Bridging the Generalization Gap: Training Robust Models on Confounded Biological Data

Tzu-Yu Liu*, Ajay Kannan*, Adam Drake, Marvin Bertin, and Nathan Wan

Freenome Inc., South San Francisco, CA

INTRODUCTION

- Statistical learning on biological data can be challenging due to confounding factors both intrinsic (e.g., diurnal variation) and extrinsic (e.g., collection and processing).
- Confounding can cause models to generalize poorly and result in inaccurate performance metrics if models are not validated thoroughly.
  - Consider all "positive" samples sourced from hospital A and all "negative" samples sourced from hospital B. It is unclear whether a classifier with good cross validation performance has learned features from the confounding signal (i.e., source hospital) or the signal of interest (i.e. positive/negative class label).
- Previous techniques to reduce confounding include:
  - Stratified sampling (Wan et al., 2018), which may yield small sample sizes
  - Normalizing data using covariate labels, e.g., Hidden Covariates with Prior (Mostafavi et al. 2018) and ComBat (Johnson et al., 2007), which often requires test set covariate labels and model re-fitting at test time
  - Domain knowledge-based sequencing technical bias correction, e.g. LOWESS (Cleveland 1979; Benjamini et al., 2012), which may not apply to generic machine learning problems
- Our objective is to find a function f that transforms the observed data X into a less confounded space. Note that confounding covariates may be used to learn f, but are not arguments of f.

METHODS

- Let X: $n\times p$ be the observed data, where n is the number of observations and p is the number of features. The underlying data generation mechanism involves several factors including the disease status and possibly several confounders. Let $Y_1, Y_2, \dots, Y_{k-1}$ represent the $k-1$ confounders, where Y_i is $n\times 1$, $i \in \{1,2,\dots,k-1\}$, e.g., the age, sex, sample source institution, etc., and let Y_k: $n\times 1$ represent the phenotype of interest, e.g., the clinical disease labels.
- **Orthonormal basis construction in confounding factor normalization (ONION):** Assume that X has been centered. The objective of ONION is to rewrite X as $X = X_c + X_n$, where X_c is associated with confounders Y_i, $i \in \{1,2,\dots,k-1\}$ and X_n is the residual after factoring out covariates. This is equivalent to constructing an orthonormal basis W for $\boldsymbol{R}^p$ to project X and deconvolve the confounders.

$$W = \begin{bmatrix} | & | & \dots & | \\ w_1 & w_2 & & w_p \\ | & | & \dots & | \end{bmatrix}, WW^T = I,$$

where $w_i \in \boldsymbol{R}^p$ denotes the i_{th} basis vector.

$$X = XWW^T = X_c + X_n,$$

where $X_c = \sum_{i<k} Xw_iw_i^T$ and $X_n = \sum_{i\geq k} Xw_iw_i^T$.

$$w_i = argmax_w w^T X^T Y_i Y_i^T Xw,$$

s.t. $\|w\| = 1$ and $w^T w_j = 0$ for $j < i < k$.

Algorithm 1 presents a sequential algorithm using power iteration and deflation to satisfy the orthogonality condition. ONION peels away layers of confounders' effects sequentially, hence the acronym.

Algorithm 1: ONION

initialization $W = [\ ]$; (an empty matrix);
$X_d = X$;
for $i \leftarrow 1$ *to* $k-1$ **do**
  if $i > 1$ **then**
    $X_d = X_d - X^T w_{i-1} w_{i-1}^T$;
  Randomly initialize u_0 ;
  $\tau = 0$;
  while *stopping criterion is not satisfied* **do**
    $\tau = \tau + 1$;
    $u_\tau = X_d^T Y_i Y_i^T X_d u_{\tau-1}$;
    $u_\tau = u_\tau / ||u_\tau||$;
  end
  $W = [W; u_\tau]$; (append u_τ to W)
end
$X_n = X - XWW^T$

- **Domain-Adversarial Neural Network (DANN):** DANN is a feed-forward neural network that shares at least one hidden layer between a target prediction network and a confounder prediction network (Ganin et al. 2016) . We train $k-1$ networks $f_{Y_i}(g(X))$, $i = 1, \dots, k-1$ to predict the $k-1$ confounders and $f_{Y_k}(g(X))$ to predict the clinical label, where g is the shared feature extractor. Let L_i be the loss between predicted and actual Y_i. Training proceeds by alternating stochastic gradient descent updates to:
  - θ_{g_s} and θ_{k_s}, g's and f_{Y_k}'s parameters at step s: $\theta_{g_{s+1}} = \theta_{g_s} - \alpha \frac{dL_k}{d\theta_{g_s}}$ and $\theta_{k_{s+1}} = \theta_{k_s} - \alpha \frac{dL_k}{d\theta_{k_s}}$.
  - θ_{g_s} and θ_{i_s}, g's and f_{Y_i}'s parameters at step s, for $i = 1, \dots, k-1$: $\theta_{g_{s+1}} = \theta_{g_s} + \alpha \sum_{i=1}^{k-1} \frac{dL_i}{d\theta_{g_s}}$ and $\theta_{i_{s+1}} = \theta_{i_s} - \alpha \frac{dL_i}{d\theta_{i_s}}$.

DATASETS

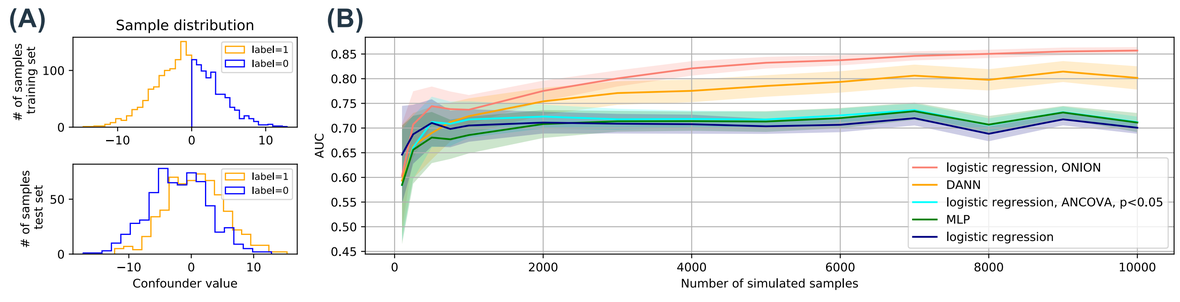
- **Confounded data simulation:** We first study confounding in a well-understood setting by simulating confounded data (**Figure 1A**).
- **Sequencing data from clinical cancer samples:** The dataset consists of whole-genome sequencing of cell-free DNA from 520 cancer patients and 214 healthy patients. Experiments are conducted with 5-fold cross validation to test confounding correction methods on data confounded (1) by biological sex and (2) by GC bias, a batch effect from the process of extracting, preparing, and sequencing cell-free DNA.

$$Z_i \sim N(0, I_d),\ i = 1, 2, ..., k$$
$$W_{x_i} :\ d \times p,\ W_{x_i}(i,j) \sim N(0,1),\ i = 1, 2, ..., k$$
$$W_{y_i} :\ d \times 1,\ W_{y_i}(i) \sim N(0,1),\ i = 1, 2, ..., k$$
$$\mathcal{E}_X \sim N(0, \sigma^2 I_p)$$
$$\mathcal{E}_{y_i} \sim N(0, \sigma^2),\ i = 1, 2, ..., k$$
$$\alpha = [\alpha_1, \alpha_2, ..., \alpha_k] \sim Dir([s_1, s_2, ..., s_k])$$
$$X = \sum_{i=1}^{k} Z_i W_{x_i} + \mathcal{E}_X$$
$$Y_i = Z_i W_{y_i} + \mathcal{E}_{Y_i},\ i = 1, 2, ..., k-1$$
$$Y_k = \mathbf{1}\{\sum_{i=1}^{k-1} \alpha_i Y_i + \alpha_k Z_k W_{y_k} + \mathcal{E}_{y_k} > 0\},$$

RESULTS

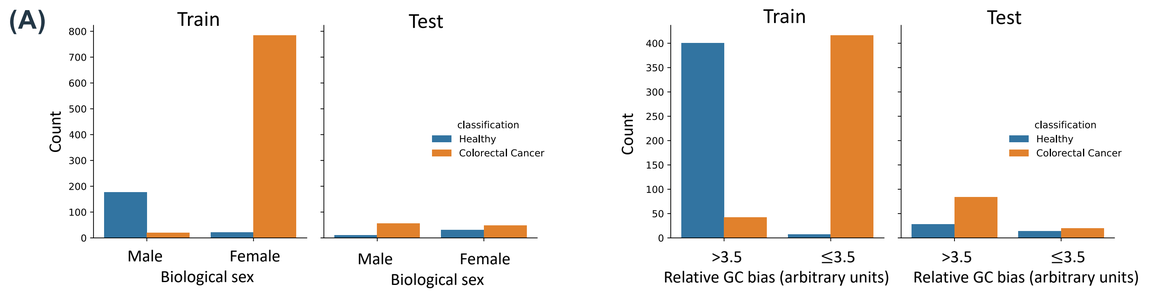
- In simulated data, the performance gaps with and without confounder correction become apparent as the sample size increases, i.e., logistic regression (logreg) vs. logreg with ONION and MLP vs. DANN. In addition, logreg with Univariate ANCOVA to filter confounded variables is marginally better than logreg (**Figure 1**).
- In empirical data, without correcting for confounders (logreg and MLP), the performance is artificially high when the test set is confounded in the same manner as the training set. However, on the entire test set, without subsampling to mimic training set confounding, logreg and MLP models perform far worse than logreg with ONION and DANN, respectively (**Figure 2**).
- In the biological sex confounding experiment, signs of the weights associated with chrX and chrY in logreg without ONION are opposite, whereas they are all centered around zero in the model with ONION (**Figure 3**).

Figure 1. Comparison of no confounder correction versus with correction on balanced test sets using simulated data. (A) Sample distribution in the training and test sets. (B) Performance as the number of samples varies.



The curves represent the mean AUC over 50 trials, and the shaded area represents +/- standard error. Number of simulated samples includes both label=1 and label=0 samples.

Figure 2. Confounding experiments using clinical cancer data. (A) Sample distribution in the training set and test sets. (B) Performance table.

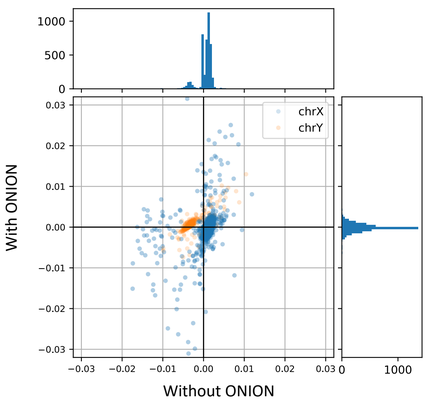


(B)

	Sex: Mean AUC (SD)		GC: Mean AUC (SD)	
Method	Entire test set	Confounded test set	Entire test set	Confounded test set
logreg	0.59 (0.04)	1.00 (0.00)	0.82 (0.08)	0.98 (0.01)
logreg, ONION	0.91 (0.01)	0.68 (0.06)	0.93 (0.02)	0.88 (0.07)
MLP	0.61 (0.06)	0.79 (0.25)	0.82 (0.11)	0.93 (0.09)
DANN	0.78 (0.07)	0.80 (0.08)	0.86 (0.05)	0.93 (0.07)
reference	0.92 (0.02)	0.95 (0.01)	0.92 (0.02)	0.86 (0.04)

AUC is the area under the receiver operating characteristic curve; SD is the standard deviation. In the sex confounding experiment, reference represents logreg trained without sex chromosomes. In the GC confounding experiment, reference represents logreg trained after LOWESS GC correction. These references serve as heuristic solutions when prior domain knowledge is given.

Figure 3. Classifier weights associated with each feature in the sex confounding experiment. Logreg without ONION relies on sex to predict cancer.



CONCLUSIONS

- If the confounding effect is not carefully corrected, one may observe inaccurate performance, with poor generalizability.
- Simulated experiments show that ONION and DANN outperform univariate ANCOVA.
- Experiments using clinical cancer data show that ONION and DANN generalize well, reducing the gap between the performance on the entire test set and a confounded, subsampled test set.
- Effective methods to mitigate the impact of confounders are subjects of ongoing research.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Lena Cheng, Riley Ennis, Signe Fransen, Girish Putcha, and David Weinberg for their extensive suggestions, feedback, and editorial support.

REFERENCES

Wan et al. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *bioRxiv*, 2018. doi: 10.1101/478065.
Mostafavi et al. Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLoS ONE*, 8(7), 2013.
Johnson et al. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.
Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.
Benjamini et al. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10): e72, 2012.
Ganin et al. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

(*: authors with equal contribution)