NeurIPS Competition Report

Amir Asiaee[1] and Kaveh Aryanpo[2]

[1]Department of Biostatistics, Vanderbilit University Medical Center, Nashville, USA
[2]Department of Informatics, Kings College, London, UK
[1]*amir.asiaeetaheri@vumc.org*
[2]*kaveh.aryan@kcl.ac.ir*

December 12, 2023

Abstract

This is the a brief project report for **Open Problems – Single-Cell Perturbations** competition which was hold as a part of Neurips 2023 content. The code is available here `https://github.com/kavaryan/sc-pertb`.

1 1000 Feet Summary

In this report, we aim to predict the impact of unseen perturbations on the transcriptome by modeling gene regulatory networks as causal systems. Understanding the cause-and-effect relationships within these networks enables us to foresee the effects of new perturbations.

1.1 Approach Based on Causal Discovery

Our methodology is rooted in the causal discovery literature, focusing on deciphering the causal relationships within a system from both observational and perturbational data. In this context, we treat the influence of drugs as a **shift** or a **soft intervention**, hypothesizing that each cell type possesses its **distinct gene regulatory network**.

1.2 Data Preparation

Each point on the provided heatmap (below) represents an **experiment** defined by a specific **drug-cell type pair**. To collate data for each experiment, we gather cells of the same type treated with the identical drug, irrespective of other variables. This necessitates the elimination of **batch effects** from cells originating from different wells but associated with the same experiment (i.e., drug-cell type pair), followed by the union of these cells into the one cell-by-gene matrix per (drug, cell) experiment.

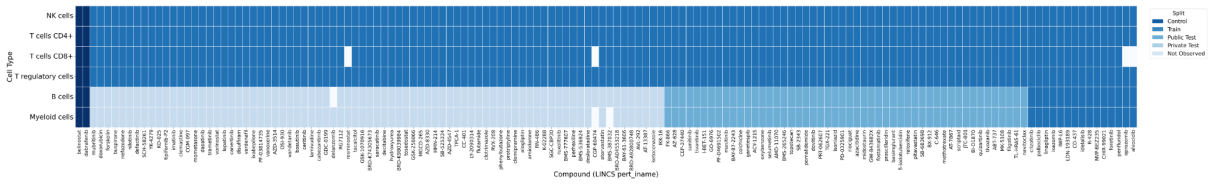

Figure 1: Each cell of this heatmap is an experiment whose data is collected from various wells and batch removed.

1.3 Structural Equation Model of Gene Regulatory Network

We postulate that every cell type has a unique gene regulatory network, which we model using a linear Gaussian structural causal model. Here, the transcriptome is defined by the structural equation $X = XB + E$, where X is an $n \times p$ matrix representing n cells and p genes, B is the gene regulatory network, and E represents noise, each row of which is sampled from $N(0, \Omega^{-1})$ with Ω being the general form of the inverse of the covariance matrix.

1.4 Drugs as Shift Interventions

Our causal model is characterized by two matrices: B and Ω. We assume that each drug **shifts the mean of the noise distribution for its direct targets**. It is crucial to note that while differential gene expression analyses (like Limma) may identify numerous genes as differentially expressed, they do not distinguish between changes caused by direct drug targets and those resulting from indirect downstream effects propagated through the gene regulatory network. To address this, we introduce a bipartite network between drugs and genes, represented by a weighted matrix A, which defines the direct targets of a drug and the extent of shift it induces. The comprehensive model of the transcriptome is thus $X = XB + DA + E$, where D is derived from the experimental setup and is an $n \times d$ matrix (d being the number of drugs, 146 in our competition), and A is the target profile of each drug.

1.5 Report Structure

Our report is structured into four main sections: batch effect correction, target learning (where we determine matrix A), structure learning (focusing on matrices B and Ω), and finally, a section detailing how to incorporate auxiliary biological information such as LINCS and ATAC-seq data into our model.

2 Batch Effect Correction

2.1 Overview

In the context of transcriptomic analysis, the observed pseudo-bulk data used for running 'limma' can be conceptualized as $\log_2(\text{pseudo-bulk-CPM}) = \text{biology} + \text{technology} + \text{noise}$. Here, the 'technology' component primarily represents the batch effect, encompassing factors such as library, plate, and donor. The 'noise' component can be dissected into two parts: one associated with biological factors and the other with batch effects. However, for simplicity, we assume the noise is entirely attributed to biological factors.

2.2 Batch Effect Correction Procedure

Our approach to correcting batch effects involves the following steps:

- Formation of Pseudo-Bulk Data: We start by aggregating the transcriptome of all cells of a given cell type in each well, forming the pseudo-bulk data.
- Application of Limma: Next, we apply 'limma' on the pseudo-bulk data, where $\log_2(\text{CPM})$ is regressed against both biological and technological factors. The biological term consists of the baseline expression of the gene (captured by the intercept of the model) and the total direct and indirect effects of the drug. The technology component represents the batch effects originating from the library, plate, and donor.
- Computing Batch Effect Corrected Biology Signal: The batch effect corrected biological signal is computed as $\text{biology} = \log_2(\text{CPM}) - \text{technology}$. This step involves adjusting the

original counts to reflect the corrected biology signal, redistributing these adjusted counts back to individual cell levels.

- Concatenation and Separation of Corrected Data: Finally, we concatenate the batch-effect-corrected cells from each experiment (characterized by the pair of cell type and drug) and store them separately for further analysis.

3 Target Learning

The objective in target learning is to discern the mechanism of action of each drug. This includes identifying the targets of each drug and quantifying the shift in expression of these targets, all of which are encapsulated in matrix A.

Ideally, with a sufficient variety of environments (cell types), we could determine drug targets by **intersecting genes that are differentially expressed in these environments compared to their unperturbed state**. However, several challenges arise:

- **Limited Environments:** With a small number of environments, there's a risk of erroneously identifying downstream genes as targets.
- **Non-Expressed Targets:** If a target gene is not expressed in a particular cell type, it complicates identification. For instance, as candidates for the target of perturbation 1, we should consider: (set of differentially expressed genes in cell type 1 compared to unperturbed cells of type 1 $\cup$ set of genes that are zero in unperturbed cells of type 1) $\cap$ (set of differentially expressed genes in cell type 2 compared to unperturbed cells of type 2 $\cup$ set of genes that are zero in unperturbed cells of type 2) $\cap \ldots$.
- **Abundant Environments:** Conversely, with too many environments, the intersection of differentially expressed genes may become empty.

To preform the differential gene expression analysis of sc-RNA seq data with two conditions (perturbed by drug i vs. unperturbed or negative control) we have multiple choices. One is to use DE analysis developed for sc-RNA seq in packages like Seurat or using bulk data method like limma. Here we use the latter. Each experiment has a cell-by-gene data matrix of batch-corrected counts and we conduct differential gene expression analysis by performing 100 random resampling of batch-corrected counts, forming 100 pseudo-bulk datasets, and then applying Limma with two terms: intercept and perturbation status (drug vs. control) variables.

Then, for each drug, we sort the significant genes (p-value ¿ 0.05) per cell type based on their inferred log fold change (the coefficient of the perturbation status term), adding all unexpressed genes in the negative control experiment of that cell type to the list, and then intersecting these lists across cell types.

Since a large number of genes are identified the intersection size can be unreasonably large. Therefore, we dynamically increase the length of the list (sorted decreasingly from the largest absolute log fold change) and compute intersections until reaching a set of 10 genes. We then generate step plots, selecting genes as targets if their corresponding step length is large, indicating persistence as part of the intersection across multiple cell types. A step is considered long if its length is more than twice the mean of all steps, based on the assumption of a Poisson distribution for step length, where twice the mean is one standard deviation away from the mean.

Finally, we select these genes as targets and their coefficients from Limma and form matrix A.

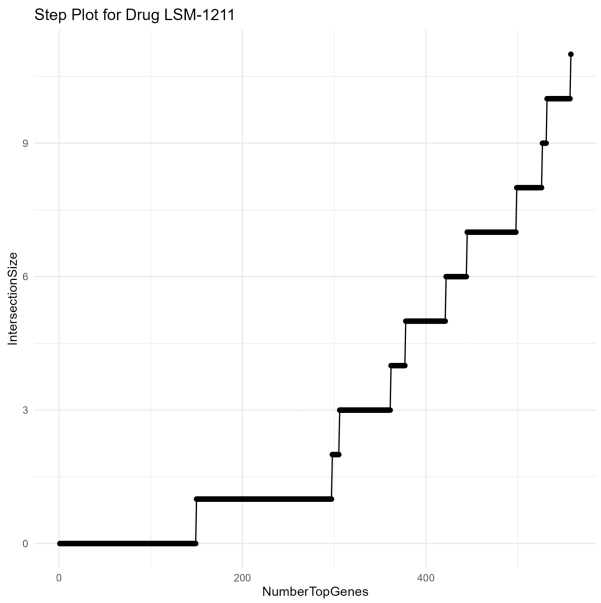

Figure 2: Step plot to determine the best set of genes for target learning.

4 Causal Structure Learning

In the realm of structure learning algorithms, there are two primary categories: score-based and constraint-based methods. Constraint-based methods rely on conditional independence tests to deduce a set of inferable conditional independencies from the data, subsequently utilizing these to infer the underlying Directed Acyclic Graph (DAG). Score-based methods, on the other hand, navigate the landscape of all structural causal models, assigning scores (typically using maximum likelihood) and seeking the graph with the highest score. A critical consideration for score-based methods is that the underlying graph should be a DAG or a stable cyclic structure. The advent of trace exponential equality constraint as a differentiable constraint, which encapsulates "DAGness," has enabled continuous optimization in score-based search. Our approach employs a score-based method with a constraint on B, innovatively allowing the graph to be cyclic and utilizing shift intervention data alongside observational data to learn B. Additionally, our method proposes a scalable strategy for learning the noise correlation structure Ω.

Our data generation model is defined as:

$$X = XB + DA + E, \quad X, E \in \mathbb{R}^{n\times p}, B \in \mathbb{R}^{p\times p}, D \in \mathbb{R}^{n\times d}, A \in \mathbb{R}^{d\times p}, \forall i \in [n] : E_{i:} \sim N(0, \Omega^{-1})$$

For causal structure learning of the gene regulatory network, our aim is to learn (B, Ω), under certain constraints. **The fundamental mathematical constraint for Ω is that it should be symmetric positive semidefinite.** Moreover, in causal discovery literature, Ω is often simplified to an identity matrix (indicating no unmeasured confounder between genes) or assumed to be diagonal. Addressing non-diagonal Ω (a more realistic assumption in GRNs) is computationally challenging unless gene selection, dimensionality reduction, or transcriptome embedding is applied.

Thus, the learning problem is reduced to learning (B, Ω, A) where B should be either a DAG or stable cyclic. **The DAGness condition on** B can be enforced by the trace exponential equality constraint:

$$h(B) \triangleq \operatorname{tr}\left(e^{B\odot B}\right) - d = 0$$

The condition for being stable cyclic is that the maximum absolute eigenvalue of B should

be less than one:

$$g(B) \triangleq \max_i |\lambda_i(B)| < 1$$

Therefore, our score-based method simplifies to the following optimization problem:

$$\arg\min_B \|(Y - XB)\Omega^{1/2}\|_F^2, \quad Y = X - DA, B \in \Theta$$

Here, the constraint set Θ is either the set of DAGs determined by $h(B) = 0$ or stable cyclic graphs where $g(B) < 1$.

4.1 Feature Selection

In our approach to simplifying the structure learning process, we initially assume that $\Omega = I$. Under this assumption, the optimization problem is effectively reduced to solving 21,000 ordinary least square (OLS) problems, subject to constraints over B. However, tackling OLS for $p = 21,000$ genes demands extensive computational resources and a large number of samples. The total number of samples, divided by the number of cell types, amounts to approximately 40,000 cells. This number drops significantly for the two test cell types (Myeloid and B cell), leading to a singular system. Consequently, we need to regularize the columns of B using l_1-norm regularization, essentially turning the problem into solving 21,000 LASSO problems.

Even with this regularization, the computational load remains intensive. To alleviate this burden, our strategy involves selecting only highly variable genes. We adopt a method similar to that used by Seurat, focusing on genes that exhibit high dispersion across all samples of the cell type under investigation. Dispersion here is defined as the variance of a gene divided by its mean. However, due to the presence of outliers, we smooth the variance estimates using the principle that variance in transcriptomic data is a smooth function of the mean. This smoothing is accomplished by fitting a loess curve to the (mean, variance) data of all genes. Then, using the predicted (smoothed) variance, we compute the dispersion and select only the top 2,000 highly dispersed genes for further analysis.

4.2 Stable Cyclic Condition

In the context of gene regulatory networks, we posit that imposing a Directed Acyclic Graph (DAG) constraint may be an unrealistic assumption, given the numerous known cycles that regulate genes and their pathways. Consequently, we adopt the stable cyclic condition: $g(B) \triangleq \max_i |\lambda_i(B)| < 1$. It's important to note that $g(B)$ represents a non-convex condition.

To address this, we propose two potential solutions:

- Convex Relaxation of $g(B)$: One approach is to relax $g(B)$ with a convex constraint, specifically $g(B) \leq \|B\|_2 \leq 1$. This would allow us to solve the problem as a convex-constrained LASSO.

- Projection of LASSO Solution: Alternatively, we can project the LASSO solution onto the non-convex set defined by $g(B) < 1$. This projection involves scaling B by $1/g(B)$, effectively shrinking all eigenvalues of B to be less than 1. To calculate $g(B) = \max_i |\lambda_i(B)|)$, we employ the "power iteration" method, which progressively approximates the maximum absolute eigenvalue with each iteration. In practical applications, approximately five iterations have proven sufficient. By opting for the projection approach, we maintain the integrity of the stable cyclic condition while managing the computational complexity associated with the non-convexity of $g(B)$. This method ensures that our model adheres to the biological reality of gene regulatory networks, encompassing both acyclic and cyclic gene interactions.

4.3 Moving Beyond Uncorrelated Noise

The presence of Ω in the objective function $\arg\min_B \|(Y - XB)\Omega^{1/2}\|_F^2$ links the LASSOs together, preventing parallelization. Moreover, Ω requires estimation. Generally, the MLE depends on both (B, Ω) and is bi-convex; it is convex with respect to one when the other is fixed. This suggests an iterative optimization strategy, beginning with $\Omega_0 = I$ to find B_1, and then incorporating it into the logdet formula to find Ω. The iterative procedure is as follows:

$$\hat{B}_t \in \arg\min_{B:h(B)=0 \text{ or } g(B)<1} \left\|(X - XB^T)\Omega_{(t-1)}^{-1/2}\right\|_F^2 + \lambda_b\|B\|_1,$$

$$\hat{\Omega}_t \in \arg\min_{\Omega\geq 0} -\log|\Omega| + \text{tr}(\hat{\Sigma}_t\Omega) + \lambda_\omega\|\Omega\|_1,$$

where $\Omega_0 = I$ is the initial value, Σ_t is the empirical covariance matrix of noises, $\text{tr}(\hat{\Sigma}_t)$ is the trace part of $\left\|(X - XB^T)\Omega_{(t-1)}^{1/2}\right\|_F^2$ that depends on Ω, and $\|\Omega\|_1 = \sum_{i,j}|\Omega_{ij}|$ is the matrix l_1-norm. In the case of fewer samples than variables ($n < p$), the regularization parameter $\lambda_\omega > 0$ ensures Ω is invertible. We can estimate $\hat{B}_t$ via the previously discussed continuous optimization and $\hat{\Omega}_t$ using methods for high-dimensional precision matrix estimation.

Addressing the computational complexity of estimating the full inverse covariance matrix Ω, we assume Ω is block diagonal, reflecting a complexity level between an arbitrary inverse covariance and a diagonal matrix. This assumption is based on the idea that genes operate in clusters with correlated noise internally, while noise correlations between clusters are negligible.

To apply this concept, we cluster genes for each cell type in the unperturbed state, focusing on genes within the same cluster to learn each block of Ω. Consequently, the logdet objective decomposes into smaller subproblems, each using genes from a single cluster to estimate the corresponding block of Ω. This block-wise approach allows for parallelization and efficient resolution using libraries such as gelnet, an implementation of Graphical Lasso suitable for large-scale problems.

5 Incorporating Auxiliary Biological Information

5.1 Biological Data: LINCS

Given that all drugs under study are sourced from the LINCS dataset, where they have been tested across multiple cell lines, we leverage information from the differentially expressed genes within these cell lines to enhance the **drug target learning procedure**. Although the data from LINCS are not derived from single-cell measurements, the diversity of environments (cell types and cell lines) enriches our capacity to accurately identify direct drug targets and the genes they influence. We utilize level 5 LINCS data to facilitate this augmentation.

5.2 Biological Data: ATAC-seq

The ATAC-seq data informs us of chromatin accessibility, indicating regions active in transcription and those that are not. Each ATAC-seq measurement comprises a peak value and its specific location on the human genome assembly GRCh38. We utilize this data to construct a prior mapping of transcription factors (TFs) in the open chromatin regions to other genes nearby. This ATAC-induced TF/other-genes mapping is integrated into our model as a prior.

The mapping process can be conducted via two methods. One approach involves analyzing the peak-address in the GRCh38 sequence to identify nearby base sequences and then searching for motifs associated with TFs. If a gene g is proximal to the peak-address, we map the TF to g. Alternatively, we can first create an (address, TF) dataset from the GRCh38 and motif data and then proceed similarly to associate TFs with nearby genes. For our purposes, we adopted

the HOMER dataset and specifically the "Human hg38 UCSC BigBed Track" BED file from `http://homer.ucsd.edu/homer/`, which provides a precomputed set of TF-gene associations.