Probabilistic Model Incorporating Auxiliary Covariates to Control FDR

Lin Qiu[1], **Nils Murrugarra-Llerena**[2] , **Vítor Silva**[3] , **Lin Lin**[4] and **Vernon M. Chinchilli**[1]

[1]**The Pennsylvania State University,** [2]**Weber State University,** [3]**Snap Inc,**

[4]**Duke University**

lin.qiu.stats@gmail.com, nmurrugarrallerena@weber.edu, vitor.silva.sousa@gmail.com,l.lin@duke.edu,vchinchi@psu.edu

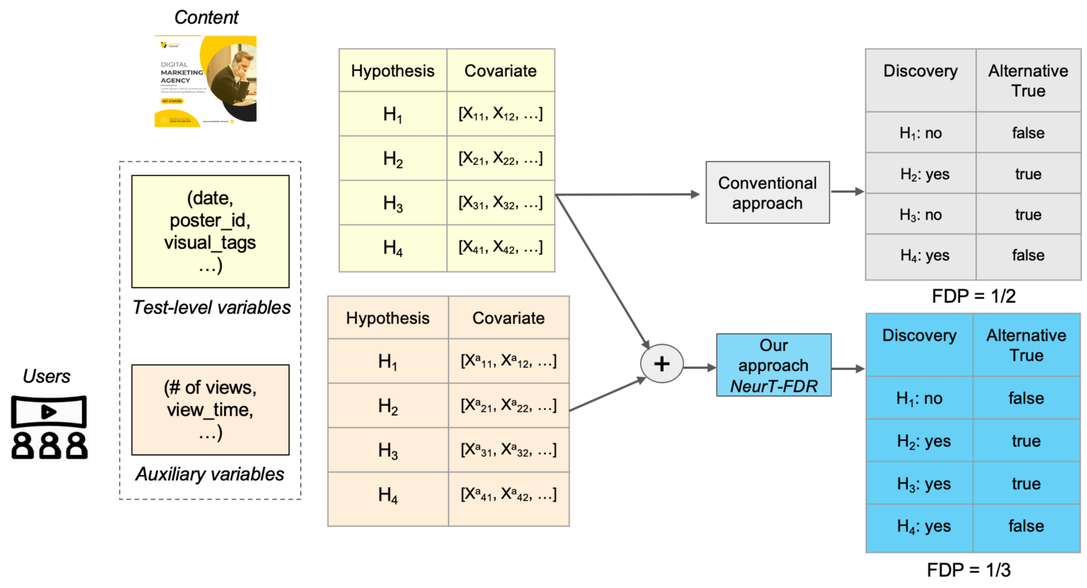
Motivation

- Multiple Hypothesis Testing (MHT) is used in different domains to discovery unique data instances.
- Existing solutions maximize the number of discoveries while controlling False Discovery Rate (FDR).
- For example, identifying most engaging snaps (social media posts).
  - **Popular**: Most engaging Snaps
  - **Normal**: Average engaging Snaps

- Lack of analysis based on complementary information, e.g., auxiliary variables.

Key idea

- MHT model that jointly learns test and auxiliary variables through a neural network.



Approach

- We optimize parameters to maximize the complete data log-likelihood.
- We employ a deep neural network to learn a β distribution that combines test-level and auxiliary covariates.
- For optimization purposes, we used Stochastic Gradient Descent (SGD) with an L2 regularizer.

$$p_\theta(z_i) = \int_0^1 (\lambda_i f_1(z_i) + (1-\lambda_i) f_0(z_i)) \times \text{Beta}(\lambda_i | \mathbf{X}_i, \mathbf{X}_i^a) d\lambda_i \, .$$

Test-level Covariates

Auxiliary Covariates

Regularizer

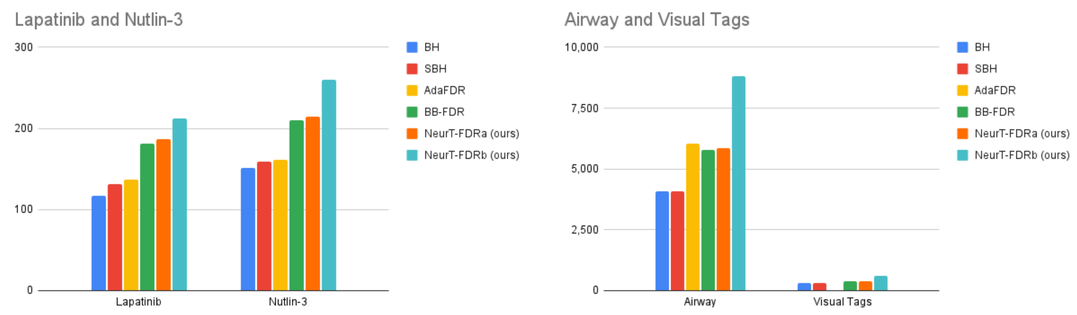
$$\underset{\theta \in \mathcal{R}^{|\theta|}}{\text{minimize}} \quad -\sum_i \log p_\theta(z_i) + \lambda_i G_{\theta_\phi}(\mathbf{X}_i)_F^2$$

Case studies and evaluation

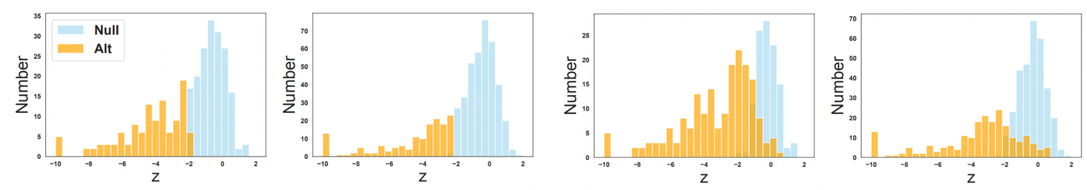
[Datasets]

- **Cancer drug screening data.** Identify how a cell line responded to a drug treatment.
- **RNA-Seq data.** Analyze RNA sequences by using log count for each gene (n=33,469) as the test-level covariate and p-value as the auxiliary covariate.
- **Snap visual tags data.** Identify top engaged contents using Snap visual tags dataset.

# of discoveries at FDR=0.1.

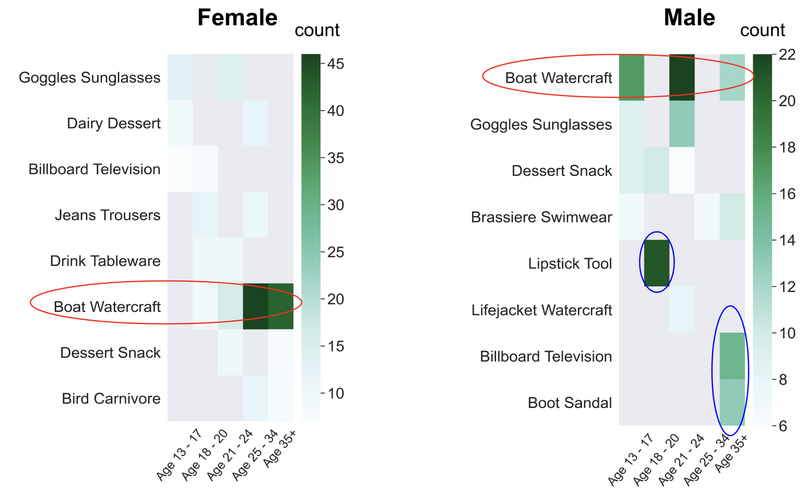


RNA-Seq: Blue and orange represents the null and alternative discoveries respectively.



(a) BH on Lapatinib (151 discoveries) (b) BH on Nutlin-3 (117 discoveries) (c) NeurT-FDRa on Lapatinib (187 discoveries) (d) NeurT-FDRa on Nutlin-3 (215 discoveries)

Snap visual tags: NeurT-FDRa discoveries.



Conclusions

- We proposed NeurT-FDR, which combines test-level and auxiliary covariates to find commonalities among these variables.
- This combination identifies more discovery under a reliable FDR of 0.1.