heterogeneous Treatment Effects (HTE)

John Cai, Weinan Wang @ Snap Inc.

1 Motivation and Introduction

At Snap, we run thousands of online controlled experiments to evaluate how feature changes affect a diverse range of users. Experimenters often propose a set of metrics (or one) as the OEC\(^1\) for success measurement and examine the ATE\(^2\) or QTE\(^3\) at pre-specified locations. In addition, to gauge how different users are affected by a treatment, we often use a pre-defined set of breakdowns (e.g., countries, age bucket, gender, platform, etc.) to estimate the HTE\(^4\).

However, this manual approach is not only cumbersome but also subject to the experimenter’s cognitive biases. Without a principled approach for global detection, experimenters waste time chasing down HTE when it does not exist. Without a way to surface and rank all HTE dimensions, experimenters rely on their pre-conceived notions of which dimensions are important, a practice that leads to confirmation bias and missed discoveries. Moreover, we also do not have a good way to precisely quantify how much each breakdown contributes to HTE. Last but not least, we need a systematic understanding of the distribution of such treatment effects beyond the average and certain pre-specified quantiles.

In this paper, we aim to tackle the aforementioned problems to enable a deeper understanding of HTE:

- **Detection**: Global test to determine whether there’s HTE present or not.
- **Surfacing**: Systematically decompose treatment effect variation, measure how much variation can be explained by known breakdowns, and surface the HTE dimensions that explain most of the variation.
- **Characterization**: Model the distribution of individual treatment effects as a mixture of different conditional distributions, enabling further interpretable understanding of user level treatment effects.

2 Related Work

A popular model-based approach to detect HTE is to test if the conditional average treatment effects (CATE) are statistically different from each other when defined over a set of covariates \([\text{Chang et al. (2015)}]\). However, this approach tests for pre-defined set of breakdowns (e.g., countries, age bucket, gender, platform, etc.) to estimate the \(QTE\).

Experimenters often propose a set of metrics (or one) as the way to estimate the HTE. Last but not least, we need a systematic understanding of the distribution of such treatment effects beyond the average and certain pre-specified quantiles.

3 Methodology and Examples at Snap

In this section, we briefly describe the methodologies involved in solving the three aforementioned problems. Without loss of generality, we focus on the case of a randomized controlled experiment with treatment indicator \(Z_i\) and continuous outcome \(Y_i\) for the \(i\)-th user. Let \(N\) be the total number of users in the study, with \(N_1\) in treatment and \(N - N_1 = N_0\) in control.

Invoking the SUTVA condition, each individual’s observed outcome can be denoted as: \(Y_i^{\text{obs}} = Z_iY_i(1) + (1 - Z_i)Y_i(0)\). Define the \(ITE\) the usual way as \(\tau_i = Y_i(1) - Y_i(0)\).

3.1 Global HTE Detection

3.1.1 Methodology

In general, following \(\text{Cox (1984)}\), we can assess treatment effect heterogeneity by examining the marginal variances of the treatment and control groups’ respective outcomes. Specifically, if the overall treatment effect is constant, i.e. \(Y_i(1) = Y_i(0) + \tau\), then \(\text{Var}(Y_i(1)) = \text{Var}(Y_i(0))\). This makes the test statistic of variance ratio \(t_{\text{var}} = \frac{\sigma^2_1}{\sigma^2_0}\) sensible.

However, we would not expect the \(t_{\text{var}}\) to follow an \(F\)-distribution since the marginal distributions of the potential outcomes are not normal for metrics at Snap. Hence, we appeal to a more general theorem that uses higher order moments:

**Theorem 1.** Assume that \(Y(z)\) has finite kurtosis \(\kappa_z\). Under the null of equal variance,

\[
t_{\text{var}} = \frac{\log \sigma^2_1 - \log \sigma^2_0}{\sqrt{(s_1 - 1)/N_1 + (s_0 - 1)/N_0}} \overset{d}{\rightarrow} \mathcal{N}(0,1)
\]

Additionally, we explore the use of a bootstrapped \(F\)-test for the variance of the mean. While the metrics are unlikely to be normally distributed, we can appeal to the central limit theorem to compute bootstrapped means, which we can then compute the variance of. This variance would be tightly related to the variance of the metrics.
3.1.2 Simulation Results for HTE Detection

To demonstrate the power of this method, we run a small simulation comparing the proposed method (Kurtosis) against the bootstrapped F-test (200 bootstraps) and the Fisher’s Randomization coupled with KS Test (FRT) in [Ding et al. (2016)]. Specifically, for a total of \(N = 5000\) users, we simulate AA period data for all as negative binomial distribution with parameter \(A_i \sim NB(\mu = 200, \phi = 0.15)\), and AB period data for control as \(Y_i(0) \sim A_i + N(0, 1)\), and treatment effect \(\tau_i \sim LN(\mu = 2, \sigma^2)\) and we vary \(\sigma\) from 0 to 0.15, simulating scenarios where there’s no treatment effect variation to strong treatment effect variation. A total of 500 simulations are conducted. We compare the power of these three procedures by utilizing the BH procedure with nominal FDR level at \(\alpha = 0.05\) on the resulting p-values and compare the proportion of detected significant results.

Going forward, we use the kurtosis-based log-variance test, as it has better power and is also robust to non-normality. Formally, we compute the two-sided p-value: \(2\Phi(-|t_{kvar}|)\) for global HTE detection on any given OEC.

3.1.3 Empirical Results at Snap

To check if the method is working as intended on real AB data, we apply the kurtosis-based log-variance test on AA and AB data. We anonymize all the metrics that we have computed the results on, and only make the distinction between whether a metric is a count-based metric (e.g. count of views) versus whether it is a days-based metric (e.g. number of active days). For the former, it is generally unbounded above 0, while for the latter, there is an upper bound in the number of days that have elapsed during the experiment.

From above, we observe that the AA p-values are generally uniformly distributed. In the AA period (pre-experiment), no treatment has been applied yet. Hence, there is no systematic difference in distribution between the treatment and control group. Therefore, we would expect the p-values to be uniformly distributed, as shown above.

For A/B data, we observe that days metric (e.g. Metric A Days, Metric D Days) tend to have a higher proportion of p-values < 0.05. Days metric measures the number of days that a user has been active over a given period. The heterogeneity in the metric could be attributed to its inherently bounded nature, as a user who is active for all the days of the experiment cannot increase their active days.

To correct for false discoveries that occur due to the metrics’ inherent distribution, it is important to apply false discovery control methods. We use the past distribution of the p-values for every metric, and apply the Benjamini-Hochberg false discovery rate control procedure from [Benjamini and Hochberg (1995)]. With this adjustment, we can accurately control the false discovery rates, and reduce the effects of a metric’s inherent susceptibility to HTE.

3.2 HTE Surfacing through Most Likely Breakdowns

3.2.1 Methodology

To surface the dimensions that contribute the most towards treatment effect variation, we construct a \(R^2\)-like measure of the ratio of explained treatment effect variation to total treatment effect variation. This measure is similar in construction to the one described in [Ding et al. (2019)]. For simplicity, suppose that the treatment effect was correlated with existing covariates \(X^T\), and that an individual takes on the following potential outcomes in control and treatment, i.e. \(Y_i(0), Y_i(1)\); let \(\epsilon_i\) and \(\tilde{\epsilon}_i\) denote the error terms in the control outcomes and the treatment effect respectively.

\[
\tau_i = X^T_i \beta + \epsilon_i, \quad Y_i(0) = X^T_i \gamma + \epsilon_i, \quad Y_i(1) = X^T_i (\gamma + \beta) + \epsilon_i + \tilde{\epsilon}_i.
\]
We begin by estimating a model of how the control outcomes vary with \( X \), using an OLS model. In our analysis, we focus on categorical variables. Hence, we estimate our model using the conditional means of each category.

\[
\hat{Y}_i(0) = X_i^T \hat{\gamma}, \quad \hat{Y}_i(1) = X_i^T (\hat{\gamma} + \hat{\beta}).
\]

The **Total Treatment Effect Variation (TEV)** can be decomposed into two components, the explained and the idiosyncratic variation. To obtain the explained treatment effect variation, we directly estimate the variation in the conditional average treatment effects, which is the difference between the predicted treatment and control outcomes (\( \hat{Y}_i(1) - \hat{Y}_i(0) \)).

Total \( \text{TEV} = \text{Var}(\tau_i) = \text{Var}(X_i^T \hat{\beta}) + \text{Var}(\epsilon_i) \), Explaned \( \text{TEV} = \text{Var}(X_i^T \hat{\beta}) = \text{Var}(\hat{Y}_i(1) - \hat{Y}_i(0)) \).

\[
\text{Idiosyncratic TEV} = \text{Var}(\epsilon_i), \quad R^2 = \frac{\text{Explaned TEV}}{\text{Total TEV}} = \frac{\text{Var}(X_i^T \hat{\beta})}{\text{Var}(X_i^T \hat{\beta}) + \text{Var}(\epsilon_i)}.
\]

Estimating the idiosyncratic TEV \( \text{Var}(\epsilon_i) \) is more complicated, as we do not observe each individual under both treatment and outcome states. Hence, we use the bounds proposed in Ding et al. (2019), but with modifications to increase the conditional average treatment effects, which is the difference between the predicted treatment and control outcomes (\( \hat{Y}_i(1) - \hat{Y}_i(0) \)).

1. Obtain the residuals after removing the predicted outcomes from the actual outcomes. This gives us: \( \hat{\epsilon}_i(0) = Y_i(0) - X_i^T \hat{\gamma} \) and \( \hat{\epsilon}_i(1) = Y_i(1) - X_i^T (\hat{\gamma} + \hat{\beta}) \). For treatment, it consists of both the treatment effect error term \( \epsilon_i \) and the control outcome error term \( \epsilon_i \), while for control, it consists of just the control outcome \( \epsilon_i \).
2. Group all residuals using each category of \( X_i^T \). This leaves us with \( M \) strata, where each strata has \( N_{m_0} \) individuals in control and \( N_{m_1} \) individuals in treatment.
3. Sort residuals within each strata s.t. \( \hat{\epsilon}_i(k) \geq \hat{\epsilon}_i(k+1) \), \( \forall k \in \{2, ..., N_{m_0} \} \) and \( \hat{\epsilon}_i(k) \geq \hat{\epsilon}_i(k+1) \), \( \forall k \in \{2, ..., N_{m_1} \} \).
4. Map each residual to a rank within each strata. WLOG, suppose that for each strata \( N_{min} = N_{m_0} \). Then, define \( S_j(1) = \{ \hat{\epsilon}_i(1) \mid (\{k \cdot (N_{min}/(N_{m_1} + 1)) + 1 \} = j \} \). The treatment group's residual used in counterfactual matching would then be the average of the set: \( \hat{\epsilon}_j = \frac{1}{|S_j|} \sum_{\hat{\epsilon}_i \in S_j} \hat{\epsilon}_i \).
5. Estimate the residual difference for both treatment and control: \( d_j(0) = \hat{\epsilon}_j(1) - \hat{\epsilon}_j(0), \quad d_j(1) = \hat{\epsilon}_j(1) - \hat{\epsilon}_j(0) \).
6. Obtain \( \text{Var}(d_i) = \frac{1}{N} \sum_{i=1}^{N} (d_i)^2 \), as \( E(d_i) = 0 \) by construction. This is the discrete analogue of the Fréchet-Hoeffding bound when we only observe the marginal distributions: \( \frac{1}{N} \sum_{i=1}^{N} (d_i)^2 \approx \int_{0}^{1} \{ F_{1}^{-1}(u) - F_{0}^{-1}(u) \}^2 du \leq \text{Var}(\epsilon_i) \) where \( F_1^{-1} \) and \( F_0^{-1} \) are the empirical CDF of \( \hat{\epsilon}_1(1) \) and \( \hat{\epsilon}_0(0) \) respectively (Aronow et al. (2014)).
7. Under local rank-preservation between \( v_j(1) \) and \( v_j(0) \) within each strata, we obtain the idiosyncratic TEV \( \text{Var}(\epsilon_i) = \text{Var}(d_i) \). Our assumption of local rank-preservation within each strata is less stringent than the assumption of global rank-preservation used in Ding et al. (2019) (**Theorem 2.3**). As a corollary, our \( R^2 \) bounds are sharper.
8. Combine the idiosyncratic TEV with the explained TEV to obtain the upper bound of \( R^2 \).

### 3.2.2 Simulation Results for \( R^2 \) Computation

To understand how well our method bounds the \( R^2 \) compared to the true \( R^2 \), we run a simulation where we know the counterfactual treatment and control outcomes. We also compare the performance of the stratified bound versus the unstratified bound (in Ding et al. (2019)), and demonstrate that our bound is tighter.

To simulate our results, we assume that the control outcomes error \( \epsilon_i \) are drawn from a negative binomial distribution of \( NB(30, 0.5) \). Negative binomial distributions are appropriate in modeling discrete count-based metrics (Hilbe (2011)). The control outcome is then modelled by: \( Y_i(0) = X_i^T \gamma + \epsilon_i \), with \( X_i \) being a categorical variable with 4 possible values.

In each category, we have 500,000 users. The treatment effect error term is then modeled using the marginal distribution \( \epsilon_i \sim NB(\mu_i, 0.5) \). Here \( \mu_i \) varies across different strata.

Subsequently, we link the control outcome error \( \epsilon_i \) and the treatment effect error \( \epsilon_i \) using a Gaussian copula with varying levels of correlation \( r \). Finally, we generate a treatment assignment vector \( Z_i \).
From above, we have two key insights. First, we observe that the upper stratified bound is tighter than the unstratified bound across all ranges of the correlation. This is expected because we induce additional rank perturbations across the strata, as captured by the different treatment effect distributions simulated. Second, we observe that as the correlation of the error terms increases from 0 to 1, the bounds of the $R^2$ estimate converge to the True $R^2$. Intuitively, as the dependence captured by the Gaussian copula increases, the treatment effect will be more locally rank-preserving, which allows the locally stratified $R^2$ to converge to the true ground-truth $R^2$.

3.2.3 HTE Surfacing applied at Snap

Using an Ads ranking experiment conducted at Snap as illustration, we examine an anonymized metric that is shown to have global HTE through the kurtosis based variance test mentioned in the previous section, and rank the upper bound of the $R^2$ measured on known popular categorical breakdowns:

From above, we see that $df_{\text{non\_friend\_recommended\_story\_view\_L7}}$ ranks the highest in explanatory power for the treatment effect variation, with an upper bound $R^2 \approx 19\%$. This metric captures the number of days in the past week that a user views a recommended story. As the experiment was designed to better utilize signals from previously recommended content, past engagement with recommended content would naturally be highly correlated with the treatment effect. Other breakdowns that are tightly related to the top breakdown, like $df_{\text{non\_friend\_story\_view\_L7}}$ and $story_{\_L7}$ are also ranked second and third respectively. These are more generic versions of the top breakdown.

3.3 Distributional Characterization of Individual Treatment Effects (ITE)

Now, we have the essential tools for detecting the existence of HTE and finding the one-dimensional covariate that can explain the greatest amount of treatment effect variation. In this section, we go beyond the decomposition of the global ATE with a mixture of CATE on the selected covariate, and characterize the global ITE distribution as a mixture of conditional distributions for each strata.

As mentioned above, we already have a sensible estimate for the residual differences $\hat{d}_i(0), \hat{d}_i(1)$, for each user, we would approximate the ITE $\tau_i$ as a combination of the estimated CATE and the idiosyncratic treatment effect:

$$\hat{\tau}_i = \mathbf{X}_i^T \hat{\beta} + \hat{d}_i(0), \text{ for } i \text{ in control};$$
$$\hat{\tau}_i = \mathbf{X}_i^T \hat{\beta} + \hat{d}_i(1), \text{ for } i \text{ in treatment}.$$

We can subsequently pool all estimates $\hat{\tau}_i$ together and surface conditional ITE distributions for each strata $\mathbf{X}_i^T$, the vertical dashed lines indicate CATE (P5 to P95 are shown for visualization):

As seen above, for the strata that $df_{\text{non\_friend\_recommended\_story\_view\_L7}}$ is 0, the ITE distribution is largely a point mass near 0, indicating a very small treatment effect with small dispersion. This is contrasted with other conditional distributions, which have some dispersion above 0. Intuitively, users with no activity in the past do not have signals that the model in the treatment can use to recommend better content, resulting in negligible benefit to these users.
As we move upwards the stratification in previous engagement days from 1 to 7, we see the ITE distribution showing more of a mixture structure of positive impact and the 0 point mass, with the positive impact’s mode increasing as well. We also observe a longer right-tail for distributions after previous engagement days reaches 4.

Furthermore, we can visualize the global ITE distribution as a weighted mixture of these empirical distributions within each strata:

For the sake of visualization, we removed the 0 cohort as it’s mostly a point mass and thus less interesting, and group strata 1 and 2 together as they are very similar in shape. We clearly see the value in decomposing the distribution of treatment effect based on the strata, as the conditional distributions not only have wildly different means, but they also have highly varying modes and tail distributions. Interestingly, we observe that at the right tail, beyond the peak of the main distribution, most of the tail treatment effects come from strata 7 users. Moreover, the results also suggest that further investigation to add sub-stratas on top of strata 5, 6, 7 could help to shed light into the composition of these conditional distributions.

A Appendix

Theorem 2. Global rank preservation implies local rank preservation

Let \( x_i \) be a series of sorted values s.t. \( x_i \geq x_{i-1} \forall i \in \{2, \ldots, N\} \). Let \( f(.) \) be a globally rank-preserving function s.t. \( f(x_i) \geq f(x_{i-1}) \forall i \in \{2, \ldots, N\} \).

Split the set of \( Q = \{x_i | i \in \{2, \ldots, N\}\} \) into any partition \( P \), where \( \cup_{i \in J} (P_j) = Q \) and \( P_j \cap P_k = \emptyset \) for \( j, k \in J \) with \( j \neq k \). A locally-rank preserving function is a function \( g(.) \) s.t. \( \forall x_m, x_m \in P, x_n \geq x_m \Rightarrow g(x_n) \geq g(x_m) \).

Since \( f(.) \) preserves the rank globally, it can be easily seen that when comparing any \( x_n, x_m \) within any partition \( P \), whenever \( x_n \geq x_m \), we will have \( f(x_n) \geq f(x_m) \). Hence, global rank preservation implies local rank preservation.

Theorem 3. Local rank preservation does not imply global rank preservation

Suppose that we have two partitions \( P_j, P_k \) where inf(\( P_j \)) > sup(\( P_k \)). Let our function \( g(.) \) have a locally rank-preserving functional form below:

\[
g(x) = \begin{cases} x + \text{sup}(P_j), & \text{if } x \in P_k \\ x, & \text{if } x \in P_j \end{cases}
\]

Within each partition, we apply a monotonically increasing function, and is hence locally rank-preserving. However, when comparing across partitions, we can easily see that \( g(x_j) < g(x_k) \forall x_j \in P_j, x_j \notin P_j \), even though \( x_j \geq x_k \). This is in contradiction to the definition of global rank preservation, which requires that \( x_j \geq x_k \Rightarrow g(x_j) \geq g(x_k) \forall x_j, x_k \in Q \).

References

Aronow, P. M., D. P. Green, and D. K. Lee (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics* 42(3), 850–871.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1), 289–300.

Chang, M., S. Lee, and Y.-J. Whang (2015). Nonparametric tests of conditional treatment effects with an application to individual treatment effect. *The Annals of Statistics* 43(3), 875–903.

Cox, D. R. (1984). Interaction. *International Statistical Review/Revue Internationale de Statistique*, 1–24.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitznik (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* 90(3), 389–405.

Ding, P., A. Feller, and L. Miratrix (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(3), 655–671.

Ding, P., A. Feller, and L. Miratrix (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association* 114(525), 304–317.
Fan, Y. and S. S. Park (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. Econometric Theory 26(3), 931–951.

Heckman, J. J., J. Smith, and N. Clements (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. The Review of Economic Studies 64(4), 487–535.

Hilbe, J. M. (2011). Negative binomial regression. Cambridge University Press.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. Statistical Science 17(3), 286–327.