

STATISTICAL METHODS FOR SINGLE CELL GENE EXPRESSION: DIFFERENTIAL
EXPRESSION, CURVE ESTIMATION AND GRAPHICAL MODELLING

By

Satabdi Saha

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2022

ABSTRACT

STATISTICAL METHODS FOR SINGLE CELL GENE EXPRESSION: DIFFERENTIAL EXPRESSION, CURVE ESTIMATION AND GRAPHICAL MODELLING

By

Satabdi Saha

This dissertation elucidates a set of statistical methods developed for analysis of single cell gene expression datasets. Gene expression profiling of single cells has led to unprecedented progress in understanding normal physiology, disease progression and developmental processes. However, despite many improvements in high throughput sequencing, various technical factors including cell-cycle heterogeneity, library size differences, amplification bias, and low RNA capture per cell lead to high noise in scRNA-seq experiments. A primary characteristic of these datasets is the presence of high number of zeroes which represents the undetectable level of expression for a transcript. Statistical methods capable of modelling novel single cell experiments are developed and new estimation strategies are proposed and validated using simulated and real data experiments.

- In Chapter 1, the motivation and underlying philosophies of single cell gene expression is reviewed. Methods for analysis of dose response experiments and gene co-expression networks are reviewed and novel statistical hypothesis to be investigated using single cell experiments are discussed.
- In Chapter 2, I analyze a unique in vivo dose response hepatic scRNAseq dataset consisting of 9 dose groups with 3 biological replicates for 11 distinct liver cell types for greater than 100K cells. A Hurdle model for multiple group data is proposed, which models the bimodality of single cell gene expression within multiple groups. Based on the model assumptions, I derive a fit for purpose Bayesian test for simultaneously testing the differences in mean gene expression and zero proportions for multiple dose groups. For comparison the counterpart likelihood-ratio test for differential expres-

sion that incorporates testing for both components is also derived. This chapter was originally published in [1].

- In Chapter 3, dose response curve estimation for single cell experiments is studied. Current protocols for genomic dose response modelling are only capable of modelling bulk and microarray datasets. A semiparametric regression model for joint dose response curve estimation for multiple cell-types while accounting for confounding covariates is proposed. A novel, scalable and efficient optimization algorithm using the MM philosophy is proposed for the estimation of both monotone and non-monotone curves. Two relevant tests of hypothesis are discussed and the proposed methods are validated using several simulated datasets.
- In Chapter 4, co-expression network estimation is studied using graph signal processing. A kernelized signed graph learning approach is developed for learning single cell gene co-expression networks, based on the assumption of smoothness of gene expressions over activating edges. Performance is assessed using real human and mouse embryonic datasets. This chapter was originally published in [2].

This work is dedicated to my husband and son. You have made me better, stronger, and more fulfilled than I could have ever imagined.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisors, Dr Tapabrata Maiti and Dr Samiran Sinha, both of whom have been extremely supportive of my study, research, and professional development. It has been inspiring to observe their scientific rigor and enthusiasm for interdisciplinary research. I am grateful for their constant encouragement, patience, guidance, and the tremendous support throughout my doctoral studies at MSU. I would like to thank Dr Sudin Bhattacharya, Dr Rance Nault and Dr Tim Zacharewski for introducing me to the field of toxicogenomics; it has always been a wonderful experience collaborating with them on highly interdisciplinary problems.

I would like to thank my committee members, Dr Shlomo Levental and Dr Lyudamila Sakhanenko for serving in my committee and being patient readers of my work. I am highly indebted to Dr Sakhanenko and Dr Levental for being wonderful teachers; their contribution towards my PhD training is immense. In addition, I am very grateful to Dr Selin Aviyente and her doctoral advisee Abdullah Karaslaanli for collaborating with me on developing signal processing based graph learning ideas, which contribute to the fourth chapter of this dissertation.

I thank Prof. Elijah Dikong and Prof. Camille Fairbourne for their advice on my teaching assistantships and for being wonderful mentors. Special thanks to the wonderful staff of our department, Sue, Teresa, Megan, Andy and Tami for their extensive support throughout the entire duration of my doctoral studies.

I sincerely thank my friends at MSU, Nilanjan, Sanket, Alex, Abhijnan, Sumegha and Phuong for their generous help and friendship. I also thank my brother and friend Subhrajit for being a constant source of positivity and encouragement. Finally and most importantly I wish to thank my mom and my husband. This dissertation would not have been possible without their endless support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	x
CHAPTER 1 INTRODUCTION	1
1.1 Motivation for exploring single cell gene expression	1
1.2 Structure of single cell gene expression datasets	2
1.3 Dose Response Experiments	3
1.3.1 National Toxicology Program's approach to genomic dose response modeling	4
1.3.1.1 Determining Adequate Signal in the Data	5
1.3.1.2 Filtering of Measured Features	6
1.3.1.3 Dose Response Curve Estimation	7
1.4 Covariation Analysis	10
CHAPTER 2 BAYESIAN SINGLE CELL RNASEQ DIFFERENTIAL GENE EXPRESSION TEST FOR DOSE RESPONSE STUDY DESIGNS	12
2.1 Single Cell Dose Response Experiments	12
2.2 Materials and Methods	14
2.2.1 Animal handling and treatment	14
2.2.2 Real scRNAseq and snRNAseq datasets	14
2.2.3 Dose-response data simulation	15
2.2.4 Single cell RNA-seq Hurdle Model	16
2.2.4.1 Hypothesis Formulation	17
2.2.4.2 Single cell Bayesian Hurdle model Analysis (scBT)	17
2.2.4.3 Multiple group Likelihood Ratio Test (LRT)	21
2.2.4.4 Linear model-based Likelihood Ratio Test (LRT linear)	22
2.2.5 Benchmarking method selection	22
2.2.5.1 Seurat Bimod	23
2.2.5.2 MAST	23
2.2.5.3 Limma-trend	25
2.2.5.4 Wilcoxon Rank Sum (WRS) Test	25
2.2.5.5 ANOVA	26
2.2.5.6 Kruskal-Wallis (KW) Test	26
2.2.6 Benchmarking and sensitivity analyses	26
2.3 Results	27
2.3.1 Performance accuracy of DE test methods	29
2.3.2 Type I error control and power	32
2.3.3 Parameter Sensitivity Analysis	33
2.3.4 Test method agreement	35

2.4	Real dose-response dataset DE analysis	36
2.5	Discussion	38
2.6	Conclusion	41
2.7	Acknowledgements	41
CHAPTER 3 SEMIPARAMETRIC DOSE RESPONSE CURVE		
ESTIMATION FOR SINGLE CELL DOSE RESPONSE		
EXPERIMENTS 42		
3.1	Single Cell Dose Response Experiments	42
3.1.1	Motivating experimental study and hypothesis of interest	44
3.1.2	Literature on dose response curve estimation	45
3.1.3	Literature on MM algorithms	47
3.2	Methods	49
3.2.1	Model, notations and assumptions	49
3.2.2	Penalized Estimation	55
3.2.3	Monotonicity constraints	56
3.2.4	Confidence interval estimation	57
3.2.5	Model Selection	58
3.3	Hypothesis Testing	58
3.4	Results	59
3.4.1	Simulation Design	59
3.4.2	Analysis	60
3.5	Discussion	72
CHAPTER 4 KERNELIZED SIGNED GRAPH LEARNING FOR		
SINGLE CELL GENE REGULATORY NETWORK		
INFERENCE 74		
4.1	Single Cell Gene Regulatory Networks	74
4.2	Background	78
4.2.1	Notations	78
4.2.2	Low- and High-frequency Signals on Unsigned Graphs	79
4.2.3	Unsigned Graph Learning	79
4.2.4	Kernels	80
4.3	Methods	81
4.3.1	Signed Graph Learning	81
4.3.2	Kernelized Signed Graph Learning	82
4.3.3	Hyperparameter Selection	84
4.3.4	Generation of simulated datasets from zero-inflated negative binomial distribution	85
4.3.5	Performance Metrics	86
4.3.5.1	AUPRC and AUROC:	86
4.3.5.2	AUPRC Activating/Inhibitory:	86
4.3.5.3	EPR	87
4.4	Results	87
4.4.1	Synthetic Datasets	88

4.4.2	Real Datasets	93
4.5	Discussion	96
4.6	Acknowledgements	99
APPENDICES		100
APPENDIX A	BAYESIAN SINGLE CELL RNASEQ DIFFERENTIAL GENE EXPRESSION TEST FOR DOSE RESPONSE STUDY DESIGNS	101
APPENDIX B	SEMIPARAMETRIC DOSE RESPONSE CURVE ESTIMATION FOR SINGLE CELL DOSE RESPONSE EXPERIMENTS	104
APPENDIX C	KERNELIZED SIGNED GRAPH LEARNING FOR SINGLE CELL GENE REGULATORY NETWORK INFERENCE	138
BIBLIOGRAPHY		147

LIST OF TABLES

Table 1.1	Parametric models for genomic dose response analysis	8
Table 2.1	Dose-response models for simulation of scRNAseq data	16
Table 3.1	Parameter estimate, asymptotic standard error, bias and mean squared error(MSE)of parameter β under full model and the intercept model. The results are averaged over 500 replicates and reported for $n_i = 100, 300$	63
Table 3.2	Parameter estimate, asymptotic standard error, bias and mean squared error(MSE)of parameter ϕ under full model and the intercept model. The results are averaged over 500 replicates and reported for $n_i = 100, 300$	63
Table 3.3	Parameter estimate, asymptotic standard error, bias and mean squared error(MSE)of parameter ψ_0 under full model and the intercept model. The results are averaged over 500 replicates and reported for $n_i = 100, 300$	63
Table 3.4	Parameter estimate, asymptotic standard error, bias and mean squared error(MSE)of parameter ψ_1 under full model and the intercept model. The results are averaged over 500 replicates and reported for $n_i = 100, 300$	72

LIST OF FIGURES

- Figure 2.1 Flow diagram of the simulation, benchmarking, and experimental data evaluation strategy presented in the manuscript. Briefly, SplattDR was developed to simulate dose-response scRNAseq data and validated based on experimental dose-response data. Simulated datasets were generated varying diverse parameters 10 times and then used to assess the performance of each test method. Each test method was also assessed using experimental data from the hepatic snRNAseq dose response dataset obtained from male mice gavaged every 4 days for 28 days with 0.01, 0.03, 0.1, 0.3, 1, 3, 10, or 30 $\mu\text{g}/\text{kg}$ TCDD. Related figures for each analysis from the main body are noted. 24
- Figure 2.2 Comparison of simulated and real dose-response data. **(A)** Relationship between gene-wise mean expression and percent zeroes for simulated and real dose-response data. Simulation data consisted of 10,000 genes and 9 dose groups based on parameters derived from experimental dose-response snRNAseq data. Black line represents a fitted model to the experimental data from which the normalized root mean square deviation (NRMSD) of simulated data was determined. **(B)** Relationship between gene-wise mean expression and variance for simulated and experimental data. NMRSD was calculated for simulated data from the fitted model represented as a black line. **(C)** Distribution of log(fold-changes) in experimental and simulated data showing the median and minimum and maximum values. **(D)** Principal components analysis of simulated data colored according to simulated dose groups. **(E)** NMRSD estimated relative to fitted model in A,B for simulated data generated from initial parameters derived from published hepatic scRNAseq (two dose; GSE148339), hepatic whole cell (whole cell; GSE129516), and peripheral blood mononuclear cell (PBMC; GSE108313) datasets. **(F)** NMRSD estimated relative to model fitted to cell-type specific experimental dose-response data when simulated from initial parameters estimated from that same cell type. Box and whisker plots show median NMRSD, 25th and 75th percentiles, and minimum and maximum values. 29

Figure 2.3 Classification performed of DE analysis tests. **(A)** ROCs estimated from simulated dose-response scRNAseq data for 9 DE test methods including all genes expressed in at least 1 cell (unfiltered). **(B)** ROCs for 9 DE test methods after filtering simulated dose-response scRNAseq data for genes expressed in only $\geq 5\%$ of cells (low levels) in at least one dose group. **(C)** Precision-recall curves (PRCs) for 9 DE test methods on unfiltered simulated dose-response scRNAseq data. **(D)** PRCs for 9 DE test methods on filtered simulated dose-response scRNAseq data. Lines represent the mean values and shaded region reflects the standard deviation for 10 independent simulations. **(E)** Precision of DE test methods. **(F)** FPR of DE test methods. **(G)** MCC for test methods. E,F,G Box and whisker plots median values, 25th and 75th percentiles, and minimum and maximum values for 10 independent simulations. Points reflects values for each independent simulation. Panels display comparisons of unfiltered and filtered datasets. 30

Figure 2.4 Evaluation of Type I and II error control. **(A)** False positive rate (FPR) of 9 differential expression test methods estimated from negative control (0% DE genes) simulated dose-response scRNAseq data including all genes expressed in at least 1 cell (unfiltered) and genes expressed in only $\geq 5\%$ of cells in at least one dose group (filtered). **(B,C)** Logistic regression models were fitted to negative control data to predict the probability of false positive identification using percent zeroes and mean expression as covariates. Lines represent the predicted probability of false positive classification with the shaded region representing the 95% confidence interval. **(D)** False negative rate (FNR) of 9 differential expression test methods estimated from positive control (100% DE genes) simulated dose-response scRNAseq data including unfiltered and filtered datasets. (E,F) Logistic regression models were fit to positive control data. Lines represent predicted probability of false negative classification with shaded region representing the 95% confidence interval. 32

Figure 2.5 Matthews correlation coefficient (MCC) from sensitivity analyses of differential expression test methods. **(A)** MCC for 9 DGEA test methods determined from simulated dose response data with varying number of cells per dose group. Simulations consisted of 5,000 genes with a probability of differential expression of 10% and 9 dose groups. **(B)** MCC for simulated data varying the cells numbers by dose group. The number of cells in each of the 9 doses groups is shown on the right. **(C)** MCC for varying proportion of differentially expressed genes. **(D)** MCC when varying the mean fold-change (location) of repressed differentially expressed genes. **(E)** MCC for varying distribution of fold-change (scale) of differentially expressed genes. **(F)** MCC for varying dropout rates calculated as in Table S3. Points represent median and error bars represent minimum to maximum values. Boxplots represent median, 25th to 75th percentile, and minimum to maximum values. Each analysis consisted of 10 replicate datasets including all genes expressed in at least 1 cell (unfiltered) and genes expressed in $\geq 5\%$ of cells in at least one dose group (filtered). 34

Figure 2.6 Agreement of differential expression test methods on experimental dose-response data. **(A)** Upset plot showing the intersection size of genes identified as differentially expressed by 9 different test methods in hepatocytes from the portal region of the liver lobule. **(B)** Intersect of differentially expressed genes in portal fibroblasts. **(C)** Intersect size in hepatic stellate cells. Vertical bars represent the intersect size for test methods denoted by a black dot. Horizontal bars show the total number of differentially expressed genes identified within each test (set sizes). Only intersects for which genes were identified are shown. Genes were considered differentially expressed when (i) expressed in $> 5\%$ of cells within any given dose group and (ii) exhibit a $|\text{fold-change}| \geq 1.5$. A heatmap in the upper left corner of each panel shows the pairwise AUCC comparisons for the 500 lowest p-values. **(D)** Relative proportion of cell types identified in each dose group of the real dataset for the cell types in A,B,C. Experimental snRNAseq data was obtained from male mice gavaged with sesame oil vehicle (vehicle control) or 0.01 – 30 $\mu\text{g/kg}$ TCDD every 4 days for 28 days. **(E)** Graph metrics for gene set enrichment analysis of portal fibroblasts grouped by similarity in gene membership. Violin plots show distribution of node-wise values for each test method. **(F)** Network visualization of significantly enriched (adjusted p-value ≤ 0.05) gene sets using the Bayes factor ranked genes of portal fibroblasts. Groups of ≥ 2 nodes were manually annotated following commonality in the gene set names. Each node represents a gene set with the size of the node representing the number of genes in a gene set, and edges connect nodes with $\geq 50\%$ overlap. 37

Figure 3.1	Results of the simulation study, illustrating the performance of Model 1 (see Simulation Design) in 500 replicates with 300 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.	64
Figure 3.2	Results of the simulation study, illustrating the performance of Model 1 (see Simulation Design) in 500 replicates with 100 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.	64
Figure 3.3	Results of the simulation study, illustrating the performance of Model 1 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to the three different cell-types. Continuous cyan, peach, darkblue, darkgreen and red lines represent the represent the true curve and the estimated log-mean curves averaged across 500 random replicates for ZINB-SPL, ZINB-GAM-Dose, ZINB-GAM-Int and NB-GAM models respectively	65
Figure 3.4	Results of the simulation study, illustrating the estimated RMSE of Model 1 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to RMSE boxplots of the NB-GAM, ZINB-GAM-Dose, ZINB-GAM-Int and ZINB-SPL models respectively for the three different cell-types plotted over 500 replicates.	65
Figure 3.5	Results of the simulation study, illustrating the performance of Model 2 (see Simulation Design) in 500 replicates for 300 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.	66
Figure 3.6	Results of the simulation study, illustrating the performance of Model 2 (see Simulation Design) in 500 replicates for 100 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.	66

Figure 3.7	Results of the simulation study, illustrating the performance of Model 2 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to the three different cell-types. Continuous cyan, peach, darkblue, darkgreen and red lines represent the represent the true simulated curves and the estimated log-mean curve averaged across 500 random replicates for ZINB-SPL, ZINB-GAM-Dose, ZINB-GAM-Int and NB-GAM models respectively	67
Figure 3.8	Results of the simulation study, illustrating the estimated RMSE of Model 2 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to RMSE boxplots of the NB-GAM, ZINB-GAM-Dose, ZINB-GAM-Int and ZINB-SPL models respectively for the three different cell-types plotted over 500 replicates.	67
Figure 3.9	Results of the simulation study, illustrating the performance of Model 3 (see Simulation Design) in 500 replicates for 300 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.	68
Figure 3.10	Results of the simulation study, illustrating the performance of Model 3 (see Simulation Design) in 500 replicates for 100 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.	68
Figure 3.11	Results of the simulation study, illustrating the performance of Model 3(see Simulation Design) in 500 replicates for sample size 300. The columns correspond to the three different cell-types. Continuous cyan, peach, darkblue, darkgreen and red lines represent the represent the true simulated curves and the estimated log-mean curve averaged across 500 random replicates for ZINB-SPL, ZINB-GAM-Dose, ZINB-GAM-Int and NB-GAM models respectively	69
Figure 3.12	Results of the simulation study, illustrating the estimated RMSE of Model 3 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to RMSE boxplots of the NB-GAM, ZINB-GAM-Dose, ZINB-GAM-Int and ZINB-SPL models respectively for the three different cell-types plotted over 500 replicates.	69

Figure 3.13	Results of the simulation study, illustrating the performance of Model 4 (see Simulation Design) in 500 replicates for 100 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals.	70
Figure 3.14	Results of the simulation study, illustrating the performance of Model 4 (see Simulation Design) in 500 replicates for 300 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals.	70
Figure 3.15	Results of the simulation study, illustrating the performance of Model 4 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to the three different cell-types. Continuous cyan, peach, darkblue, darkgreen and red lines represent the represent the true simulated curves and the estimated log-mean curve averaged across 500 random replicates for ZINB-SPL, ZINB-GAM-Dose, ZINB-GAM-Int and NB-GAM models respectively	71
Figure 3.16	Results of the simulation study, illustrating the estimated RMSE of Model 4 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to RMSE boxplots of the NB-GAM, ZINB-GAM-Dose, ZINB-GAM-Int and ZINB-SPL models respectively for the three different cell-types plotted over 500 replicates.	71
Figure 4.1	Euclidean distances (left, normalized to $[0, 1]$) and correlations (right) between expressions of gene pairs in curated datasets studied in Section 4.4. Values are calculated only for gene pairs that are connected in the ground truth GRNs and they are reported separately for activating and inhibitory edges. Only inhibitory edges are reported for VSC, since its GRN includes only inhibitory edges.	77
Figure 4.2	Performance of scSGL and state-of-the-art methods on curated datasets as measured by AUPRC for activating and inhibitory edges. x-axis indicates dropout ratio in the dataset.	88
Figure 4.3	Performance of various methods for synthetic datasets with varying number of genes (top row), dropout ratio (middle row) and number of cells (bottom row).	89

Figure 4.4	Scalability analysis of different methods. Run time of benchmarking methods are calculated using BEELINE pipeline [3]. Run time of scSGL includes kernel construction and optimization procedure. All methods are run on the same computer. Results of GENIE3 for 2000 genes are not reported due to its high run time.	92
Figure 4.5	Performance of methods for two real-world scRNAseq datasets. Inferred graphs are compared to three different gene regulatory databases.	94
Figure 4.6	The subnetworks of 24 lineage specific genes in hESC (A) and 19 well known marker genes in mESC (B). We report results of scSGL- <i>r</i> as it has the highest AUPRC ratio in Figure 4.5. For clarity, only those edges whose absolute edge weight fall into the top 1 percentile are shown. Node sizes are proportional to their degrees.	95
Figure 4.7	UpSet plot that shows intersection between the top 1000 edges by scSGL with 3 kernels and benchmarking methods in hESC and mESC datasets.	97
Figure A.1	Principal components analysis of cell types identified in a real hepatic dose-response snRNAseq dataset. Points represent a distinct cell and colors reflect the dose group.	101
Figure A.2	Comparison of fold-change distribution in simulated and real dose-response snRNAseq data where the log-normal mean (facLoc) and standard deviation (facScale) were varied as well as the percentage of differentially expressed genes and proportion of downregulated DE genes. A total of 5000 genes were simulated or sampled from real data and the fold-change for the highest dose group was calculated. The Kullback-Leibler Divergence (KLD) intrinsic discrepancy (ID) was used to evaluate the similarity in distributions.	102
Figure A.3	Benchmarking scores for data simulated using the Splatter [4] wrapper Splattdr with default initial parameters. A total of 4500 cells and 5000 genes were simulated across 9 dose groups with a probability of being differentially expressed of 10%, 50% of which were downregulated. (A) Ground truth was used to estimated the False Positive Rate (FPR), True Positive Rate (TPR), False Negative Rate (FNR), True Negative Rate (TNR), precision, balanced accuracy, and F1 score. Boxplots and whiskers represent values for 10 replicate simulations. (B) The area-under the concordance curve (AUCC) was calculated as previously described (ref) for the 100 most significant genes ($K = 100$). Heatmap represents the pairwise AUCC for each DE analysis grouped by similarity.	103

Figure C.1	AUROC values of various methods for synthetic datasets with three different topologies (random, modular and hub) and varying number of genes (top row), dropout ratio (middle row) and number of cells (bottom row).	141
Figure C.2	EPR values of various methods for synthetic datasets with three different topologies (random, modular and hub) and varying number of genes (top row), dropout ratio (middle row) and number of cells (bottom row).	142
Figure C.3	AUROC and EPR ratios of methods for two real-world scRNAseq datasets. Inferred graphs are compared to three different gene regulatory databases.	143
Figure C.4	Performance of scSGL and state-of-the-art methods on curated datasets as measured by AUPRC ratios for activating and inhibitory edges. Each column corresponds to a synthetic network. Abbreviations: LI, linear; CY, cycle; LL, linear long; BF, bifurcating; BFC, bifurcating converging and TF, trifurcating.	144
Figure C.5	Edges detected using scSGL- <i>r</i> between 24 Lineage marker genes of hESC at different time points of the differentiation process. Only edges whose absolute edge weights fall into top 10 percent are shown. Edge thicknesses are proportional to their weights, and node sizes are proportional to their degrees.	145

CHAPTER 1

INTRODUCTION

1.1 Motivation for exploring single cell gene expression

Cells are the fundamental unit of life: Millions of cells in our body coordinate to perform basic physiological functions essential for maintaining life. Yet, several critical biological processes such as the genesis, development and fate commitment of early-stage embryos, are determined by the biology of a single cell. The ability to investigate single-cell states have allowed exploration of critical biological systems with unprecedented resolution. It is now possible to easily profile individual cells instead of cell populations, which advances our fundamental understanding of the intrinsic cellular heterogeneity and dynamics [5, 6, 7].

However, despite many improvements in high throughput sequencing, various technical factors including cell-cycle heterogeneity, library size differences, amplification bias, and low RNA capture per cell lead to high noise in single-cell RNAseq (scRNAseq) experiments. Recent technologies are capable of sequencing millions of cells, but often generate highly sparse expression datasets due to shallow sequencing. The presence of these issues result in substantial noise that often obscures the true biological signal and renders the application of traditional statistical models for analysis of single cell datasets unsuitable [8, 9].

The availability of gene expression data on a single cell level has led to a more complete picture of the average state of an organism across time and space. Denoting Y as the vector of gene expression measurement and Z as the set of covariates, significant efforts have been made in modelling $E(Y|Z)$ to study differential expression of genes, ie. to determine if the observed difference or change in expression levels of genes between experimental conditions is statistically significant.

Compared to the bulk genomic protocols, single cell gene expression allows inference on cellular heterogeneity. Bulk experiments dealt with data averaged over cells, hence measuring

cell-cell variation is unfamiliar ground in gene expression methods. Grouping cells into predefined cell-types (assume that cells within a cell-type group are independent replicates), estimating cell-type specific models and quantifying the statistical differences therewith can lead to important insights into cell-type specific behaviour. One might be interested in modelling the functional form of $E(Y|Z)$ for different cell-types and quantify the variability between the estimated functions.

Finally one might desire inference on the co-expression of Y , which explains how the expression of one gene covaries with the others, an example studied in coexpression network estimation. This might help us understand how genes can regulate each other, and render deeper biological insights.

1.2 Structure of single cell gene expression datasets

Unlike microarray and bulk RNAseq datasets, scRNAseq data exhibits excess zero values due to the low per cell RNA input, biases in capture and amplification, transcriptional bursts, and other technical factors [10, 11, 12]. This no expression (zero values), due to a conflation of both biological and technical factors, results in an excessive number of zeroes leading to bimodal expression. Consequently, single cell models consider the gene expression distribution as a mixture of an unexpressed (zero) and a positively (non-zero) expressed population [13, 11, 14, 1]. The statistical model for Y is formulated as

$$\begin{aligned} Y_{i,j}|R_{i,j} &\sim f(\theta) \\ Y_{i,j} &= 0 \text{ with probability } 1 \text{ when } R_{i,j} = 0 \end{aligned} \tag{1.1}$$

where $f(\theta)$ denotes the distribution of gene j in cell i , given that it is detectable. The indicator variable $R_{i,j}$ indicates the presence or absence of expression for gene j in cell i . Most existing methods use the Poisson or the negative binomial distribution to model the read counts of individual genes directly [10, 13, 11, 15]. Another approach that is frequently used is to assume that $\log(Y_{i,j})$ follows a normal distribution, which leads to a more tractable mathematical theory than count distributions have [14, 16, 17, 1]. Additionally bulk RNA-seq

data have been historically modelled using both count and continuous distributions and the logarithmic transformation used to make the data continuous is well accepted in the field of bulk genomics [18, 19, 20]. There is also an ongoing debate in the field of single cell genomics as to whether the excess zeros present in the data need to be modelled differently assuming a sampling process different from the process used to model the true biological expression. Recent research has argued that count distributions like the negative binomial or a Poisson distribution with a Gamma prior on the mean, that eventually leads to a negative binomial posterior is enough for modelling the excess zeros in the expression data. What we observed in our experimental datasets is a very high percentage of zeroes, often close to 60 – 70% and the experimental scientists were more convinced that such severe percentage of zeroes were more a result of the technical errors than the actual biological process. We conducted several experiments to test this hypothesis [1] and concluded that a zero induced distribution would best model our experimental data [21, 22]. Therefore throughout the entire dissertation we stick to zero-induced distributions. However all our proposed methods could be well adapted to work without the zero-inflation part.

1.3 Dose Response Experiments

The central idea of toxicology lies in understanding the molecular mechanisms of action that makes a treatment toxic. An important tool for studying the mechanisms of action is to understand the response of a living organism to a range of doses of the toxic substance. The dose–response relationship, or the dose response curve, describes the magnitude of the response of an organism, as a function of exposure to a stressor (usually a toxin) after a certain exposure time [23]. Developing dose response models is crucial for determining safe, hazardous and beneficial levels of pollutants, drugs, foods and several other substances to which humans or other organisms are exposed [24]. These conclusions often form the basis for public policy. The U.S. Environmental Protection Agency (EPA) has developed extensive guidance and reports on dose–response modeling and assessment, as well as software [25].

The potency value derived from the dose response model is the benchmark dose (BMD). BMD is the dose or concentration that produces a predetermined change in the response rate of an adverse effect. This predetermined change in response is called the benchmark response (BMR), and is generally assumed to be the standard deviation of the response at zero dose [26]. Dose-dependent gene expression profiling (aka genomic dose-response studies (GDRS)) of drugs and chemicals has been proposed as an alternative test to rodent bioassays to assess human health risks. Application of single-cell RNA sequencing (scRNAseq) for the evaluation of chemicals, drugs, and food contaminants presents the rare opportunity of accounting for cellular heterogeneity in pharmacological and toxicological responses[27].

1.3.1 National Toxicology Program’s approach to genomic dose response modeling

The National Toxicology Program (NTP) proposes the use of the approach outlined below for both in vivo and in vitro genomic dose-response studies (GDRS) [26]. The steps in the process include (1) designing a dose response experiment with a sufficient number of doses that can effectively capture the shape of a dose response curve and accurately determine the benchmark dose (BMD) for all dose responsive genes, (2) designing a statistical test of hypothesis to determine whether the modelled data shows any effect to the treatment. This test acts a preliminary step for filtering out genes that aren’t dose responsive. (3) Conducting a second trend test for identifying genes that exhibit a biologically plausible response to the treatment, (4) fitting parametric dose response models derived from the Environmental Protection Agency (EPA) BMD software to identify a biological potency estimate, ie. benchmark dose (BMD) for each gene exhibiting a dose-related response to treatment, (5) grouping genes into predefined sets defined by gene ontologies and calculating the composite BMD of the gene set and (6) finally providing a biological explanations for the selected set of genes and BMD estimates. As pointed out by NTP, following all the steps for GDRS modelling leads to consistent modelling and facilitates the use of use of genomic

dose-response data in risk assessment. Assuming that the experimental process is adequate, I mainly focus on the statistical considerations of items 2-4.

1.3.1.1 Determining Adequate Signal in the Data

The first step of the GDRS protocol seeks to determine whether the signal in the data is adequate to model and likely to yield minimally reproducible findings [28]. Simply put, we are interested in finding whether a gene is dose responsive. Statistically, this can be formulated as an ANOVA problem. There is a population of interest for which there is a true quantitative outcome Y for each of the k levels of dose D . The population outcomes for each group have mean parameters $\mu_{1,j}, \dots, \mu_{K,j}$ with no restrictions on the pattern of means. The population variances for the outcome for each of the k groups defined by the levels of the explanatory variable all have the same value, σ_j^2 . We are interested in testing the null hypothesis that “nothing interesting is happening.”. For one-way ANOVA, we use $H_0 : \mu_1 = \mu_2 = \dots \mu_K$, which states that all of the population means are equal, without restricting what the common value is. The alternative must include everything else, which can be expressed as “at least one of the k population means differs from all of the other”. Use of a pre-filtering test to determine dose responsiveness helps to avoid modeling of data with no statistically plausible signal. Modeling data with no statistically significant effect is likely to yield unreproducible results with highly inaccurate estimates of BMD values. Further, given the huge number of genes (usually in the order of millions) it is an unnecessary computational burden to model genes with no statistically plausible signal.

An associated problem that is studied in gene expression studies is to compare two biological conditions in order to find differentially expressed (DE) genes. A gene is defined as DE if it is transcribed into different amounts of mRNA molecules per cell under the two conditions. However, since we do not observe the true levels of expression, carefully designed statistical tests help biologists understand to what extent a gene is DE. Currently, there are dozens of differential gene expression analysis (DGEA) approaches for scRNAseq data;

developed based on differences in assumptions, statistical methodologies, and study designs [29, 30, 31, 32, 33, 34]. A recent comparison of 36 approaches demonstrated acceptable performance for common bulk RNAseq tools such as edgeR and limma-trend, and MAST for single cell experiments, as well as common statistical tests such as the Wilcoxon rank sum (WRS) and the t-test [32]. However, most methods have been developed primarily for two group comparisons whereas our study designs consists of multiple groups. The use of two sample tests for multiple group study designs elevate the type I error rate warranting further investigation of these methods for multiple group dose-response study designs [35]. Chapter 2 proposes the first multiple group statistical testing framework designed exclusively for single cell dose response designs. The test statistic is derived for both the frequentist and Bayesian set-up and benchmarked against selected bulk and single cell DE analysis methods capable of accomodating multiple group designs. A further detailed background and analysis is delayed until Chapter 2.

1.3.1.2 Filtering of Measured Features

The second step of GDRS protocol applies a statistical trend test and effect size to filter genes. Statistically, it assumes that the experimenter has apriori knowledge that the response to the treatment has a known direction, and in most cases, the response increases in magnitude as the dose is increased. Therefore the statistical analysis is based on the assumption that the responses are monotonically ordered. Thus the test of hypothesis translates to $H_0 : \mu_1 = \mu_2 = \dots \mu_K$, versus $\mu_1 \leq \mu_2 \leq \dots \mu_K$. Several statistical tests have been proposed for testing this monotonically ordered alternative hypothesis. The test recommended by the US EPA and NTP is the William's Trend test [36] to identify monotonic trends.

Williams' test is a step-down trend test for testing several treatment levels with a zero control in a one -way ANOVA design with normally distributed errors of homogeneous variance. Let there be K groups including the control and let the zero dose level be indicated with $k = 1$ and the treatment levels as $2 \leq \dots \leq K$, then the following $K - 1$ hypotheses

are tested:

$$\begin{array}{ll}
H_{0,K-1} : \bar{x}_1 = \mu_2 = \dots = \mu_K, & H_{a,K-1} : \bar{x}_1 \leq \mu_2 \leq \dots \mu_K, \bar{x}_1 < \mu_K \\
H_{0,K-2} : \bar{x}_1 = \mu_2 = \dots = \mu_{K-1}, & H_{a,K-2} : \bar{x}_1 \leq \mu_2 \leq \dots \mu_{K-1}, \bar{x}_1 < \mu_{K-1} \\
\vdots & \vdots \\
H_{0,1} : \bar{x}_1 = \mu_2, & H_{a,1} : \bar{x}_1 < \mu_2,
\end{array}$$

The first step is to determine the maximum likelihood (ML) estimates $\widehat{\mu}_k$ under the alternative hypothesis that there is a response. This is estimated using the pool adjacent violators algorithm (PAVA) [37]. If the response is a decrease in mean, then the above formulation applies after a change of the inequality sign. The k^{th} test statistic $\bar{t}_i$ is calculated as follows:

$$\bar{t}_i = \frac{\widehat{\mu}_k - \bar{x}_1}{s^2 \sqrt{1/n_k - 1/n_1}}$$

where s^2 is an unbiased estimator of error variance of Y with ν degrees of freedom. The critical t -values as given in the tables of [36] for $\alpha = 0.05$ (one-sided) and are looked up according to the degree of freedoms (ν) and the order number of the dose level. Regarding genomic dose-response modeling, NTP acknowledges limitations to using a more traditional trend test such as the Williams' Test that identifies only monotonic trends.

1.3.1.3 Dose Response Curve Estimation

The third step of the genomic dose-response analysis focuses on fitting dose-response curves to each gene that exhibits a response to treatment as determined by the filtering approach described above. To model the data, parametric models such as polynomial, linear, power, Hill and exponential 2, 3,4 dose-response models are fitted to the measured features. Equations describing the models are given in the table below. A detailed description of all the models can be found in the EPA's Benchmark Dose Software (BMDS) Version 3.2 user guide. The current GDRS modelling softwares [28] are only capable of modelling bulk RNA-sequencing datasets, with no statistical framework or recommendation existing for dose response modelling of single cell datasets. The bulk gene expression data is usually log transformed and presumed to have a normal distribution and each model is run assuming constant variance.

Model	Formula
Linear	$\mu(dose) = \gamma + \beta dose$
Polynomial (order- q)	$\mu(dose) = \gamma + \beta_1 dose + \beta_2 dose^2 + \dots + \beta_q dose^q$
Hill	$\mu(dose) = \gamma + (v * dose^n)/(k^n + dose^n)$
Exponential 2 or 3	$\mu(dose) = a * \exp(sign(b * dose^d))$
Exponential 3 or 4	$\mu(dose) = a * (c - (c - 1) * \exp(dose^d))$
Power	$\mu(dose) = \gamma + \beta dose^\delta$

Table 1.1 Parametric models for genomic dose response analysis

Let us denote a sample of n independent and identically distributed (IID) pairs of variables, $\mathcal{B} = \{(Z_i, Y_i)\}, i = 1, \dots, n$, as the data. Z_i can be further segmented as $Z_i = \{(D_i, X_i)\}$ where D_i represents the dose vector and X_i represents the vector of other confounding covariates such as age, gender etc. The main interest of dose response studies lies in modelling $E(Y_i)$ as a function of the dose vector; ie. $E(Y_i|D_i) = \zeta(D_i, \theta)$, where θ is the vector of unknown parameters. In general, the methods for estimating dose response can be classified as parametric or nonparametric. Parametric methods assume a model $\zeta(D_i, \theta)$ for the DR curve where θ is the vector of unknown parameters. Considerable amount of statistical methodology exists in parametric modelling of dose-response studies, employing parametric models, such as Logistic, Exponential, and Gompertz as well as others [38] and the US EPA [28] gives guidelines on how to employ these methods. These models are generally fit using standard techniques such as maximum likelihood (ML) or restricted maximum likelihood (RML) and their functional forms are monotonic allowing for tractable estimation of relevant BMD's. It is well known that if a parametric model is correctly specified then the ML/RML estimators are efficient. However, in many cases, it is difficult to correctly specify the parametric form of the dose-response curve because the biological mechanism of drug action or toxicity may be complex and the form of the dose-response curve is unknown apriori [39]. When the parametric model is misspecified, the corresponding curve estimate may be severely biased. In addition, fitting of extremely flexible polynomial models with high orders to match difficult non-monotonic curve shapes, can lead to severe overfitting. The standard workaround is to fit multiple parametric models, each of which may fit the data reasonably, but produce

a range of BMD estimates that accounts for model uncertainty in the estimation process. Accounting for this uncertainty using model averaging (MA) has received recent attention in the literature [40]. However, despite simulation studies suggesting that model-averaged estimates provide BMD estimates that have low bias and nominal coverage properties, these estimates may fail to adequately describe the true uncertainty for DR curves on the edge of the MA model space, hence resulting in inaccurate inference. Further, as the MA results are available on limited number of parametric forms, models that make fewer assumptions on parametric forms may allow for better estimation of the dose response curves [39].

To enhance the robustness of the estimation of the dose-response curve, many nonparametric methods have been proposed. [41] proposed MLE estimation procedures under the assumption that the dose response curve is sigmoid and non-decreasing. [42] proposed a method for non-parametric estimation of the dose response curve under monotonicity constraints using B-splines regression. [43] used kernel estimators to obtain estimates for the dose-response curve under general shape restrictions. In the Bayesian setup, several approaches have been proposed for non-parametric estimation of dose response curves under monotonicity constraints. From a Bayesian perspective, one specifies a suitable prior on the regression function that induces monotonicity and then inference is based on the posterior distribution. [44] employed an additive model with a prior imposed on the slope of the piecewise-linear functions. [45] adopted mixture modelling of shifted and scaled probability distribution functions. [46] propose a Gaussian process model with a posterior projection approach for shape-constrained curves. In contrast to the parametric methods, all the non-parametric approaches are flexible and learn the shape of the dose response curves based on the data. To the best of our knowledge, none of the existing literature has considered building a non-parametric regression model for both unconstrained and monotonicity constrained estimation of dose response curves for single cell experiments while accounting for technical confounders. Incorporation of such a flexible modelling structure will allow for more accurate modeling of a variety of non-monotonic responses. Further, the incorporation

of a flexible nonparametric modeling approach to describe the shape of the diversity of dose-response behaviors observed in genomic dose-response will lead to the estimation of more accurate BMD's.

1.4 Covariation Analysis

A gene co-expression network (GCN) is an undirected graph, where nodes correspond to different genes, and edges connecting the nodes denote the co-expression relationships between genes. GCNs can help people learn the functional relationships between genes and infer and annotate the functions of unknown genes. GCN reconstruction attempts to infer this co-expression network from high-throughput data using statistical and computational approaches. Multiple methods encompassing varying mathematical concepts have been proposed during the last decade to infer GCNs using gene expression data from bulk population sequencing technologies, which accumulate expression profile from all cells in a tissue. These methods can be broadly classified into two groups: the first group infers a static GCN, considering steady state of gene expression, while the second group uses temporal measurements to capture the expression profile of the genes in a dynamic process. A thorough evaluation of the static and dynamic models used in bulk GCN reconstruction can be found in [47, 48]. In the static group, several computational approaches ranging from correlation networks [49], Gaussian graphical models [50], Bayesian networks [51], regression analyses [52, 53], and information theoretical approaches [54, 55] have been used for network inference from population-level data. Similarly, several models have been proposed for the analysis of population level dynamic gene expression data [56, 57, 58]. In this dissertation we will mainly focus on estimating a static co-expression network. Mathematically we are interested in determining the relationship between Y_i and Y_j . The first step is to construct a symmetric adjacency matrix A , where $A_{i,j}$ is a weighted adjacency score in the range from 0 to 1 between genes i and j . $A_{i,j}$ measures the level of association between gene expression vectors Y_i and Y_j . Clustering methods can then be applied to search for gene modules based on the

resulting distance matrix [49]. The modules can serve to understand functional relationships between known disease genes and candidate genes. Gene modules can also be used to detect regulatory genes and study the regulatory mechanisms in various organisms [59]. Further background will be delayed until Chapter 4.

CHAPTER 2

BAYESIAN SINGLE CELL RNASEQ DIFFERENTIAL GENE EXPRESSION TEST FOR DOSE RESPONSE STUDY DESIGNS

2.1 Single Cell Dose Response Experiments

Single-cell transcriptomics enables researchers to investigate homeostasis, development and disease at unprecedented cellular resolution [60, 6, 61, 62, 63]. As with any new innovative technology, diverse tools soon follow to address specific applications and unique challenges. Currently, there are dozens of differential gene expression analysis (DGEA) approaches for single-cell RNAseq (scRNAseq) data; developed based on differences in assumptions, statistical methodologies, and study designs [29, 30, 31, 32, 33, 34]. A recent comparison of 36 approaches demonstrated acceptable performance for common bulk RNAseq tools such as edgeR and limma-trend, and MAST for snRNAseq, as well as common statistical tests such as the Wilcoxon rank sum (WRS) and the t-test [32]. However, most methods have been developed primarily for two group comparisons whereas study designs typical of pharmacology and toxicology experiments such as dose-responses consist of multiple groups. The use of two sample tests for multiple group study designs elevate the type I error rate warranting further investigation of these methods for multiple group dose-response study designs [35].

Dose-response studies are used to derive the efficacy and/or safety margins such as effective dose and the point of departure (POD). Significant efforts by the toxicology and regulatory communities have suggested that acute (< 14 days) and sub-acute (14–28days) transcriptomic studies as viable alternative to the current standard 2-year rodent bioassay that significantly reduces the time and resources needed to assess risk [64, 65, 26]. Gene expression profiling at single-cell resolution could further support such evaluations by identifying cell-specific dose-dependent responses indicative of an adverse event. The U.S. National Toxicology Program (NTP) recently reported a robust DGEA approach is essential to de-

iving biologically relevant PODs [26]. However, concerns regarding the inclusion of false positives that produce less conservative POD estimates potentially leads to incorrect classification of mode-of-action(MoA), thus highlighting the importance of controlling type I error rates [66, 67].

Unlike microarray and bulk RNAseq datasets, single-cell RNAseq (scRNAseq) data exhibits excess zero values due to the low per cell RNA input, biases in capture and amplification, transcriptional bursts, and other technical factors[12]. This no expression (zero values), due to a conflation of both biological and technical factors, results in an excessive number of zeroes in an otherwise continuous measure [14]. Therefore, traditional tests of differential gene expression, based on the assumption of a normal distribution, fail to correctly model the bimodality of single cell gene expression[16]. Consequently, scRNAseq test methods usually consider the gene expression distribution as a mixture of a unexpressed (zero) and a positively (non-zero) expressed population [14, 16, 68]. For example, the Seurat Bimod approach tests for differential gene expression using a likelihood ratio test designed for the said mixture population. MAST extends the Seurat Bimod test to a two-part generalized linear model structure capable of incorporating covariates [14, 16]. Given the improved performance of MAST [32, 14, 16], we hypothesized that multiple group tests developed assuming the same distributional framework would be most favorable for dose-response study designs. Furthermore, a Bayesian approach which considers prior knowledge is anticipated to minimize type I error rates [69, 70].

The aim of the presented study is to evaluate the performance of existing and novel DGEA test methods on dose-response scRNAseq datasets. To reduce the rate of false positives we propose a novel, multiplicity corrected, Bayesian multiple group test (scBT) designed exclusively for DGEA of dose-response scRNAseq data. Two other fit-for-purpose frequentist multiple group tests are also examined: (i) a multiple group extension of the Seurat Bimod test and (ii) a simple extension of test (i) to a generalized linear model framework. Existing and proposed methods are benchmarked on simulated and real experimental dose-response

datasets. Using simulated datasets we were able to investigate the influence of various parameters such as number of cells, and illustrate how using different test methods can aid in gaining biological insight on the role of individual cell types on the pathophysiological consequences of exposure.

2.2 Materials and Methods

2.2.1 Animal handling and treatment

Male C57BL/6 mice aged postnatal day (PND) 25 were obtained from Charles Rivers Laboratories (Kingston, NY) were housed and treated as previously described [71]. Mice were housed in Innovive cages (San Diego, CA) with ALPHA-dri bedding (Shepherd Specialty Papers, IL) at 23°C, 30-40% relative humidity, and a 12:12 h light:dark cycle. Aquavive water (Innovive) and Harlan Teklad 22/5 Rodent Diet 8940 (Madison, WI) was provided ad libitum. On PND 29, randomly assigned mice were gavaged at Zeitgeber time (ZT) 0 with 0.1 mL sesame oil vehicle (Sigma-Aldrich, St. Louis, MO), 0.01, 0.03, 0.1, 0.3, 1, 3, 10 or 30 μ g/kg TCDD every 4 days for 28 days (7 total administered doses). At day 28 mice were euthanized by CO₂ asphyxiation and livers were immediately flash frozen in liquid nitrogen and stored at -80°C. All animal procedures were approved by the Michigan State University (MSU) Institutional Animal Care and Use Committee (IACUC) and reporting of in vivo experiments follow the Animal Research: Reporting of In Vivo Experiments (ARRIVE) [72] and Minimum Information about Animal Toxicology Experiments (MIATE) guidelines (<https://fairsharing.org/FAIRsharing.wYScsE>).

2.2.2 Real scRNAseq and snRNAseq datasets

Hepatic single-nuclei RNA-sequencing (snRNAseq) was performed as previously described using the 10× Genomics Chromium Single Cell 3' v3.1 kit (10X Genomics, Pleasanton, CA) [73]. Briefly, nuclei were isolated using EZ Lysis Buffer (Sigma-Aldrich), homogenized by disposable Dounce homogenizer, washed, filtered using a 70- μ m cell strainer. The nuclei

pellet was resuspended in buffer containing DAPI (10 $\mu\text{g}/\text{ml}$), filtered using a 40- μm strainer, and immediately sorted using a BD FACSAria IIu (BD Biosciences, San Jose, CA) at the MSU Pharmacology and Toxicology Flow Cytometry Core (facs.iq.msu.edu/). Sequencing (150-bp paired end) was performed at a depth of 50,000 reads/cell using a NovaSeq6000 at Novogene (Beijing, China). CellRanger v3.0.2 (10x Genomics) was used to align reads to mouse gene models (mm10, release 93) including introns and exons to consider both pre-mRNA and mature mRNA gene models. Seurat was used to integrate and log-normalize expression data [74]. The data is available on the Gene Expression Omnibus (GEO) at accession ID GSE184506 and R package versions are listed in Supplementary Table S1. Additional real datasets were publicly available. Hepatic whole-cell generated using the 10X Genomics platform was obtained from GEO (GSE129516)[63]. Hepatic single-nuclei processed as the dose-response data for control and high dose TCDD treatment (0 and 30 $\mu\text{g}/\text{kg}$) was obtained from GEO (GSE148339). Peripheral blood mononuclear cell (PBMC) data also generated using the 10x Genomics platform and Seurat was obtained from the SeuratData R package [74].

Gene set enrichment analysis of experimental data was performed using the fgsea v1.14 R package on gene lists sorted by significance values (e.g. P-value). Gene sets from BIO-CARTA, KEGG, PANTHER and WIKIPATHWAYS were obtained from the Gene Set Knowledgebase (GSKB; <http://ge-lab.org/gskb/>) and filtered for gene sets containing 15–250 genes. Gene sets were agglomerated based on overlap of gene membership and only those showing $\geq 50\%$ overlap were considered similar for subsequent network analyses. Visualization and calculation of measures of centrality were determined using igraph v1.2.7. Gene sets were considered enriched when adjusted P-value < 0.05

2.2.3 Dose-response data simulation

To simulate dose-response scRNAseq datasets we developed a wrapper for the Splatter R package[75]. Splatter simulates counts using parameters estimated from real data to set the

mean expressions, variance, and outlier probability. Other parameters such as the number of cells, genes, probability of being differentially expressed, mean fold-change of DE genes (location) and standard deviation of fold-change of DE genes (scale) were manually assigned to best reflect real data. The wrapper (SplattDR) leverages the group simulation feature of Splatter by applying a multiplicative factor estimated using dose-response models in 2.1 based on the US EPA Benchmark Dose Software[28]. SplattDR R package is available at (github.com/zacharewskilab/splattdr).

Model	Formula
Hill	$\mu(dose) = \gamma + (v * dose^n)/(k^n + dose^n)$
Exponential 2 or 3	$\mu(dose) = a * \exp(sign(b * dose^d))$
Exponential 3 or 4	$\mu(dose) = a * (c - (c - 1) * \exp(dose^d))$
Power	$\mu(dose) = \gamma + \beta dose^\delta$

Table 2.1 Dose-response models for simulation of scRNAseq data

2.2.4 Single cell RNA-seq Hurdle Model

We model the log-normalized gene expression matrix using a hurdle distribution wherein the rate of gene expression is assumed to follow a Bernoulli distribution and conditional on a cell expressing the gene, the log-normalized expression level is assumed to follow a Gaussian distribution[14]. We denote $Y_{i,j}$ to be the log-normalized expression value of gene j in cell i , for $i = 1, \dots, n$ and $j = 1, \dots, p$. To characterize the bimodal properties of single cell data, for a given cell, a gene is defined to be either positively expressed or undetected. Define $R_{ij} = I[Y_{ij} > 0]$ to be the indicator variable denoting the presence or absence of expression for gene j in cell i . Following [14], the log-normalized gene expression values are modeled as follows:

$$\begin{aligned}
Y_{i,j}|R_{i,j} &\sim Normal(\mu_j, \sigma_j^2), \\
Y_{i,j} &= 0 \text{ with probability } 1 \text{ when } R_{i,j} = 0, \\
R_{i,j} &\sim Bernoulli(\omega_j),
\end{aligned} \tag{2.1}$$

where μ_j and σ_j^2 denote the mean and variance of the gene expression level, conditional on the gene being expressed and ω_j denotes the rate of gene expression of gene j across all cells.

2.2.4.1 Hypothesis Formulation

We now assume that our data has been collected under K conditions (doses), and denote the data by $D_{k,o} \equiv \{(Y_{k,i,j}, R_{k,i,j}), i = 1, \dots, n_k\}$. The underlying populations for the sample data $D_{k,o}$ for $k = 1, 2, \dots, K$, dose groups are assumed to be identified by the parameters $(\mu_{k,j}, \sigma_j^2, \omega_{k,j})$. The aim of this study is to test for difference in gene expression patterns between the different dose groups. Traditionally one would perform an ANOVA test to detect changes in mean across groups for samples with continuous measurements. However, to account for the bimodality in single cell gene expression distribution, the test should detect for changes in μ_j and ω_j simultaneously, as both could drive differential gene expression. Therefore we define,

$$H_0 : \quad \mu_{1,j} = \mu_{2,j} = \dots \mu_{K,j} = \mu_j \text{ and } \omega_{1,j} = \omega_{2,j} = \dots \omega_{K,j} = \omega_j. \quad (2.2)$$

versus the alternative

$$H_a : \quad \begin{aligned} &\mu_{k,j} \text{ is different for at least one } k \text{ and} \\ &\omega_{k,j} \text{ is different for at least one } k, k = 1, \dots, K \end{aligned}$$

2.2.4.2 Single cell Bayesian Hurdle model Analysis (scBT)

Given the single cell RNA-seq hurdle model structure, we assume that a priori, given σ_j^2 , $\mu_{k,j} \sim \text{Normal}(m_{k,0}, \tau_{k,\mu} \sigma_j^2)$, $\sigma_j^2 \sim \text{IG}(a_\sigma, b_\sigma)$, $\omega_{k,j} \sim \text{Beta}(a_{k,\omega}, b_{k,\omega})$, where IG is the inverse gamma distribution with shape a_σ and scale b_σ and $m_{k,0}$, $\tau_{k,\mu}$, a_σ , b_σ , $a_{k,\omega}$, $b_{k,\omega}$ are the hyperparameters. Given the large number of gene-wise model fits arising from a single cell experiment, there is a pressing need to allow for a parallel structure whereby the same model is fitted to each gene. The prior distributions on the parameters describe how the unknown coefficients $\mu_{k,j}$, $\omega_{k,j}$ and σ_j^2 vary across the genes and the dose groups while allowing for information borrowing between the genes. Now, based on the model assumptions, we propose a Bayesian test for simultaneously testing the differences in mean gene expression

and dropout proportions as formulated in 2.2.4.1. Under the null hypothesis the marginal likelihood is written as

$$\begin{aligned}
\mathcal{L}_{H_0,j} &= \int \int \int \prod_{k=1}^K \left\{ \prod_{i=1}^{n_k} \left[\frac{1}{\sqrt{2\pi}\sigma_j} \exp \left\{ -\frac{(Y_{k,i,j} - \mu_j)^2}{2\sigma_j^2} \right\} \omega_j \right]^{R_{k,i,j}} (1 - \omega_j)^{1-R_{k,i,j}} \right\} \\
&\times \pi(\mu_j | \sigma_j^2) \pi(\sigma_j^2) \pi(\omega_j) d\mu_j d\sigma_j^2 d\omega_j \\
&= \frac{1}{(2\pi)^{(\sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j})/2}} \times \frac{1}{\sqrt{1 + \tau_\mu \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j}}} \\
&\times \frac{1}{\Gamma(a_\sigma) b_\sigma^{a_\sigma}} \times \frac{\Gamma(a_\sigma + (\sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j})/2)}{(1/b_\sigma + \mathcal{A}_{tot}/2)^{a_\sigma + (\sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j})/2}} \\
&\times \frac{Beta(a_\omega + (\sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j}), b_\omega + \sum_{k=1}^K n_k - (\sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j}))}{Beta(a_\omega, b_\omega)}, \tag{2.3}
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{A}_{tot} &= \left\{ \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j} Y_{k,i,j}^2 + \frac{m_0^2}{\tau_\mu} \right\} - \left\{ \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j} + \frac{1}{\tau_\mu} \right\}^{-1} \\
&\times \left\{ \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j} Y_{k,i,j} + \frac{m_0}{\tau_\mu} \right\}^2.
\end{aligned}$$

Under the alternative hypothesis we compute the marginal likelihood without any restriction on the K means $\mu_{k,j}$ and the dropout parameter $\omega_{k,j}; k = 1, 2, \dots, K$. Particularly, we assume that $\mu_{k,j} \sim Normal(m_{k,0}, \tau_{k,\mu} \sigma_j^2)$, and $\sigma_j^2 \sim IG(a_\sigma, b_\sigma)$, $\omega_{k,j} \sim Beta(a_{k,\omega}, b_{k,\omega}); k = 1, 2, \dots, K$. Now, the marginal likelihood under the alternative hypothesis is given by

$$\begin{aligned}
\mathcal{L}_{H_a,j} &= \int \cdots \int \left\{ \prod_{k=1}^K \prod_{i=1}^{n_k} \left[\frac{1}{\sqrt{2\pi}\sigma_j} \exp \left\{ -\frac{(Y_{k,i,j} - \mu_{k,j})^2}{2\sigma_j^2} \right\} \omega_{k,j} \right]^{R_{k,i,j}} (1 - \omega_{k,j})^{1-R_{k,i,j}} \right\} \\
&\times \prod_{k=1}^K \left\{ \pi(\mu_{k,j}) \pi(\omega_{k,j}) \right\} \pi(\sigma_j^2) \prod_{k=1}^K \left\{ d\mu_{k,j} d\omega_{k,j} \right\} d\sigma_j^2 \\
&= \frac{1}{(2\pi)^{(\sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j})/2}} \times \frac{1}{\prod_{k=1}^K \sqrt{1 + \tau_{k,\mu} \sum_{i=1}^{n_k} R_{k,i,j}}} \\
&\times \frac{1}{\Gamma(a_\sigma) b_\sigma^{a_\sigma}} \times \frac{\Gamma(a_\sigma + \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j}/2)}{(1/b_\sigma + \sum_{k=1}^K \mathcal{A}_k/2)^{a_\sigma + \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j}/2}}
\end{aligned} \tag{2.4}$$

$$\times \prod_{k=1}^K \frac{Beta(a_{k,\omega} + \sum_{i=1}^{n_k} R_{k,i,j}, b_{k,\omega} + n_k - \sum_{i=1}^{n_k} R_{k,i,j})}{Beta(a_{k,\omega}, b_{k,\omega})} \quad (2.5)$$

Now, under the assumption that $a_\omega = a_{k,\omega}$, $b_\omega = b_{k,\omega}$, $\tau_\mu = \tau_{k,\mu}$, for $k=1,2, \dots K$, and $r! \sim \sqrt{2\pi r}(r/e)^r$ we have,

$$\begin{aligned}
\frac{\mathcal{L}_{H_{0,j}}}{\mathcal{L}_{H_{a,j}}} &= \left(\sqrt{2\pi}e\right)^{1-K} \times \frac{\prod_{k=1}^K \sqrt{1 + \tau_{k,\mu} \sum_{i=1}^{n_k} R_{k,i,j}}}{\sqrt{1 + \tau_\mu \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j}}} \\
&\times \left(\frac{1/b_\sigma + \sum_{k=1}^K \mathcal{A}_k/2}{1/b_\sigma + \mathcal{A}_{tot}/2}\right)^{a_\sigma + \frac{1}{2}(\sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j})} \\
&\times \sqrt{\frac{(a_\omega + \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j} - 1) \times (b_\omega + \sum_{k=1}^K n_k - \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j} - 1)}{(a_\omega + b_\omega + \sum_{k=1}^K n_k - 1)}} \\
&\times \frac{(a_\omega + \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j} - 1)^{(a_\omega + \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j} - 1)}}{(a_\omega + b_\omega + \sum_{k=1}^K n_k - 1)^{(a_\omega + b_\omega + \sum_{k=1}^K n_k - 1)}} \\
&\times (b_\omega + \sum_{k=1}^K n_k - \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j} - 1)^{(b_\omega + \sum_{k=1}^K n_k - \sum_{k=1}^K \sum_{i=1}^{n_k} R_{k,i,j} - 1)} \\
&\times \prod_{k=1}^K \sqrt{\frac{(a_\omega + b_\omega + n_k - 1)}{(a_\omega + \sum_{i=1}^{n_k} R_{k,i,j} - 1) \times (b_\omega + n_k - \sum_{i=1}^{n_k} R_{k,i,j} - 1)}} \\
&\times \prod_{k=1}^K \left\{ \frac{(a_\omega + b_\omega + n_k - 1)^{(a_\omega + b_\omega + n_k - 1)}}{(a_\omega + \sum_{i=1}^{n_k} R_{k,i,j} - 1)^{(a_\omega + \sum_{i=1}^{n_k} R_{k,i,j} - 1)}} \right. \\
&\times \left. \frac{1}{(b_\omega + n_k - \sum_{i=1}^{n_k} R_{k,i,j} - 1)^{(b_\omega + n_k - \sum_{i=1}^{n_k} R_{k,i,j} - 1)}} \right\} \\
&\times \left\{ \frac{\Gamma(a_\omega)\Gamma(b_\omega)}{\Gamma(a_\omega + b_\omega)} \right\}^{(K-1)}
\end{aligned}$$

The Bayes factor is then defined as

$$BF_{01,j} = \frac{\mathcal{L}_{H_{0,j}}}{\mathcal{L}_{H_{a,j}}} \times \frac{\pi(H_a)}{\pi(H_0)} \quad (2.6)$$

where $\pi(H_a)$ and $\pi(H_0)$ are the prior probabilities for the alternative and null model, respectively. The hyperparameters are obtained by maximising the marginal likelihood under the null and the alternative hypothesis. Detailed derivations of the likelihood function and the Bayes Factor are provided in Supplementary Material. Using the test of hypothesis described in Equation (2.2) scBHM conducts a test of DE for each gene independently. To control for multiplicity we adopt the FDR correction approach discussed in [76]. The rejection threshold is estimated in terms of the posterior probabilities of the null hypothesis, $p(H_{0,j}|D_j)$. For a target FDR α , the procedure rejects all hypotheses with $p(H_{0,j}|D_j) < \zeta$

, where $p(H_{0,j}|D_j) = [1 + \frac{1}{BF_{01,j}}]^{-1}$ and ζ is the largest value such that $\frac{C(\zeta)}{J(\zeta)} \leq \alpha$ where, $J(\zeta) = \{j : p(H_{0,j}|D_j) \leq \zeta\}$ and $C(\zeta) = \sum_{j \in J(\zeta)} p(H_{0,j}|D_j)$.

2.2.4.3 Multiple group Likelihood Ratio Test (LRT)

To carry out a direct performance comparison with scBT, we extend the Seurat Bimod [14] for multiple groups. Assuming that all the K groups have the same variance σ_j^2 and omitting the index j for clarity, the likelihood ratio test can be defined as;

$$\Lambda(Y, R) = \frac{\sup_{\theta \in H_0} L(\theta|Y, R)}{\sup_{\theta \in H_a} L(\theta|Y, R)}$$

where the likelihood can be written as;

$$L(\theta|Y, R) = \prod_k \omega_k^{e_k} (1 - \omega_k)^{n_k - e_k} \prod_{i \in C_k} f(Y_{i,k}|\mu_k, \sigma^2)$$

where Y and R represent the gene observation vector and the gene indicator vector across K dose groups, $\theta = \{\mu_k, \sigma^2, \pi_k, k = 1, \dots, K\}$ is the vector of unknown parameters, C_k is the set of cells expressing the gene in group k (*i.e.* $C_k = \{i : R_{i,k} = 1\}$), $e_k = \sum_i R_{i,k}$ is the number of cells expressing the gene in group k and f is the density function of the normal distribution with parameters μ_k and σ^2 . Therefore we can write,

$$\begin{aligned} \Lambda(Y, R) &= \frac{\sup_{\{\omega, \mu, \sigma^2\}} \omega^{(\sum_k e_k)} (1 - \omega)^{(\sum_k n_k - \sum_k e_k)} \prod_k \prod_{i \in C_k} N(Y_{ik}|\mu, \sigma^2)}{\sup_{\{\omega_k, \mu_k, \sigma^2; k=1, \dots, K\}} \prod_k \omega_k^{e_k} (1 - \omega_k)^{(n_k - e_k)} \prod_k \prod_{i \in C_k} N(Y_{ik}|\mu_k, \sigma^2)} \\ &= \frac{\sup_{\{\omega\}} \omega^{(\sum_k e_k)} (1 - \omega)^{(\sum_k n_k - \sum_k e_k)}}{\sup_{\{\omega_k, k=1, \dots, K\}} \prod_k \omega_k^{e_k} (1 - \omega_k)^{(n_k - e_k)}} \\ &\quad \times \frac{\sup_{\{\mu, \sigma^2\}} \prod_k \prod_{i \in C_k} N(Y_{ik}|\mu, \sigma^2)}{\sup_{\{\mu_k, \sigma^2; k=1, \dots, K\}} \prod_k \prod_{i \in C_k} N(Y_{ik}|\mu_k, \sigma^2)} \\ &= \prod_k \left\{ \frac{\sum_k e_k}{\sum_k n_k} \right\}^{e_k} \cdot \left\{ \frac{1 - \frac{\sum_k e_k}{\sum_k n_k}}{1 - \frac{e_k}{n_k}} \right\}^{n_k - e_k} \cdot \left\{ 1 + \frac{\sum_k e_k (Y_k^+ - \bar{Y}^+)^2}{\sum_k \sum_{i=1}^{e_k} (Y_{ik}^+ - \bar{Y}_k^+)^2} \right\}^{-\frac{\sum_k e_k}{2}} \\ &= \Lambda_b(R) \cdot \Lambda_n(Y^+) \end{aligned}$$

where Λ_b is a binomial LRT, Λ_n is a normal LRT, Y^+ is the set of positive Y values, $\bar{Y}_k^+$ is the mean of the positive Y values in the k^{th} group and $\bar{Y} = \frac{1}{K} \sum_{k=1}^K \bar{Y}_k$. Thus, it can be shown

that the combined LRT is the product of a binomial and a normal LRT statistic Λ_b and Λ_n , both of which are derived using classical statistical theory. Applying classical asymptotic results about LRTs, $-2\log\Lambda(Y, R)$ converges to a χ^2 distribution with $(2K - 2)$ degrees of freedom under H_0 . We note here that the sample size for the χ^2 statistic is not n , but $n_+ = \sum_{k=1}^K \omega_k n_k$ and n_+ is sufficiently large for our simulation and real data experiments.

2.2.4.4 Linear model-based Likelihood Ratio Test (LRT linear)

The generalized linear model approach MAST was identified as one of the top performing tests for pairwise differential expression testing [32, 16]. Deriving from their approach, the LRT multiple test is extended to a generalized linear model framework, where the mean and the dropout proportions are modelled as a linear function of the dose groups (assumed to be a continuous covariate). Using the same distributional assumptions defined in 2.1 we fit a logistic regression model for the discrete variable R and a Gaussian linear model for the continuous variable Y conditional on $(R = 1)$ independently, as follows: $E(Y_{ij}|R_{ij} = 1) = m_{0,j} + m_{1,j} * d_i$ and $\text{logit}\{P(R_{ij} = 1)\} = \psi_{0,j} + \psi_{1,j} * d_i$, where d represents the continuous dose groups. Under this modelling approach, the null hypothesis described in Equation 2.2 can be rewritten as $H_0 : E(Y_{ij}|R_{ij} = 1) = m_{0j}$ and $\text{logit}\{P(R_{i,j} = 1)\} = \psi_{0j}$. The regression models are fit using the *lm* and *brglm* functions in the *stats* and *brglm* R packages. The likelihood ratio test statistic is computed using the same statistical theory discussed for the LRT multiple test and it asymptotically follows a χ^2 distribution under H_0 .

2.2.5 Benchmarking method selection

Our fit-for-purpose tests were benchmarked to existing differential expression testing methods or their multiple group equivalent based on previously reported performance, ability to consider multiple groups, or whether they served as foundation for the scBT and multiple group LRT (LRT multiple) tests developed here. Seurat Bimod served as foundation for the

scBT and LRT multiple tests as previously outlined, and MAST was identified as one of the top performing test for two group comparisons [32]. Similarly, limma-trend performed well for two sample comparisons and can consider multiple groups. The Wilcoxon Rank Sum test was identified as providing excellent balance between its ability to identify DE genes and speed, and is the default test for the Seurat R package for scRNAseq analysis. It was also reported that the t test performed well and therefore we included the ANOVA and Kruskal-Wallis (KW) tests, a parametric and non-parametric alternative of the t test for multiple group comparisons. All tests were run without correction for batch effects or other nuisance covariates. Multiplicity for each test was controlled using FDR correction [31]. All tests, including scBT, LRT multiple and LRT Linear are available in our scBT R package (github.com/satabdisaha1288/scBT). R session information is listed Supplementary Table S1. A flow diagram outlines our benchmarking approach (Figure 2.1).

2.2.5.1 Seurat Bimod

Seurat Bimod test [14] is a pairwise differential gene expression testing approach developed assuming the single cell RNA-seq hurdle model framework. The test is formulated as H_0 : the mean and the dropout parameters of the gene vector under two dose groups are equal versus H_a : the mean and the dropout parameters differ over the two groups. The LRT based test statistic $-2\log\Lambda(y, r)$ converges to a χ^2 distribution with two degrees of freedom under H_0 . The computations are carried out using the R Package *Seurat*.

2.2.5.2 MAST

MAST [16] proposes a two-part generalized linear model for differential expression analysis of scRNAseq data. The first part models the rate of gene expression using logistic regression $\text{logit}(\omega_{ij}) = X_i\beta_j^\omega$ and the second part uses a linear model to express the positive gene-expression Y_{ij} , conditional on R_{ij} as $\mu_{ij} = X_i\beta_j^\mu$; where β_j^ω and β_j^μ are the coefficients of the covariates used in the logistic and linear regression model respectively. A test with an

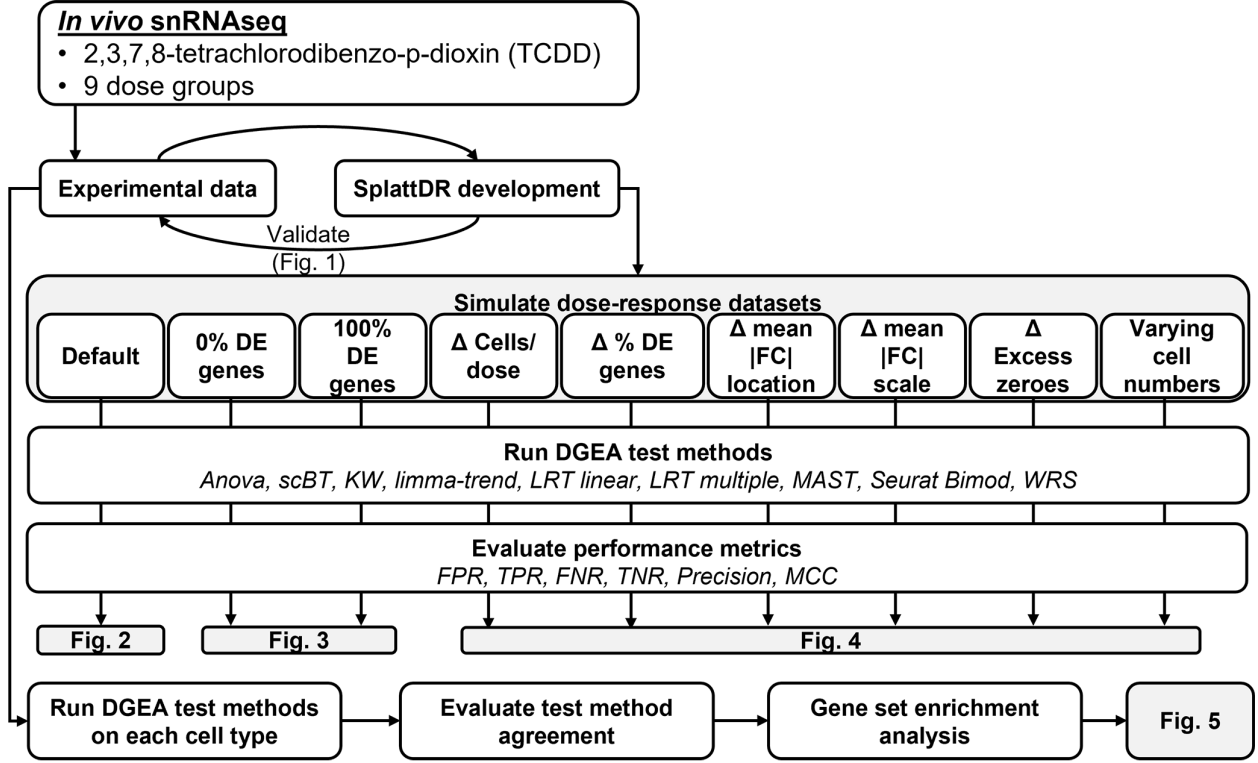


Figure 2.1 Flow diagram of the simulation, benchmarking, and experimental data evaluation strategy presented in the manuscript. Briefly, SplattDR was developed to simulate dose-response scRNAseq data and validated based on experimental dose-response data. Simulated datasets were generated varying diverse parameters 10 times and then used to assess the performance of each test method. Each test method was also assessed using experimental data from the hepatic snRNAseq dose response dataset obtained from male mice gavaged every 4 days for 28 days with 0.01, 0.03, 0.1, 0.3, 1, 3, 10, or 30 $\mu\text{g}/\text{kg}$ TCDD. Related figures for each analysis from the main body are noted.

asymptotic χ^2 null distribution is employed for identifying DEGs and multiplicity is controlled using FDR correction[77]. Despite the fact that LRT-linear and MAST have the same hurdle regression framework, the estimation process for the two methods has some significant differences. First, to achieve shrinkage of the continuous variance, MAST assumes a gamma prior distribution on the precision (inverse of variance) parameter and estimates its posterior maximum likelihood estimator (MLE) and uses that in place of the regular MLE of the precision parameter. Second, it fits a Bayesian logistic regression model for the discrete component by assuming Cauchy distribution priors centered at zero for the regression parameters. This is

done to deal with cases of “linear separation” where the parameter estimates diverge to $\pm\infty$ and the Fisher information matrix becomes singular. And finally, it considers the cellular detection rate defined as $CDR_i = \frac{1}{p} \sum_{j=1}^P R_{ij}$ to be a covariate in both the logistic and linear regression models. LRT linear on the other hand simply fits the non-Bayesian linear and the logistic regression models without considering variance shrinkage or adjustment for additional covariates.

2.2.5.3 Limma-trend

Limma-trend [20] proposes a linear model based differential expression approach for modelling RNA-seq experiments of arbitrary complexity. Their framework models the mean gene expression as a function of several continuous and categorical covariates. A separate linear model is fitted for each gene, but the gene-wise models are linked by global parameters using the parametric empirical Bayes approaches [78]. The global variance estimated by the empirical Bayes procedure also incorporates a mean variance trend, allowing better modelling of low abundance genes. Finally, test of differential gene expression is carried out by testing the significance of one or more coefficients of the fitted linear model.

2.2.5.4 Wilcoxon Rank Sum (WRS) Test

WRS [79] test is a non-parametric test commonly used for pairwise DGE testing. The test is formulated as H_0 , the distributions of the gene vector under two dose groups are equal versus H_a : the distributions are not equal. The test involves the calculation of the U statistic, which for large samples is approximately normally distributed. Since this is a pairwise test, a union is taken over all the genes found to be DE in each of the pairwise tests. The computations are carried out using the *wilcox.test* function in R package *stats* and multiplicity is controlled using FDR correction.

2.2.5.5 ANOVA

Analysis of variance (ANOVA) [80] is very commonly used for testing the differences among means in multiple groups. For a fixed gene j , it is assumed that the observed gene vector $y_{k,i,j}$ for cell i is grouped by dose. Assuming that $Y_{k,i,j} \sim \text{Normal}(\mu_{k,j}, \sigma_j^2)$, ANOVA aims to test the null hypothesis $H_0 : \mu_{1,j} = \mu_{2,j} = \dots \mu_{K,j} = \mu_j$ versus $H_a : \mu_{k,j}, i = 1, \dots, n; j = 1, \dots, p; k = 1, \dots, K$ is different for at least one k . The test statistic is computed using the *aov* function in R package *stats* and it follows a F-distribution with $(K - 1)$ and $(n - K)$ degrees of freedom. Multiplicity is controlled by applying FDR correction on the obtained p-values.

2.2.5.6 Kruskal-Wallis (KW) Test

KW [81] test extends the WRS test for multiple groups. It is also a non-parametric extension of the ANOVA test. The test is formulated as; H_0 , the distributions of the gene vector under K dose groups are equal versus H_a : the distributions are not equal. The computation of the KW test statistic is carried out using the *kruskal.test* function in R- package *stats* and it asymptotically follows a χ^2 distribution with $K - 1$ degrees of freedom. Multiplicity is controlled by applying FDR correction on the obtained p-values.

2.2.6 Benchmarking and sensitivity analyses

Benchmarking of DE test methods was performed on simulated datasets based on initial parameters derived from real dose-response snRNAseq data. The probability of differential expression was set to 10% with a 50% probability of being down-regulated, equally distributed among the dose-response models in Table 1. Batch parameters were used to include sample variation associated with data obtained from 3 individuals in each dose group. A total of 5,000 genes were simulated for 4,500 cells (500 per dose group) using the same doses as the real dataset. Sensitivity analyses varied each of the following parameters according

to values is supplementary Table 1: cell abundance equally distributed among dose groups, varying cell numbers in each dose group, percent DE genes, proportion of downregulated DE genes, fold-change location or scale, and dropout rate. Each simulation was replicated 10 times using a different initial seed. Method concordance was determined as area under the concordance curve (AUCC) for the top 100- or 500-ranked genes in simulated and real datasets, respectively, as previously described [82].

2.3 Results

For benchmarking of DGEA methods, a ground truth is required. Existing simulation tools such as PowSimR, SymSim, SPsimSeq and Splatter are commonly used for power analyses, evaluating DE analysis methods, and testing cell clustering strategies [75, 83, 84, 85]. Tools such as SymSim and Splatter are also capable of simulating cell trajectories and model differentiation processes. Trajectories which exhibit non-linear changes over time or across different developmental stages are not unlike dose-response effects which change over a continuum of doses. However, dose-responsive changes commonly follow defined trajectories such as Hill, exponential, power, and linear models [28]. To simulate dose-response scRNAseq data we developed a wrapper for the Splatter scRNAseq data simulation tool named SplattDR. SplattDR modified the Splatter grouped data simulation strategy by adjusting counts from means defined by one of the dose-response functions outlined in the Materials and Methods.

To demonstrate the modeling capability of SplattDR, 10000 gene expression responses were simulated with a 10% probability of being differentially expressed, equally distributed across the dose-response models. Parameters used in Splatter were initially estimated from our experimental single nuclei RNAseq (snRNAseq) dose-response dataset. The simulated data compared to the experimental data showed the relationship between the mean expression, percentage of zeroes, and mean variance were consistent (Figures 2.2A, B). Estimation of the normalized root mean square deviation (NRMSD) from a curve fit to the experimental

data indicated excellent concordance.

The distribution of $\log(\text{fold-changes})$ between vehicle (dose 0) and the highest simulated dose (dose 9; 30 $\mu\text{g/kg}$) showed a more even distribution within a similar range compared to experimental data which was skewed towards induction (Figure 2.2C).

However, the gene induction skew was captured by modulating the parameters affecting the probability of differential expression and the proportion of differentially repressed genes (Supplementary Figure A.1). Principal components analysis (PCA) of the simulated data clearly showed the dose-dependent characteristics of scRNAseq data with distinct clusters increasing in separation with increasing dose (Figure 2.2D) which was also resolved by PCA within the experimental data (Supplementary Figure A.2).

To our knowledge, no other published in-vivo dose-response scRNAseq datasets are available limiting the number of datasets to estimate initial parameters for simulation to date. To investigate whether existing datasets generated using a different study design (e.g. whole cells or different tissue source) could be used to derive initial parameters, we also simulated 10 000 genes starting with parameters estimated from (i) a two-dose liver snRNAseq (GSE148339), (ii) whole cell liver scRNAseq (GSE129516) and (iii) peripheral blood mononuclear cells (PBMC; GSE108313) datasets. When compared to a model fit for experimental data to determine the relation between mean expression and percent zeroes or mean variance, the NRMSD for data simulated from these datasets were between 1 and 10% with data simulated from whole cell data differing the most from the model fit (Figure 2.2E). We then explored whether parameters estimated from distinct cell types could replicate the characteristics of that same cell type (Figure 2.2F). Not surprisingly, using initial parameters derived from individual cell types in the experimental dose-response data had lower NRMSD than those derived from the whole cell dataset. Notably, when data derived from a lower abundant cell sub-type was used to estimate starting parameters, the dose-response characteristics for that cell subtype was also poorly modeled (Figures 2.2E, 2.2F)

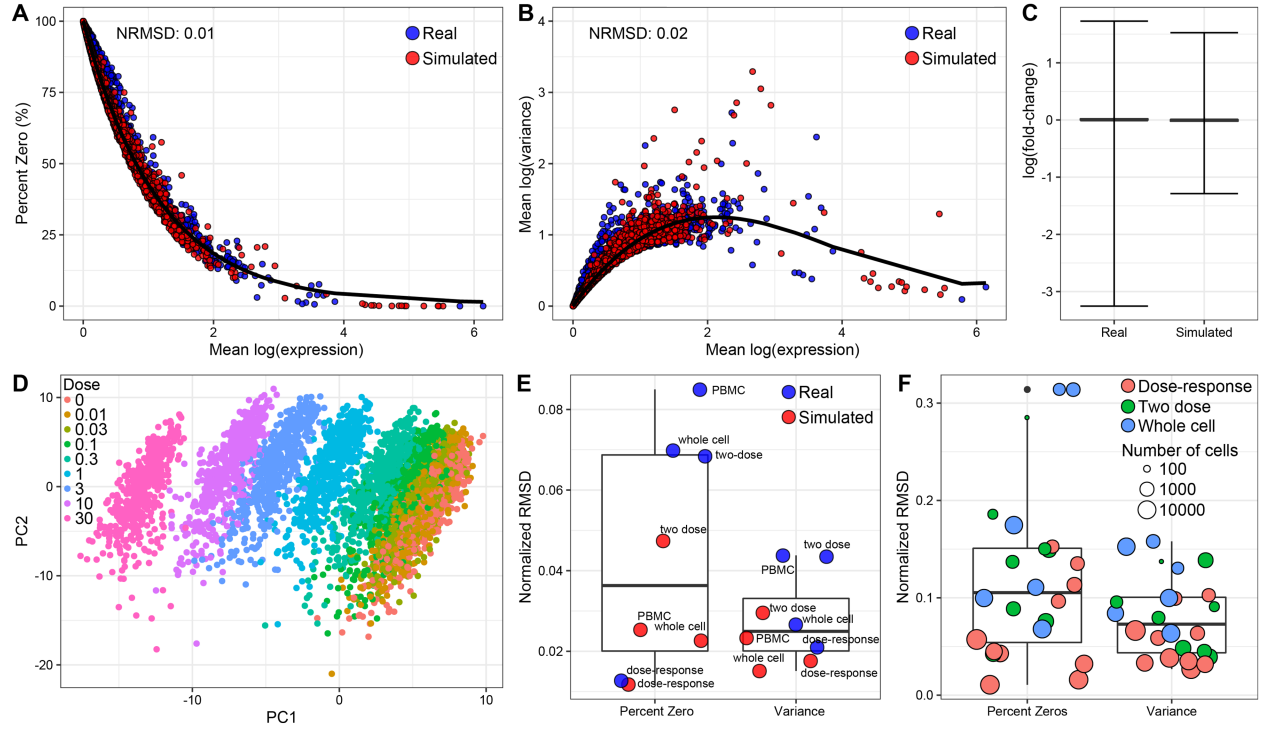


Figure 2.2 Comparison of simulated and real dose-response data. **(A)** Relationship between gene-wise mean expression and percent zeroes for simulated and real dose-response data. Simulation data consisted of 10,000 genes and 9 dose groups based on parameters derived from experimental dose-response snRNAseq data. Black line represents a fitted model to the experimental data from which the normalized root mean square deviation (NRMSD) of simulated data was determined. **(B)** Relationship between gene-wise mean expression and variance for simulated and experimental data. NMRSD was calculated for simulated data from the fitted model represented as a black line. **(C)** Distribution of log(fold-changes) in experimental and simulated data showing the median and minimum and maximum values. **(D)** Principal components analysis of simulated data colored according to simulated dose groups. **(E)** NMRSD estimated relative to fitted model in A,B for simulated data generated from initial parameters derived from published hepatic scRNAseq (two dose; GSE148339), hepatic whole cell (whole cell; GSE129516), and peripheral blood mononuclear cell (PBMC; GSE108313) datasets. **(F)** NMRSD estimated relative to model fitted to cell-type specific experimental dose-response data when simulated from initial parameters estimated from that same cell type. Box and whisker plots show median NMRSD, 25th and 75th percentiles, and minimum and maximum values.

2.3.1 Performance accuracy of DE test methods

We evaluated the performance of several differential gene expression analysis methods on simulated datasets consisting of nine dose groups of 500 cells each (4500 total) and 5000

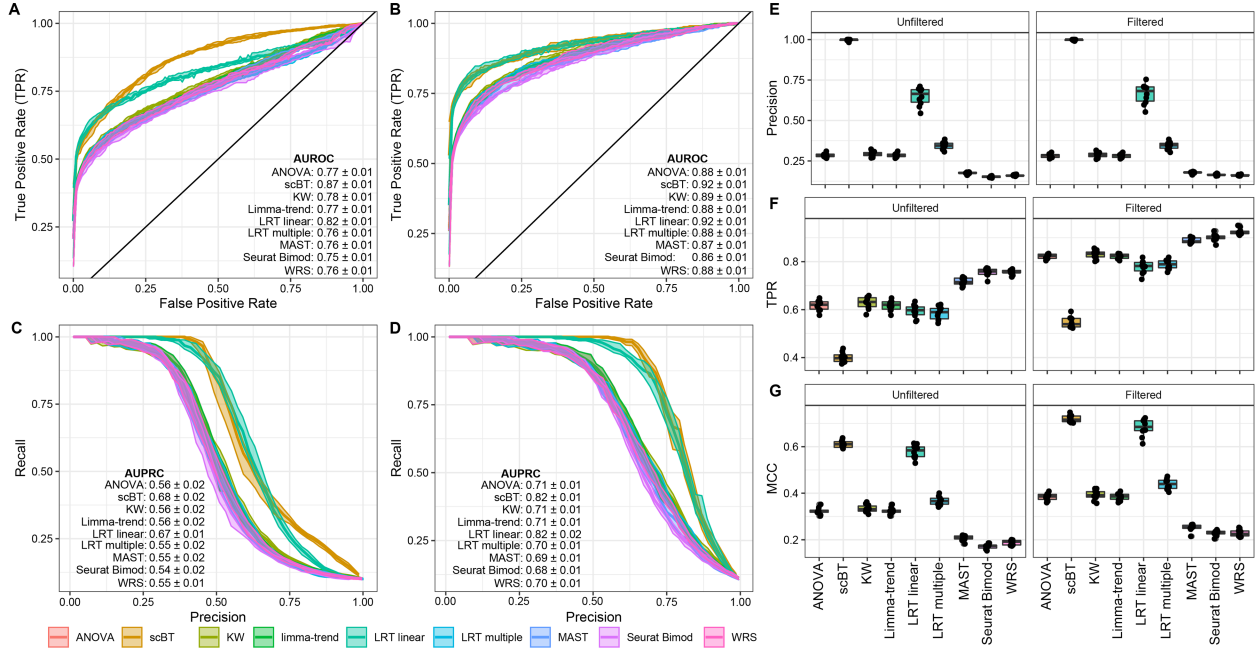


Figure 2.3 Classification performed of DE analysis tests. (A) ROCs estimated from simulated dose-response scRNAseq data for 9 DE test methods including all genes expressed in at least 1 cell (unfiltered). (B) ROCs for 9 DE test methods after filtering simulated dose-response scRNAseq data for genes expressed in only $\geq 5\%$ of cells (low levels) in at least one dose group. (C) Precision-recall curves (PRCs) for 9 DE test methods on unfiltered simulated dose-response scRNAseq data. (D) PRCs for 9 DE test methods on filtered simulated dose-response scRNAseq data. Lines represent the mean values and shaded region reflects the standard deviation for 10 independent simulations. (E) Precision of DE test methods. (F) FPR of DE test methods. (G) MCC for test methods. E,F,G Box and whisker plots median values, 25th and 75th percentiles, and minimum and maximum values for 10 independent simulations. Points reflects values for each independent simulation. Panels display comparisons of unfiltered and filtered datasets.

genes with a 10% probability of being differentially expressed (500 differentially expressed genes). Selection criteria for test inclusion are outlined in the Materials and Methods section and included 9 test methods; ANOVA [80], single-cell Bayes hurdle model test (scBT), Kruskal–Wallis (KW) [81], limma-trend [20, 78], likelihood-ratio test (LRT) linear and multiple, MAST [16], Seurat bimod [14] and WRS [79]. With ground truth from simulated data, the sensitivity, specificity, and precision for each test method was computed. Area under the receiver-operating characteristic curve (AUROC) was used to measure test performance for correctly classified differentially expressed genes.

In unfiltered data, AUROC scores showed similar performance for most tests except scBT which had the largest AUROC among all test methods (Figure 2.3A). To account for the inherent class imbalance between differentially expressed and non-differentially expressed classes the area under the precision-recall curves (AUPRC) was also calculated. Similar to AUROCs, AUPRCs identified scBT as the best performing test (Figure 2.3C). In most standard differential expression testing pipelines genes expressed at low levels are removed to minimize false detection rates. Following filtering of genes expressed in $\geq 5\%$ of cells in any dose group, scBT was consistently ranked as the best test based on AUROC and AUPRC scores. The performance of LRT linear test also improved, with comparable AUROC and AUPRC scores relative to scBT, suggesting LRT linear is poorly suited for genes expressed at low levels (Figures 2.3B–D).

AUROC and AUPRC reflect the performance of each test method with varying significance (i.e. P-value) thresholds. In the standard pipeline a fixed threshold is used, typically a P-value ≥ 0.05 after adjustment for multiple hypothesis testing (i.e. Bonferroni correction). For each method except scBT, the performance at an adjusted P-value ≥ 0.05 significance criteria was evaluated. In scBT analysis, a gene was considered differentially expressed when the estimated posterior probabilities of the null hypothesis, $p(H_0, j|D_j)$, was less than α , where the value was chosen to achieve a target FDR of 0.05. scBT significantly outperformed all other tests in precision rates irrespective of low expression filtering (Figure 2.3E). However, scBT was less effective in identifying true positives (Figures 3F). Applying the filtering criteria improved the recall rates, but the precision rates remain largely unchanged (Figure 2.3E, 2.3F). Test method classification performance scores were estimated as the Matthews correlation coefficient (MCC) which is well suited for unbalanced data [86]. We see that the scBT and LRT linear tests performed best for this metric on both unfiltered and filtered data (Figure 2.3G).

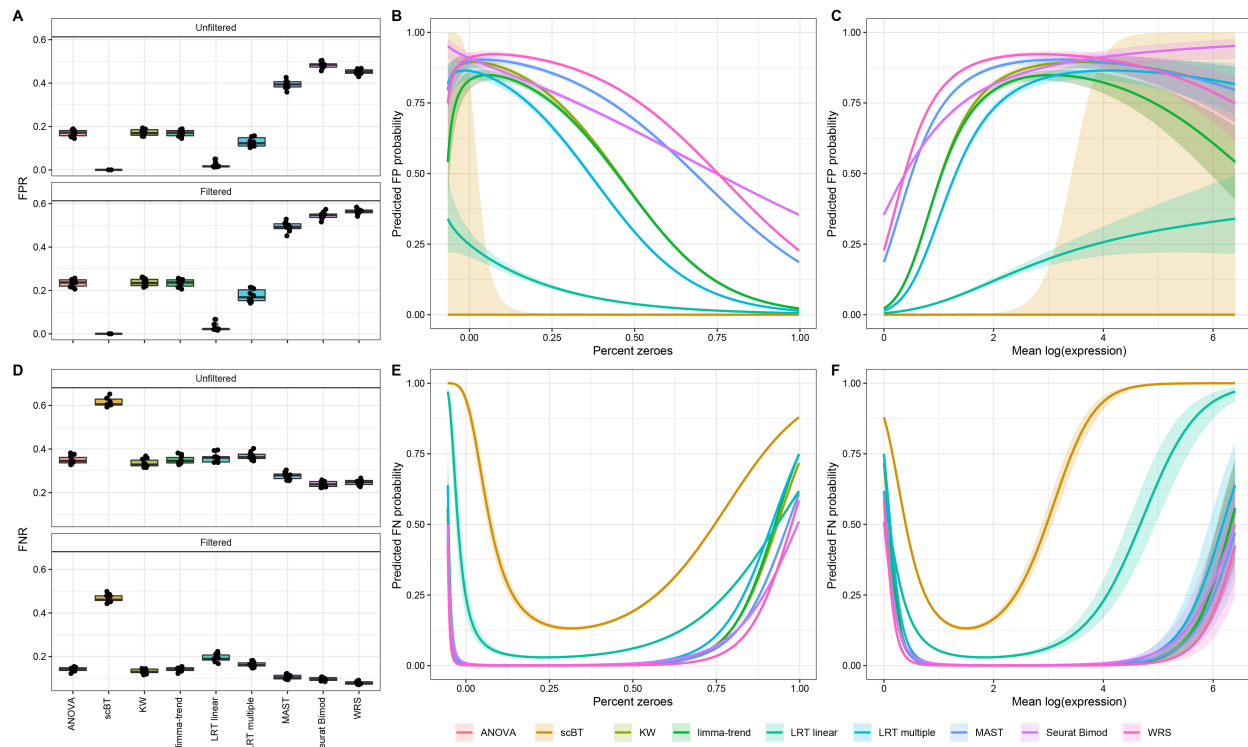


Figure 2.4 Evaluation of Type I and II error control. **(A)** False positive rate (FPR) of 9 differential expression test methods estimated from negative control (0% DE genes) simulated dose-response scRNAseq data including all genes expressed in at least 1 cell (unfiltered) and genes expressed in only $\geq 5\%$ of cells in at least one dose group (filtered). **(B,C)** Logistic regression models were fitted to negative control data to predict the probability of false positive identification using percent zeroes and mean expression as covariates. Lines represent the predicted probability of false positive classification with the shaded region representing the 95% confidence interval. **(D)** False negative rate (FNR) of 9 differential expression test methods estimated from positive control (100% DE genes) simulated dose-response scRNAseq data including unfiltered and filtered datasets. **(E,F)** Logistic regression models were fit to positive control data. Lines represent predicted probability of false negative classification with shaded region representing the 95% confidence interval.

2.3.2 Type I error control and power

To investigate test performance in controlling type I errors (false positives), DGEA methods on simulated datasets were examined with 0% DE genes (i.e. negative control). Using the threshold for the computed posterior null probabilities, scBT identified only one false positive gene in 2 of 10 simulations (Figure 2.4 A). ANOVA, scBT, KW, limma-trend and LRT linear

had false positive rates (FPRs) below 3% indicating better performance compared to two group tests. After filtering for genes with low expression levels, scBT still correctly identified all the non-differentially expressed genes and was the best performing test. These are the same tests that had a better FPR control in initial simulations (Figure 2.4). To explore whether mean expression or percentage of zeroes influenced type I error rates, a logistic regression model was fit to negative control data. We predicted the probability for each gene to be identified as differentially expressed in the negative control data. While the curve for scBT is missing since few false positives were identified, the predicted FPR for all the other tests except LRT linear were also high for highly expressed genes with few zeroes (Figures 2.4B, 2.4C). Next, a positive control dataset with 100% differentially expressed genes was simulated to evaluate test performance for detecting true positives. All tests except scBT exhibited a false negative rate (FNR) $\geq 40\%$ (Figure 2.4D). The best performing tests for FNR also had high FPR. Logistic model regression fitting for false negative classification of genes shows that the false negative rates were highest when the mean expression was either too high or too low for all tests (Figures 2.4E, 2.4F).

2.3.3 Parameter Sensitivity Analysis

Experimental scRNAseq datasets will vary between cell types, cell composition, and responses depending on the target tissue, treatment, number of cells sequenced, and more. For example, some distinct cell types are very abundant (e.g. hepatocytes), with others present at lower levels (e.g. portal fibroblasts) in hepatic scRNAseq datasets. Moreover, treatments such as exposure to a xenobiotic, can elicit dose-dependent changes in relative proportions of cell types such as the infiltration of immune cells (26). We investigated the impact by changing cell abundance from 25 to 2000 cells per dose group and observed an increase in the false positive rate (FPR) when increasing the number of cells. The scBT and LRT linear tests were less sensitive to an increase in the FPR as cell abundance increased while the total positive rates (TPR + FPR) increased with cell abundance for all methods.

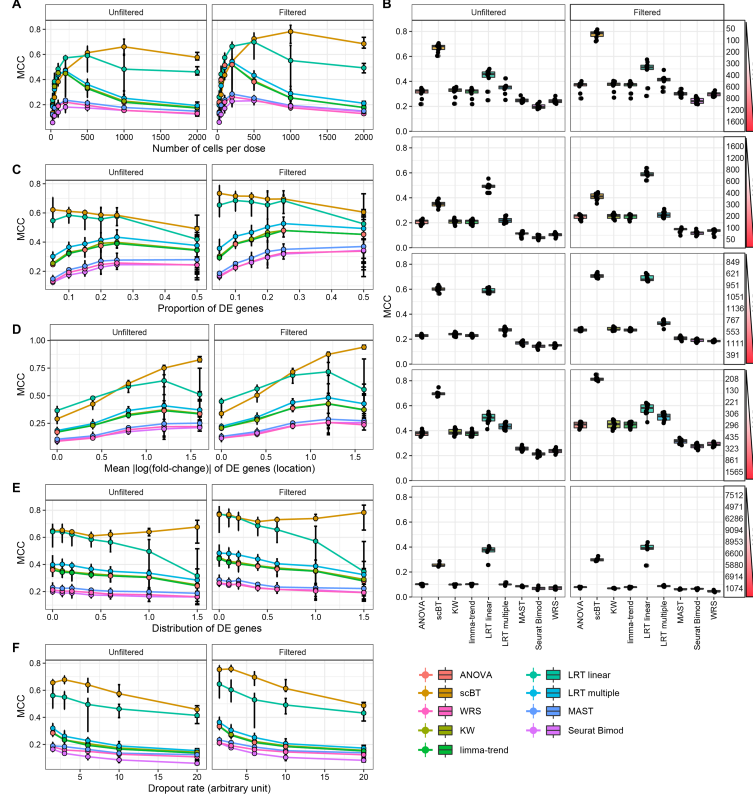


Figure 2.5 Matthews correlation coefficient (MCC) from sensitivity analyses of differential expression test methods. **(A)** MCC for 9 DGEA test methods determined from simulated dose response data with varying number of cells per dose group. Simulations consisted of 5,000 genes with a probability of differential expression of 10% and 9 dose groups. **(B)** MCC for simulated data varying the cells numbers by dose group. The number of cells in each of the 9 doses groups is shown on the right. **(C)** MCC for varying proportion of differentially expressed genes. **(D)** MCC when varying the mean fold-change (location) of repressed differentially expressed genes. **(E)** MCC for varying distribution of fold-change (scale) of differentially expressed genes. **(F)** MCC for varying dropout rates calculated as in Table S3. Points represent median and error bars represent minimum to maximum values. Boxplots represent median, 25th to 75th percentile, and minimum to maximum values. Each analysis consisted of 10 replicate datasets including all genes expressed in at least 1 cell (unfiltered) and genes expressed in $\geq 5\%$ of cells in at least one dose group (filtered).

Although all tests exhibited comparable performance at low cell numbers (≥ 500), as cell numbers increased scBT outperformed all other tests in both precision and MCC score (Figures 2.5 A) Comparison of AUROCs and AUPRCs across cell numbers showed that ANOVA, KW, limma-trend, and LRT linear tests performed best for a small number of cells, but the increase in AUROC was steeper for scBT.

It was also evident from the experimental snRNAseq dataset that the number of cells per dose group was not fixed. We evaluated the performance of the test methods when the number of cells dose-dependently increased or decreased, and when the number of cells per dose group were taken from experimental data. Notably, while scBT had the best MCC for increasing number of cells per dose, LRT linear performed better than scBT when the number of cells decreased before and after filtering for genes expressed at low levels (Figure 2.5B). The shift in MCC between increasing and decreasing cell numbers for scBT appears to be driven by a concomitant decrease in TPR and increase in FNR.

Increasing the proportion of differentially expressed genes led to an improvement in MCC except for scBT and LRT linear, though these tests maintained the top MCC scores as well as AUROC and AUPRC (Figure 2.5C). As the magnitude of the effect increased, LRT linear performed best at the low end while scBT exhibited the greatest improvement in MCC (Figure 2.5D). Conversely, while the MCC decreased for most tests when modulating the fold-change scale of differentially expressed genes, scBT improved and was more stable (Figure 2.5 E). As the proportion of unexpressed genes increased, the FPR increased with precision decreasing for all tests. However, scBT was least affected, and maintained the highest MCC among all tests (Figure 2.5F).

2.3.4 Test method agreement

To assess agreement between tests, the area under the concordance curve (AUCC) for each pair of tests for the top 100 genes ranked by adjusted P-value was calculated as previously described [32, 82]. All methods showed excellent concordance ($AUCC \geq 0.77$) with LRT

linear showing the poorest consistency compared to all other tests while the limma-trend and ANOVA tests showed perfect agreement with an AUCC of 1 (Supplementary Figure A.3). Pairwise differential gene expression comparisons between Seurat Bimod, MAST and WRS had $AUCC > 0.95AUCCs$ while the multiple group tests ANOVA, LRT multiple, KW, and scBT clustered together with AUCC ranging between 0.9 and 1. In the absence of nuisance covariates, MAST and Seurat Bimod provided similar results, as expected given their similar mixture normal model structure. Likewise for ANOVA and limma-trend, both of which rely on normality assumptions for testing differential gene expression.

2.4 Real dose-response dataset DE analysis

Without ground truth for experimental data, the performance of the differential expression test methods was examined by first evaluating the agreement for each identified cell type (Figures 2.6). Genes in the experimental dataset were considered differentially expressed when expressed in $\geq 5\%$ of cells in at least one dose group and had a $|\text{fold-change}| \geq 1.5$. In hepatocytes, the most abundant cell type, fewer than 5 genes were not detected in all test methods, with the majority missed by the WRS test (Figure 2.6A). Upon closer examination, those genes were not expressed in control hepatocytes. Not surprisingly, for all cell types, the largest intersection was between all tests indicating strong agreement within all test methods. Only a few tests identified a subset of unique genes as differentially expressed, which accounted for a very small fraction. For example, LRT linear identified 12 unique differentially expressed genes in portal fibroblasts, one of the least abundant cell types (Figure 2.6B). LRT linear was the best performing test for low cell numbers indicating that the 12 unique differentially expressed genes may in fact be true positives. Consistent with simulations of varying cell numbers (Figure 2.6A), 24 genes were not identified as differentially expressed by the scBT method for stellate cells which exhibit a dose-dependent decrease in numbers (Figures 2.6C, D). Although scBT outperformed other tests in most scenarios, it under performed in this scenario. Nevertheless, when ranking genes by significance level (i.e.

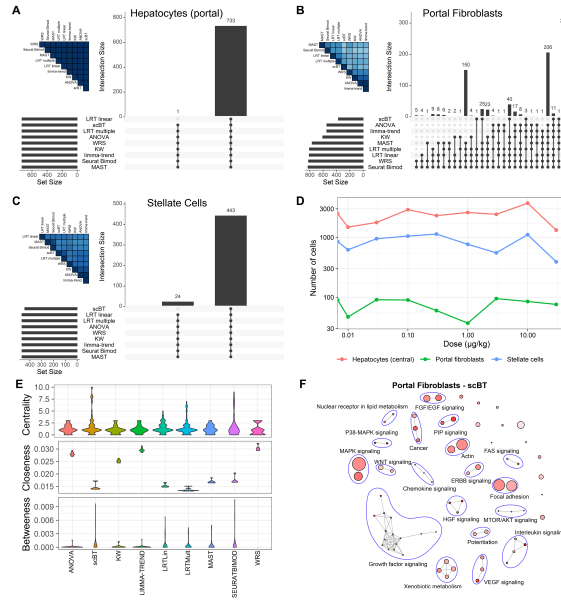


Figure 2.6 Agreement of differential expression test methods on experimental dose-response data. **(A)** Upset plot showing the intersection size of genes identified as differentially expressed by 9 different test methods in hepatocytes from the portal region of the liver lobule. **(B)** Intersect of differentially expressed genes in portal fibroblasts. **(C)** Intersect size in hepatic stellate cells. Vertical bars represent the intersect size for test methods denoted by a black dot. Horizontal bars show the total number of differentially expressed genes identified within each test (set sizes). Only intersects for which genes were identified are shown. Genes were considered differentially expressed when (i) expressed in $> 5\%$ of cells within any given dose group and (ii) exhibit a $|\text{fold-change}| \geq 1.5$. A heatmap in the upper left corner of each panel shows the pairwise AUCC comparisons for the 500 lowest p-values. **(D)** Relative proportion of cell types identified in each dose group of the real dataset for the cell types in A,B,C. Experimental snRNAseq data was obtained from male mice gavaged with sesame oil vehicle (vehicle control) or 0.01 – 30 $\mu\text{g/kg}$ TCDD every 4 days for 28 days. **(E)** Graph metrics for gene set enrichment analysis of portal fibroblasts grouped by similarity in gene membership. Violin plots show distribution of node-wise values for each test method. **(F)** Network visualization of significantly enriched (adjusted p-value ≤ 0.05) gene sets using the Bayes factor ranked genes of portal fibroblasts. Groups of ≥ 2 nodes were manually annotated following commonality in the gene set names. Each node represents a gene set with the size of the node representing the number of genes in a gene set, and edges connect nodes with $\geq 50\%$ overlap.

P-values), AUCC were high for all pairwise comparisons.

To explore the biological insight gained by using the test methods, gene set enrichment analysis was performed by ranking genes following significance values (adjusted P-value or Bayes factor) on gene sets from BIOCARTA, KEGG, PANTHER and WIKIPATHWAYS. Gene sets were grouped based on their similarity in gene membership into a network for which centrality measures can be estimated. An examination of portal fibroblasts, which exhibited the most disagreement among test methods (Figure 2.6B), showed that multiple group test methods, particularly scBT had improved centrality metrics (centrality – number of edges; closeness – steps required to access other nodes; and betweenness – number of paths that go through a node) (Figure 2.6E). Visualization of significantly enriched terms identified enriched functions associated with growth factor and immune cell signaling in addition to expected terms such as xenobiotic metabolism and nuclear receptors involved in lipid metabolism (Figure 2.6F). Alternatively, WRS which did not find as many connected groups of functions, was largely limited to those identified by scBT except for the hormone signaling and tryptophan metabolism clusters. While there is no ground truth from real data, greater agreement between similar gene sets from disparate sources suggests that multiple group tests such as scBT provide more reliable findings [1]. However, all the test methods produce comparable gene set enrichment results as expected since the most robust changes were identified by all the test methods.

2.5 Discussion

The goal of this study was to compare the performance of newly developed DGEA test methods for dose-response experiments to existing analysis methods. Using simulated data to generate ground truth, we evaluated the performance of nine differential expression testing methods which were broadly classified as either fit-for-purpose, multiple group, or two group tests. Criteria for test method selection was based on previous benchmarking efforts for two group study designs identifying MAST, limma-trend, WRS, and t-test as the best performers

[32, 87]. ANOVA and KW tests were also included for evaluating multiple group comparisons, and Seurat Bimod, for having the same modelling framework as scBT, LRT multiple and LRT linear tests. The test methods were ranked from best to worse (1-9) based on type I error rate, type II error rate, MCC, AUROC and AUPRC (Figure 7, Supplementary Table S4).

While several scRNAseq tools have been developed [75, 83, 84, 85], none are developed to simulate dose-response models commonly identified in toxicological and pharmacological datasets [28, 88]. Our SplattDR wrapper for the Splatter package (28) was able to show that simulated data can effectively emulate key experimental scRNAseq data characteristics when simulation parameters were estimated from various Unique Molecular Identifier (UMI)-based datasets. In agreement with a previous report, technical and biological factors, such as cell type, does appear to influence gene dropout rates (18). We primarily focused on 10× Genomics UMI data given the unavailability of real experimental dose-response data generated using other platforms.

Overall, test method performance was consistent with their intended application. For example, fit-for-purpose tests scBT and LRT linear consistently ranked higher followed by multiple groups tests such as KW and LRT multiple. scBT exhibited the best overall performance with excellent FPR control and top ranked MCC while LRT linear struck a balance between type I and type II error rates. The scBT results are not surprising as Bayes factor-based tests have proven to be conservative and consequently more appropriate when false positives are of concern [69, 70]. In the context of investigating chemical or drug MoAs, false positives have the potential to lead to wasted effort and resources in attempts to validation and support findings [89]. Moreover, when assessing a large number of genes, a 5% FP rate ($P - value \geq 0.05$) can result in hundreds of FPs that skew MoA classifications [67].

A single test method was not expected to outperform all other tests under all conditions as previously demonstrated when comparing pairwise testing [29, 32, 87]. Therefore, we assessed the strengths and limitations of each test method by varying parameters likely to change

within and across various experimental datasets. The number and relative abundance of cell types is known to be affected by disease or treatment, and the distribution of differential expression influenced by the chemical, drug, or food contaminant being evaluated [63, 73]. scBT consistently ranked at the top under most scenarios, particularly when the mean and standard deviation of the fold-change for differentially expressed genes varied. However, scBT under performed in MCC when the number of cells decrease in a dose-dependent manner which would be expected in treatments which alter cell population sizes (e.g. inflammation). Under these circumstances LRT linear outperformed all other tests with scBT performing similar to the other test methods as evident when 24 differentially expressed genes were not identified by scBT within experimental data for stellate cells which experienced a dose-dependent decrease in relative abundance following TCDD treatment. Although excluding genes expressed at low levels generally improved the performance of all test methods, the comparative performance of test methods did not significantly change in most cases. We did not have access to experimental scRNAseq dose-response data, however, we expect that the scBT would perform equally well as with experimental snRNAseq data as the elevated number of zeroes are common to both types of data. Major differences between these types of data are (i) biases in gene detection and (ii) overall counts [73]. Given the higher overall counts in scRNAseq data, test method such as scBT may even perform better.

DGEA provides biological information regarding the effects of exposure to chemicals, drugs, and food contaminants. As expected, gene set enrichment analyses did not dramatically differ in the enriched pathways which are driven by the most robust responses such as xenobiotic metabolism. However, when integrating gene sets from disparate sources we found gene sets that partially overlap in gene membership were consistently identified by multiple group test methods. For example, several gene sets related to growth factors and cell proliferation were identified by scBT but not WRS. Portal fibroblasts are implicated in proliferation of cholangiocytes and the secretion of growth factors during development. Enrichment of these terms suggests a functional role consistent with the induction of bile

duct proliferation by TCDD (45,46). In contrast, WRS identified enrichment associated with tryptophan as well as oxytocin/thyrotropin-releasing-hormone pathways which has not been linked to the effects of TCDD on portal fibroblasts. Although ground truth for the complete experimental dataset is not available, the use of test methods such as scBT reduce experimental noise to identify leads warranting further analysis.

2.6 Conclusion

Collectively, our findings suggest that scBT and LRT linear fit-for-purpose tests are better suited for the differential expression analysis of dose-response studies and when false positives are of greater concern than false negatives. Moreover, consistent with previous benchmarking efforts, we show that common non-parametric tests such as KW out-perform test methods developed for scRNAseq data when the study involves comparisons between multiple groups. Ultimately, each test method performs optimally under diverse scenarios. While the importance of controlling type I error rates is acknowledged, a balance must be struck with type II error rates. The tradeoff should be determined based on the individual research question being investigated. It may even be reasonable to apply different test methods to distinct cell types based on dropout rates, cell abundance, and changes in relative cell proportions given the strengths and weaknesses of each test method.

2.7 Acknowledgements

This chapter is based on my published paper [1]. I would like to thank the joint first author Dr Rance Nault, and co-authors Dr Samiran Sinha, Dr Tapabrata Maiti, Dr Sudin Bhattacharya, Jack Dodson and Dr Tim Zacharewski for their support and advice. This work was funded by National Human Genome Research Institute [R21 HG010789]; National Institutes of Environmental Health Sciences Superfund Research Program [P42 ES004911] and NSF [DMS 1945824].

CHAPTER 3

SEMIPARAMETRIC DOSE RESPONSE CURVE ESTIMATION FOR SINGLE CELL DOSE RESPONSE EXPERIMENTS

3.1 Single Cell Dose Response Experiments

Gene expression profiling of single cells has led to unprecedented progress in understanding normal physiology, disease progression and developmental processes. In contrast to bulk RNA-sequencing (RNAseq), gene expression profiling of single-cell allows investigation of cell-specific heterogeneity and can be used to assess changes in response to drugs and chemicals for individual cell populations. Dose-dependent gene expression profiling (aka genomic dose-response studies (GDRS)) of drugs and chemicals has been proposed as an alternative test to rodent bioassays to assess human health risks and application of single-cell RNA sequencing (scRNAseq) for the evaluation of chemicals, drugs, and food contaminants presents the opportunity of accounting for cellular heterogeneity in pharmacological and toxicological responses. Although dose response modelling of bulk RNA-seq datasets have been long-used for generating quantitative estimates of risks associated with such responses, no statistical study yet exists for the estimation of dose response curves for scRNA-seq datasets.

Unlike microarray and bulk RNA-seq datasets that record gene expression measurements averaged over many cells, scRNAseq allows gene expression measurement on a cellular level hence enabling biological investigation at the cellular level. However, despite many improvements in high throughput sequencing, various technical factors including cell-cycle heterogeneity, library size differences, amplification bias, and low RNA capture per cell lead to high noise in scRNA-seq experiments. Recent technologies are capable of sequencing millions of cells, but often generate highly sparse expression datasets due to shallow sequencing. The presence of these issues result in substantial noise that often obscures the true biological signal and renders unsuitable, the application of traditional statistical models for analysis of

single cell datasets. Several statistical models have been proposed for the analysis of single cell datasets, with most focusing on zero-inflated count data distributions like Poisson or negative binomial, to account for the overdispersion. Continuous hurdle models like MAST have also been proposed that treat the zero sampling process to be completely different from the sampling process of true biological expression, which is generally positive. Several statistical studies have been designed to investigate problems related to differential gene expression, cell-type clustering, denoising or imputation, gene regulatory network construction, trajectory inference and other statistical problems studied earlier in the context of bulk and microarray datasets. However, no rigorous statistical framework has yet been designed for dose response modelling for scRNA-seq datasets that are invaluable for deriving cell-type specific efficacy and/or safety margins such as effective dose and the point of departure (POD) of several toxicological/ pharmacological responses.

Regulatory communities have suggested that acute (< 14 days) and sub-acute (14–28 days) transcriptomic studies as viable alternative to the current standard 2-year rodent bioassay that significantly reduces the time and resources needed to assess risk. Furthermore, single cell transcriptomic datasets could bolster such investigations by identifying cell-specific dose-dependent responses indicative of an adverse event. The U.S. National Toxicology Program (NTP) recently reported a robust DGEA approach is essential to deriving biologically relevant PODs. In characterizing biologic and public health significance, and the need for possible regulatory interventions, it is important to efficiently estimate dose response functions while accounting for cell-specific heterogeneity and known experimental confounders. Motivated by the lack of a single framework for single cell dose-response estimation and trend testing, while accounting for covariates this chapter proposes a complete statistical framework for the same.

3.1.1 Motivating experimental study and hypothesis of interest

In this work we analyze a unique in vivo dose response hepatic scRNAseq dataset consisting of 9 dose groups with 3 biological replicates for 11 distinct liver cell types for greater than 100K cells. The 9 dose groups represent 9 doses of varying levels of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD). The persistent organic pollutant and potent agonist of the aryl hydrocarbon receptor (AHR), TCDD induces hepatic lipid accumulation (steatosis) that progresses to steatohepatitis with fibrosis. In humans, TCDD and related compounds are associated with dyslipidemia and inflammation and in mice, AHR activation by TCDD elicits cell-specific and spatially resolved histological and gene expression responses. Our interest lies in estimation of cell-type specific dose response curves for genes of interest in order to derive finer point of departures (PODs) for use towards informed decision-making on the safety and toxicity of drugs and chemicals. We hypothesize that some genes of interest would exhibit different POD's for different cell-types and our interest not only lies in estimating cell-specific DR curves but also in testing whether there exists a difference in cell-type specific POD's. This would be particularly hard if the genes exhibit slight changes in POD's over a handful of cell-populations while exhibiting the same POD over others. In those cases we would like to investigate how the POD estimates change in comparison to existing ones established via bulk and microarray based DR studies. Further, it could be possible that celltypes exhibit completely different DR curve shapes for some genes whereas there are slight shifts and slope changes for cell-type specific DR curves for other genes. It is essential to allow the DR curve to change according to cell-types for a particular gene in order to capture these varying curve shapes. It is also essential to test whether these estimated changes between celltypes are statistically significant, in order to ensure accurate POD estimation. Therefore in this work we propose 1) a semi-parametric regression model for flexible modelling of dose response curves, while accounting for high percentage of zeroes and effects of additional covariates 2) a scalable and computationally efficient MM algorithm for the estimation of regression parameters, 3) an approach to incorporate monotonicity assumptions, if supported by prior

beliefs, within the estimation paradigm, 4) tests of hypothesis for testing the presence of a dose response trend and whether the trend varies for different cell-types and 5) estimation of cell-type aware POD's.

3.1.2 Literature on dose response curve estimation

The primary interest in dose response studies lies in assessing the association between a continuous predictor, D , and a response variable, Y , while adjusting for covariates, $X = (X_1, \dots, X_q)^\top$. There are a number of papers in the existing literature on dose response curve estimation, but none that solves the problem underlying our motivating examples. Most of the existing literature focusses on analyzing data from bulk RNA-seq datasets, microarrays or traditional environmental toxicity studies that focuses on continuous or dichotomous endpoints. Let us denote a sample of n independent and identically distributed (IID) pairs of variables, $\mathcal{B} = \{(Z_i, Y_i)\}, i = 1, \dots, n$ as the data. Z_i can be further segmented as $Z_i = \{(D_i, X_i)\}$ where D_i represents the dose vector and X_i represents the vector of other confounding covariates such as age, gender etc. The main interest of dose response studies lies in modelling $E(Y_i)$ as a function of the dose vector; ie. $E(Y_i|D_i) = \zeta(D_i, \theta)$, where θ is the vector of unknown parameters.

In general, the methods for estimating dose response can be classified as parametric or nonparametric. Parametric methods assume a model $\zeta_i(D, \theta)$ for the DR curve where θ is the vector of unknown parameters. Considerable amount of statistical methodology exists in parametric modelling of dose-response studies, employing parametric models, such as Logistic, Exponential, and Gompertz as well as others (Holland-Letz and Kopp-Schneider 2015) and the US EPA (US EPA 2012) gives guidelines on how to employ these methods. These models are generally fit using standard techniques such as maximum likelihood (ML) or restricted maximum likelihood(RML) and their functional forms are monotonic allowing for tractable estimation of relevant POD's. It is well known that if a parametric model is correctly specified then the ML/RML estimators are efficient. However, in many cases,

it is difficult to correctly specify the parametric form of the dose–response curve because the biological mechanism of drug action or toxicity may be complex and the form of the dose–response curve is unknown apriori. When the parametric model is misspecified, the corresponding curve estimate may be severely biased. In addition fitting of extremely flexible polynomial models with high orders, to match difficult non-monotonic curve shapes can lead to severe overfitting. The standard workaround is to fit multiple parametric models, each of which may fit the data reasonably, but produce a range of POD estimates that accounts for model uncertainty in the estimation process. Accounting for this uncertainty using model averaging (MA) has received recent attention in the literature. However, despite simulation studies suggesting that model-averaged estimates provide POD estimates that have low bias and nominal coverage properties, these estimates may fail to adequately describe the true uncertainty for DR curves on the edge of the MA model space, hence resulting in inaccurate inference. Further, as the MA results are available on limited number of parametric forms, models that make fewer assumptions on parametric forms may allow for better estimation of the dose response curves.

To enhance the robustness of the estimation of the dose–response curve, many nonparametric methods have been proposed. [41] proposed MLE estimation procedures under the assumption that the dose response curve is sigmoid and non-decreasing. [42] proposed a method for non-parametric estimation of the dose response curve under monotonicity constraints using B-splines regression. [43] used kernel estimators to obtain estimates for the dose–response curve under general shape restrictions. In the Bayesian setup, several approaches have been proposed for non-parametric estimation of dose response curves under monotonicity constraints. From a Bayesian perspective, one specifies a suitable prior on the regression function that induces monotonicity and then inference is based on the posterior distribution. [44] employed an additive model with a prior imposed on the slope of the piecewise-linear functions. [45] adopted mixture modelling of shifted and scaled probability distribution functions. [46] propose a Gaussian process model with a posterior projection

approach for shape-constrained curves. In contrast to the parametric methods, all the non-parametric approaches are flexible and learn the shape of the dose response curves based on the data. To our knowledge none of the existing literature has considered building a semi-parametric regression model for both unconstrained and monotonicity constrained estimation of dose response curves for single cell experiments while accounting for technical confounders and whose development is the primary goal of this chapter.

3.1.3 Literature on MM algorithms

A large number of statistical and machine learning problems require the computation of

$$\min_{\theta \in \Theta} \mathcal{R}(\theta, \mathcal{B}) \quad \text{or} \quad \hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \mathcal{R}(\theta, \mathcal{B}) \quad (3.1)$$

where $\mathcal{R}(\theta, \mathcal{B})$ is a risk function defined over the observed data $\mathcal{B}$ and is dependent on some parameter $\theta \in \times$. Common risk functions that are used in practice are the negative log-likelihood functions, which can be expressed as

$$\mathcal{R}(\theta, \mathcal{B}) = -\frac{1}{n} \sum_{i=1}^n f(z_i; y_i; \theta)$$

where $f(z_i, y_i, \theta)$ is a density function over the support of Z and Y . Simple maximum likelihood estimation problems (or alternatively minimization of the negative log-likelihood) can be solved analytically, but most practical maximum likelihood and least squares estimation problems must be solved numerically. The task of computing 3.1 may be complicated by various factors which include the lack of differentiability of $\mathcal{R}$ or difficulty in obtaining closed-form solutions to the first-order condition equation $\nabla_{\theta} \mathcal{R} = \mathbf{0}$, where ∇_{θ} is the gradient operator with respect to θ , and $\mathbf{0}$ is a zero vector. The double duty MM acronym that stands for majorize in case of minimization problems and minorize for maximization problems provides an unifying algorithm for simplifying the computation of a difficult form of 3.6 via iterative minimization of surrogate functions [90]. Simplification is attained by (a) avoiding computationally expensive inversion of large matrices, (b) linearizing the optimiza-

tion problem, (c) parameter uncoupling, (d) cleverly dealing with equality and inequality constraints or (e) turning a non-differentiable surface into a smooth one.

Let $\Theta^{(r)}$ represent a fixed value of the parameter Θ , and let $h(\Theta|\Theta^{(r)})$ denote a real-valued function of Θ whose form depends on $\Theta^{(r)}$. The function $h(\Theta|\Theta^{(r)})$ is said to majorize a real-valued function $f(\Theta)$ at the point $\Theta^{(r)}$ provided

$$\begin{aligned} h(\Theta|\Theta^{(r)}) &\geq f(\Theta) \quad \text{for all } \Theta \\ h(\Theta^{(r)}|\Theta^{(r)}) &= f(\Theta^{(r)}) \end{aligned} \tag{3.2}$$

Therefore the surface $h(\Theta|\Theta^{(r)})$ lies above the surface $f(\Theta)$ and is tangent to it at the point at $\Theta = \Theta^{(r)}$. The function $h(\Theta^{(r)}|\Theta^{(r)})$ is said to minorize $f(\Theta)$ at $\Theta^{(r)}$ if $-h(\Theta^{(r)}|\Theta^{(r)})$ majorizes $-f(\Theta)$ at $\Theta^{(r)}$. $\Theta^{(r)}$ represents the current iterate in optimizing the surface $f(\Theta)$. In a majorize-minimize MM algorithm, the algorithm minimizes the majorizing function $h(\Theta|\Theta^{(r)})$ rather than the actual function $f(\Theta)$. If $\theta^{(r+1)}$ denotes the minimizer of $h(\Theta^{(r)}|\Theta^{(r)})$, then the MM procedure forces $f(\Theta)$ towards a minimum. The inequality

$$\begin{aligned} f(\Theta^{(r+1)}) &= h(\Theta^{(r+1)}|\Theta^{(r)}) + f(\Theta^{(r+1)}) - h(\Theta^{(r+1)}|\Theta^{(r)}) \\ &\leq h(\Theta^{(r)}|\Theta^{(r)}) + f(\Theta^{(r)}) - h(\Theta^{(r)}|\Theta^{(r)}) \\ &= f(\Theta^{(r)}) \end{aligned} \tag{3.3}$$

follows directly from the fact $h(\Theta^{(r+1)}|\Theta^{(r)}) \leq h(\Theta^{(r)}|\Theta^{(r)})$ and definition 3.2. The descent property 3.3 allows for remarkable numerical stability and the MM algorithm also applies to maximization rather than minimization with straightforward changes.

MM algorithms, which present a generalization of the EM (expectation– maximization) algorithms [91] have been shown to effectively solve a variety of optimization problems in machine learning, statistical estimation and signal processing. A comprehensive treatment on the theory and implementation of MM algorithms can be found in [92]. Summaries and tutorials on MM algorithms for various problems can be found in [93, 90, 94, 95, 92] Some theoretical analyses of MM algorithms can be found in [96, 97, 98].

3.2 Methods

3.2.1 Model, notations and assumptions

It is understood that scRNA-seq data are collected from many cells of different types when cells are exposed to different dose levels. Since the data for every gene will be analyzed separately, let us denote the data from a given gene by $\{Y_{i,j}, D_{i,j}, X_{i,j} \mid i = 1, \dots, I, j = 1, \dots, n_i\}$. Here i and j are used to denote the suffix for the cell-type and a cell within a specific type of cell. Here $Y_{i,j}$ denotes the response (scRNA-seq expression) from the j th cell of i th cell type, $D_{i,j}$ and $X_{i,j}$ denote the dose-level and a set of covariates for the corresponding cell. Since the scRNA-seq expression contain excessive amount of zeros, we adopt a zero-inflated negative binomial distribution:

$$Y_{i,j} \sim \omega_{i,j} NB(m_{i,j} = s_{i,j}\mu_{i,j}, \phi) + (1 - \omega_{i,j})I(Y_{i,j} = 0). \quad (3.4)$$

Model (3.4) implies that the response is assumed to come from a two-component mixture of two distributions. One component has a mass fully concentrated on zero, and the other component is the Negative Binomial distribution with the mean of $m_{i,j}$ and the scale of $\exp(\phi)$, where ϕ is on the real line. We define the mean $m_{i,j} = s_{i,j}\mu_{i,j}$ where $s_{i,j}$ represents the cell-specific biases and $\mu_{i,j}$, expected transcript count. The scale factors $s_{i,j}$ are computed using scran [99], an approach that uses pooling across single cells to normalize scRNA-seq sequencing data with high number of zeroes. Further $\mu_{i,j}$ is assumed to be function that depends on the dose D_j and covariates $X_{i,j}$. Specifically, we assume that

$$\mu_{i,j} = \exp\{\zeta_i(D_{i,j}) + X_{i,j}^\top \beta\}, \quad (3.5)$$

where ζ_j is nonparametric function of the dose specific to i th type of cells. Our goal is to estimate this function ψ_i . The covariate $X_{i,j}$ is assumed to exert a linear effect through the regression parameter β . Our method implicitly assumes that ζ_i can be well approximated between boundary points a and b by a cubic B spline basis with some number of knots. We will denote knot configurations by pairs (K, τ) , where the number of knots K is a

non-negative integer and the knot locations are given by the K dimensional vector $k = (k_1, \dots, k_K)$, for $a < d_{(1)} < \tau_1 \leq \dots \leq \tau_K < d_{(n)} < b$ for $k = 1, \dots, K$ denote the k^{th} function in a cubic B-spline basis with natural boundary constraints, i.e. linear outside $[a, b]$.

Therefore the dose response curve is approximated as:

$$\zeta_i(D) = \sum_{m=1}^M \gamma_{i,m} B_m(D),$$

with $M = d + K + 1$. With the above model for the dose-response curve, we can write

$$\mu_{i,j} = \exp\left\{\sum_{m=1}^M \gamma_{i,m} B_m(D_{i,j}) + X_{i,j}^\top \beta\right\}. \quad (3.6)$$

Furthermore, lab experiments indicate that the chance of zero inflation tends to decrease with the increasing dose level. This is because a lot of genes are induced, thereby increasing the mean while simultaneously reducing the dropouts. Therefore we propose to model the inflation parameter as a function of the mean of the negative binomial component (3.4), which means we model $\omega_{i,j}$ in terms of $s_i \mu_{i,j}$

$$\text{logit}(\omega_{i,j}) = \psi_0 + \psi_1 \log(s_i \mu_{i,j}), \quad (3.7)$$

where ψ_0 and ψ_1 are two unknown parameters to be estimated. Define $\Theta = (\phi, \beta, \gamma_{1,1}, \dots, \gamma_{1,M}, \dots, \gamma_{J,1}, \dots, \gamma_{J,M}, \psi_0, \psi_1)^\top$. Then the log-likelihood function is

$$\begin{aligned} \ell(\Theta) = & \sum_{i=1}^I \sum_{j=1}^{n_i} \left(I(Y_{i,j} > 0) \left[Y_{i,j} \log(s_i \mu_{i,j}) - \{Y_{i,j} + \exp(\phi)\} \log\{s_i \mu_{i,j} + \exp(\phi)\} \right. \right. \\ & + \phi \exp(\phi) + \log \left[Y_{i,j} + \exp(\phi) \right] - \log(Y_{i,j}!) - \log \left[\exp(\phi) + \log(\omega_{i,j}) \right] \\ & \left. \left. + I(Y_{i,j} = 0) \log \left[(1 - \omega_{i,j}) + \omega_{i,j} \left\{ \frac{\exp(\phi)}{s_i \mu_{i,j} + \exp(\phi)} \right\}^{e^\phi} \right] \right] \right). \end{aligned}$$

The goal is estimating the parameter by maximizing $\ell(\Theta)$. The maximization of $\ell(\Theta)$ is difficult as the dimension of Θ could be large. Therefore, we develop an MM algorithm to ease the complexity of this optimization. The first step of the MM algorithm is developing a minorization function that is easy to optimize compared to $\ell(\Theta)$ and it is presented in the following theorem. Define $\Theta^{(0)} = (\phi^{(0)}, \beta^{(0)}, \gamma_{1,1}^{(0)}, \dots, \gamma_{1,M}^{(0)}, \dots, \gamma_{J,1}^{(0)}, \dots, \gamma_{J,M}^{(0)}, \psi_0^{(0)}, \psi_1^{(0)})^\top$.

Theorem 1. *The minorizing function of $\ell(\Theta)$ is $\ell^\dagger(\Theta|\Theta^{(0)})$, such that*

$$\begin{aligned}\ell^\dagger(\Theta|\Theta^{(0)}) &= \ell_1^\dagger(\phi|\Theta^{(0)}) + g_1(\psi_0|\Theta^{(0)}) + g_2(\psi_1|\Theta^{(0)}) + \sum_{\iota=1}^I \ell_{3,\iota}^\dagger(\gamma_\iota|\Theta^{(0)}) \\ &\quad + \ell_4^\dagger(\beta|\Theta^{(0)}) + \ell_5^\dagger(\Theta^{(0)}),\end{aligned}\tag{3.8}$$

$\ell(\Theta) \geq \ell^\dagger(\Theta|\Theta^{(0)})$ with equality holding when $\Theta = \Theta^{(0)}$, and the terms of (3.8) are

$$\begin{aligned}\ell_1^\dagger(\phi|\Theta^{(0)}) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left[I(Y_{i,j} > 0) \{Y_{i,j} + \exp(\phi)\} + \exp(\phi) \Gamma_{i,j}^{(0)} \right] \right. \\ &\quad \left. \left\{ \frac{\exp(\phi^{(0)}) - \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} + \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \right. \\ &\quad \times \left[\phi \exp(\phi) - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{ \exp(\phi) - \exp(\phi^{(0)}) \}^2 \right. \\ &\quad \left. \left. - \exp(\phi) \log \{ s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \} \right] \right. \\ &\quad \left. + I(Y_{i,j} > 0) \left\{ \log \left[Y_{i,j} + \exp(\phi) \right] - \log \left[\exp(\phi) \right] \right\} \right), \\ g_1(\psi_0|\Theta^{(0)}) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[\left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \psi_0 - \frac{\omega_{i,j}^{(0)}}{2M+5} \exp \{ (\psi_0 - \psi_0^{(0)})(2M+5) \} \right] \\ g_2(\psi_1|\Theta^{(0)}) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \left[\psi_1 \log(s_i) - 0.5(M+1)(\psi_1 - \psi_1^{(0)})^2 \right. \right. \\ &\quad \left. \left. + \psi_1 \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \right] \right. \\ &\quad \left. - \frac{\omega_{i,j}^{(0)}}{2M+5} \left(\exp \{ \log(s_i)(\psi_1 - \psi_1^{(0)})(2M+5) \} \right. \right. \\ &\quad \left. \left. + 0.5(M+1) \exp \{ (\psi_1 - \psi_1^{(0)})^2 \} \right. \right. \\ &\quad \left. \left. + \exp \left[(\psi_1 - \psi_1^{(0)}) \left\{ \sum_{m=1}^M \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} (2M+5) \right] \right) \right\}, \\ \ell_{3,i}^\dagger(\gamma_i|\Theta_0) &= \sum_{j=1}^{n_i} \left\{ \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \right. \\ &\quad \times \left[-0.5 \sum_{m=1}^M (\gamma_{i,m} - \gamma_{i,m}^{(0)})^2 B_m^2(D_{i,j}) + \psi_1^{(0)} \sum_{m=1}^M (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right] \\ &\quad \left. - \frac{\omega_{i,j}^{(0)}}{2M+5} \left(0.5 \sum_{m=1}^M \exp \{ (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})(2M+5) \}^2 \right) \right\}\end{aligned}$$

$$\begin{aligned}
& + \sum_{m=1}^M \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)})\psi_1^{(0)} B_m(D_{i,j})(2M+5)\} \Big) \Big\} \\
& + \sum_{j=1}^{n_i} \left(I(Y_{i,j} > 0) Y_{i,j} \sum_{m=1}^M \gamma_{i,m} B_m(D_{i,j}) \right. \\
& - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& \left. \left. + \Gamma_{i,j}^{(0)} \exp(\phi^{(0)}) \right\} \right. \\
& \times \frac{\mu_{i,j}^{(0)}}{M+1} \sum_{m=1}^M \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})(M+1)\} \\
& - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \\
& \times \left[-2(\mu_{i,j}^{(0)})^2 \sum_{m=1}^M (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right. \\
& \left. \left. + \frac{(\mu_{i,j}^{(0)})^2}{M+1} \sum_{m=1}^M \exp\{2(M+1)(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})\} \right] \right),
\end{aligned}$$

$$\begin{aligned}
\ell_4^\dagger(\beta|\Theta_0) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \left[-0.5 \{X_{i,j}^\top (\beta - \beta^{(0)})\}^2 \right. \right. \\
& \left. \left. + \psi_1^{(0)} X_{i,j}^\top (\beta - \beta^{(0)}) \right] \right. \\
& - \frac{\omega_{i,j}^{(0)}}{2M+5} \left(0.5 \exp[\{X_{i,j}^\top (\beta - \beta^{(0)}) (2M+5)\}^2] \right. \\
& \left. \left. + \exp\{X_{i,j}^\top \psi_1^{(0)} (\beta - \beta^{(0)}) (2M+5)\} \right) \right\} \\
& + \sum_{i=1}^I \sum_{j=1}^{n_i} \left(I(Y_{i,j} > 0) Y_{i,j} X_{i,j}^\top \beta \right. \\
& - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) + \Gamma_{i,j}^{(0)} \exp(\phi^{(0)}) \right\} \\
& \times \frac{\mu_{i,j}^{(0)}}{M+1} \exp\{X_{i,j}^\top (\beta - \beta^{(0)}) (M+1)\} \\
& - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \\
& \times \left[-2(\mu_{i,j}^{(0)})^2 X_{i,j}^\top (\beta - \beta^{(0)}) + \frac{(\mu_{i,j}^{(0)})^2}{M+1} \exp\{2(M+1) X_{i,j}^\top (\beta - \beta^{(0)})\} \right] \Big),
\end{aligned}$$

where, $\Gamma_{i,j}^{(0)} = I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})/\{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})\}$,
 $\omega_{i,j}^{(0)} = \exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}/[1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}]$, and
 $G(\phi, \mu_{i,j}) = [\exp(\phi)/\{s_j\mu_{i,j} + \exp(\phi)\}]^{e^\phi}$.

The noticeable fact is that the minorizing function separates all the parameters. Consequently, the follow-up calculation becomes extremely simple compared to the case of non-separated parameters in the joint update of all the parameters in any iterative procedure. Next, parameters are estimated by the Newton-Raphson method. This technique of updating the parameters is known as gradient MM algorithm [93]. Let $\Theta^{(t)}$ and $\Theta^{(t+1)}$ be the parameter value at the t th and $(t+1)$ th iterations, respectively, then in the gradient MM approach, $\Theta^{(t+1)} = \Theta^{(t)} - \left\{ \frac{\partial^2 \ell^t(\Theta|\Theta^{(t)})}{\partial \Theta \partial \Theta^\top} \Big|_{\Theta=\Theta^{(t)}} \right\}^{-1} \frac{\partial \ell^t(\Theta|\Theta^{(t)})}{\partial \Theta} \Big|_{\Theta=\Theta^{(t)}}$. Starting with an initial value of Θ , this update is repeated until the parameters converge with a specified tolerance. Note that in the above update, instead of the log-likelihood, we are using the minorizing function. Because of the separation of the parameters, $\{\partial^2 \ell^t(\Theta|\Theta^{(t)})/\partial \Theta \partial \Theta^\top\}$ turns out to be diagonal matrix. Specifically,

$$\begin{aligned} \frac{\partial \ell^t(\Theta|\Theta^{(0)})}{\partial \phi} \Big|_{\Theta=\Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(-\frac{\exp(\phi^{(0)})}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left[Y_{i,j}I(Y_{i,j} > 0) + \exp(\phi^{(0)}) \right. \right. \\ &\quad \times \left. \left. \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \right] \right. \\ &\quad + I(Y_{i,j} > 0) \left\{ \frac{\partial}{\partial \phi} \log \left[Y_{i,j} + \exp(\phi) \right] \Big|_{\phi=\phi^{(0)}} \right. \\ &\quad \left. \left. - \frac{\partial}{\partial \phi} \log \left[\exp(\phi) \right] \Big|_{\phi=\phi^{(0)}} \right\} \right. \\ &\quad + \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \left[\exp(\phi^{(0)}) + \exp(\phi^{(0)})\phi^{(0)} \right. \\ &\quad \left. \left. - \exp(\phi^{(0)})\log \left\{ s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \right] \right), \\ \frac{\partial \ell^t(\Theta|\Theta^{(0)})}{\partial \psi_0} \Big|_{\Theta=\Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} - \omega_{i,j}^{(0)} \right\}, \\ \frac{\partial \ell^t(\Theta|\Theta^{(0)})}{\partial \psi_1} \Big|_{\Theta=\Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ \log(s_i) + \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \\ &\quad \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} - \omega_{i,j}^{(0)} \right\}, \end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell^\dagger(\Theta|\Theta^{(0)})}{\partial \beta} \Big|_{\Theta=\Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} X_{i,j} \left[\psi_1^{(0)} \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} - \omega_{i,j}^{(0)} \right\} + I(Y_{i,j} > 0) Y_{i,j} \right. \\
&\quad \left. - \frac{s_j \mu_{i,j}^{(0)}}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} \right. \right. \\
&\quad \left. \left. + I(Y_{i,j} > 0) \exp(\phi^{(0)} + \Gamma_{i,j}^{(0)} \exp(\phi^{(0)})) \right\} \right], \\
\frac{\partial \ell^\dagger(\Theta|\Theta^{(0)})}{\partial \gamma_{i,m}} \Big|_{\Theta=\Theta^{(0)}} &= \sum_{j=1}^{n_i} B_m(D_{i,j}) \left[\psi_1^{(0)} \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} - \omega_{i,j}^{(0)} \right\} + I(Y_{i,j} > 0) Y_{i,j} \right. \\
&\quad \left. - \frac{s_j \mu_{i,j}^{(0)}}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)} \right. \right. \\
&\quad \left. \left. + \Gamma_{i,j}^{(0)} \exp(\phi^{(0)}) \right\} \right].
\end{aligned}$$

Also,

$$\begin{aligned}
\frac{\partial^2}{\partial \phi^2} \ell^\dagger(\Theta|\Theta^{(0)}) \Big|_{\Theta=\Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(-\frac{Y_{i,j} \exp(\phi^{(0)})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \right. \\
&\quad \left. \left. + \frac{\partial^2}{\partial \phi^2} \log \left| Y_{i,j} + \exp(\phi) \right| \Big|_{\Theta=\Theta^{(0)}} - \frac{\partial^2}{\partial \phi^2} \log \left| \exp(\phi) \right| \Big|_{\Theta=\Theta^{(0)}} \right) \\
&\quad + \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \left[2 \exp(\phi^{(0)}) + \exp(\phi^{(0)}) \phi^{(0)} \right. \\
&\quad \left. - \exp(\phi^{(0)}) \log \left\{ s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \right. \\
&\quad \left. \left. - \frac{3 \exp(2\phi^{(0)}) + s_j \exp(2\phi^{(0)})}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right] \right\}, \\
\frac{\partial^2}{\partial \psi_0^2} \ell^\dagger(\Theta|\Theta_0) \Big|_{\Theta=\Theta^{(0)}} &= (2M + 5) \sum_{i=1}^I \sum_{j=1}^{n_i} \omega_{i,j}^{(0)}, \\
\frac{\partial^2}{\partial \psi_1^2} \ell^\dagger(\Theta|\Theta^{(0)}) \Big|_{\Theta=\Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(-(M + 1) \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \right. \\
&\quad \left. - \frac{\omega_{i,j}^{(0)}}{2M + 5} \left[(2M + 5)^2 \times \{\log(s_i)\}^2 + M + 1 \right. \right. \\
&\quad \left. \left. + \left\{ \sum_{m=1}^M \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\}^2 (2M + 5)^2 \right] \right), \\
\frac{\partial^2}{\partial \gamma_{i,r}^2} \ell^\dagger(\Theta|\Theta^{(0)}) \Big|_{\Theta=\Theta^{(0)}} &= - \sum_{j=1}^{n_i} B_r^2(D_{i,j}) \left(\left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} + (2M + 5) \right\} \right.
\end{aligned}$$

$$\begin{aligned}
& \times (1 + \psi_1^{(0),2})\omega_{i,j}^{(0)} \Big\} \\
& + \frac{s_j\mu_{i,j}^{(0)}(M+1)}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left[\left\{ I(Y_{i,j} > 0)Y_{i,j} + I(Y_{i,j} > 0)\exp(\phi^{(0)}) \right. \right. \\
& \left. \left. + \Gamma_{i,j}^{(0)}\exp(\phi^{(0)}) \right\} \right. \\
& \left. + 2\mu_{i,j}^{(0)} \left\{ I(Y_{i,j} > 0) + \Gamma_{i,j}^{(0)} \right\} \right] \Bigg),
\end{aligned}$$

A detailed proof of the theorem is provided in the Appendix for Chapter 3.

3.2.2 Penalized Estimation

In our proposed spline model, regression is performed by choosing a set of knots and by finding the spline defined over these knots. In this case, the number of knots has an important influence on the resulting fit; not having enough knots leads to underfitted regression and too many knots lead to overfitted model. Choosing the position of knots is also an issue since uniformly distributed knots can lead to overfitting in an area where there are few points and underfitting in an area where there are many points. The most widely used spline regression methods overcome this difficulties by using a penalization approach. In smoothing splines, knots are set at each data point and the wiggliness of the spline is controlled by penalizing over its integrated squared second order derivative $\int \psi''(x)^2 dx$. Substantial statistical methodology exists in the literature for finding the best number and location of knots; we refer to [100] for a review. To this effect, one can choose a large number of equally spaced initial knots and estimate the model by maximizing a penalized likelihood function. We follow [101] Chapter 5.5.3, and set for $k = 1, \dots, K$, $\tau_k = \frac{k+1}{K+2}th$ sample quantile of the unique dose levels. The default choice for $K = \min(0.25 \times \text{the number of unique dose levels}, 35)$. This knot placement procedure has been shown to perform well in a lot of scenarios. Alternatively the statistician may choose K based on visual inspection of the data scatterplot to better depict the complexity of the regression function relative to the noise in the data. The penalized log-likelihood function

is $\ell_p(\Theta) = \ell(\Theta) - \sum_i \lambda_i \mathcal{P}(\gamma_i)$, where $\lambda_i > 0$ is the smoothing parameter and $\mathcal{P}(\gamma_i)$ denotes the penalty function for the i th dose-response curve. The minorizing function for $\ell_p(\Theta)$ is simply $\ell^\dagger(\Theta|\Theta^{(0)}) - \sum_i \lambda_i \mathcal{P}(\gamma_i)$. This will lead to a minor adjustment to the expressions when taking derivatives of the minorizing function with respect to γ_i . Procedures such as cross validation and generalized cross validation are used for the selection of the smoothness parameter λ ([101] Chapter 5)

3.2.3 Monotonicity constraints

Our primary focus is inference on the unknown regression function $\log(\mu_i)$. Equation (1) works fine when we have no prior information on the shape of the regression function. However, a high percentage of genes are expected to have a higher chance of an adverse response at increasing dose levels, after adjusting for important confounding factors, such as age and sex. Additionally, incorporating monotonicity constraints while assessing the dose response function has been shown to improve estimation efficiency and power to detect trends [102]. To impose monotonicity constraints on Equation (1), we need to assume that $\psi_i \in \Theta^+$ is an isotonic regression function with $\Theta^+ = \{\zeta_i(d_1) \leq \zeta_i(d_2) \forall (d_1, d_2) \in \mathcal{R}^2, d_1 < d_2\}$. Currently we express the log-mean function as $\log(\mu_{i,j}) = \sum_{m=1}^M \gamma_{i,m} B_m(D_{i,j}) + X_{i,j}^\top \beta$. To make this function monotone increasing with respect to dose d , we need $\gamma_{i,m} > 0$, for $m = 1, \dots, M$ [103]. Therefore we write

$$\begin{aligned} \gamma_{i,m} &= \exp(\eta_{i,m}) \\ \log(\mu_{i,j}) &= \sum_{m=1}^M \exp(\eta_{i,m}) B_m(D_{i,j}) + X_{i,j}^\top \beta \end{aligned}$$

According to theorem 5.9 of [104], $\zeta_i(D) = \sum_{m=1}^M \gamma_{i,m} B_m(D)$, is a class of monotone nondecreasing splines, since the monotonicity of the B-splines is guaranteed by the nondecreasing order of co- efficients.

Thus our goal now is estimation of parameters

$$\Theta = (\phi^{(0)}, \beta^{(0)}, \eta_{1,1}^{(0)}, \dots, \eta_{1,M}^{(0)}, \dots, \eta_{J,1}^{(0)}, \dots, \eta_{J,M}^{(0)}, \psi_0^{(0)}, \psi_1^{(0)})^\top$$

Accounting for this parameter transformation in the minorization function, only changes the first and second derivative calculation with respect to $\gamma_{i,m}$ for $m = 2, \dots, M$. Now, the first derivative of the minorization function with respect to $\eta_{i,m}$ will be

$$\frac{\partial}{\partial \eta_{i,m}} g_{5,i}(\eta_i | \Theta^{(0)}) = \left(\frac{\partial}{\partial \gamma_{i,m}} g_{3,i}(\exp(\eta_i) | \Theta^{(0)}) + \frac{\partial}{\partial \gamma_{i,m}} \ell_{3,i}^\dagger(\exp(\eta_i) | \Theta^{(0)}) \right) \times \exp(\eta_{i,m})$$

Define

$$H_{m,m'}(\gamma_{i,m}) = \frac{\partial^2}{\partial \gamma_{i,m}^2} g_{3,i}(\gamma_i | \Theta^{(0)}) + \frac{\partial^2}{\partial \gamma_{i,r}^2} \ell_{3,i}^\dagger(\gamma_i | \Theta^{(0)})$$

Then the second derivatives with respect to $\eta_{i,m}$ can be written as

$$\begin{aligned} H_{m,m'}(\eta_{i,m}) &= H_{m,m'}(\exp(\eta_{i,m})) \times \exp(\eta_m) \times \exp(\eta_{m'}) \\ &\quad + \frac{\partial}{\partial \eta_{i,m}} g_{5,i}(\eta_i | \Theta^{(0)}) \times I(m = m') \end{aligned}$$

The rest of the computations remain exactly the same as before.

3.2.4 Confidence interval estimation

The gradient of the penalized log-likelihood function is

$$\frac{\partial \ell_p(\Theta)}{\partial \Theta} = \frac{\partial \ell(\Theta)}{\partial \Theta} - \sum_i \lambda_i \frac{\partial}{\partial \Theta} \mathcal{P}(\gamma_i) = \sum_{i=1}^I \sum_{j=1}^{n_i} U_{i,j},$$

where $U_{i,j} = \{\partial \ell_{i,j}(\Theta) / \partial \Theta - (\lambda_i / n_i) \partial \mathcal{P}(\gamma_i) / \partial \Theta\} |_{\Theta = \hat{\Theta}}$.

Let $H = \partial^2 \ell_p(\Theta) / \partial \Theta \partial \Theta^\top |_{\Theta = \hat{\Theta}}$, and define $A = -(H / \sum_{i=1}^I n_i)^{-1}$. Then the variance-covariance of $\sqrt{\sum_{i=1}^I n_i} \hat{\Theta}$ can be estimated by

$$V_0 = A \left(\frac{1}{\sum_{i=1}^I n_i} \sum_{i=1}^I \sum_{j=1}^{n_i} U_{i,j} U_{i,j}^\top \right) A^\top$$

. The square root of the diagonal elements of V_0 divided by the square root of the total sample size, $\sum_i n_i$, is the standard error of the estimator of each component of the Θ vector. An important component of the model is estimation of dose-response relation for a gene in a specific cell type. Suppose the interest is in the i th cell type, so the dose-response relation

estimator for the i th cell type is $\widehat{G}_i(D^*) = \sum_{m=1}^M B_m(D^*) \widehat{\gamma}_{i,m}$ at the dose level D^* . We can re-write $\widehat{G}_i(D^*) = L^\top \widehat{\Theta}$, where

$$L^\top = (0, 0, 0, 0, B_1(D^*), \dots, B_M(D^*), 0, \dots, 0)$$

Therefore,

$$\text{var}(\widehat{G}_i(D)) = \text{var}(L^\top \widehat{\Theta}) = L^\top \left(\frac{V_0}{\sum_{i=1}^I n_i} \right) L.$$

Then the 95% pointwise CI for $G_i(D^*)$ is $\widehat{G}_i(D^*) \pm 1.96 \times \sqrt{\text{var}(\widehat{G}_i(D^*))}$.

3.2.5 Model Selection

In case, the user is not sure whether placing monotonicity constraints betters the model estimation, we suggest fitting both the unconstrained and constrained versions of the models and selecting the model with the lowest AIC where $AIC = 2\theta - 2\log(l(\widehat{\Theta}))$, where θ is the number of estimated parameters and $\log(l(\widehat{\Theta}))$ is maximum value of the likelihood function for the model. In case of penalized estimation, the value of the optimal λ can be chosen using cross-validation and then the value of the log-likelihood can be calculated using the MLE estimates obtained via MM algorithm for penalized estimation.

3.3 Hypothesis Testing

There are two tests of hypothesis that we wish to conduct for the model described above:

1. Test if there is any effect of dose on the response for a given cell-type. Suppose that we are interested in the i th cell type, then set $H_{0,1} : \gamma_{i,2} = \gamma_{i,3} = \dots \gamma_{i,M}$ versus the alternative $H_{a,1} : \text{at least one of the components of } \gamma_i \text{ is different from the rest of the components.}$
2. To test if the dose-response curves differ by the cell-types we set $H_{0,2} : \gamma_1 = \dots = \gamma_I$ versus the alternative $H_{a,2} : \text{at least one of } \gamma_i \text{ is different from the rest.}$

For both tests, we use likelihood ratio test. The likelihood ratio test statistic is given by

$$\Lambda_\psi = 2 \left\{ \ell(\hat{\Theta}) - \ell(\hat{\Theta}_{H_0}) \right\}$$

where $\hat{\Theta}$, calculated via the MM algorithm, denotes the unconstrained MLEs of Θ while $\hat{\Theta}_{H_0}$ is the constrained MLE of Θ under H_0 , which can be obtained by

$$\Theta_{H_0}^{(t+1)} = \Theta_{H_0}^{(t)} - \left\{ \frac{\partial^2 \ell^\dagger(\Theta_{H_0} | \Theta_{H_0}^{(t)})}{\partial \Theta_{H_0} \partial \Theta_{H_0}^\top} \Big|_{\Theta_{H_0} = \Theta_{H_0}^{(t)}} \right\}^{-1} \frac{\partial \ell^\dagger(\Theta_{H_0} | \Theta_{H_0}^{(t)})}{\partial \Theta_{H_0}} \Big|_{\Theta_{H_0} = \Theta_{H_0}^{(t)}}$$

where

$$\begin{aligned} \ell^\dagger(\Theta_{H_0} | \Theta_{H_0}^{(0)}) &= \ell_1^\dagger(\phi_{H_0} | \Theta_{H_0}^{(0)}) + g_1(\psi_{0H_0} | \Theta_{H_0}^{(0)}) + g_2(\psi_{1H_0} | \Theta_{H_0}^{(0)}) + \sum_{i=1}^I \ell_{3,i}^\dagger(\gamma_{iH_0} | \Theta_{H_0}^{(0)}) \\ &\quad + \ell_4^\dagger(\beta_{H_0} | \Theta_{H_0}^{(0)}) + \ell_5^\dagger(\Theta_{H_0}^{(0)}) \end{aligned}$$

For problem 1), the null model, $\psi_i(D)$ does not vary with D , so is a constant. Therefore, $\mu_{i,j} = \exp(\gamma^* + X_{i,j}^T \beta)$. Under the null hypothesis $H_{0,1}$, Λ_ψ is asymptotically follows a chi-squared distribution with degrees of freedom $(M - 1)$. Under the null hypothesis $H_{0,2}$, $\gamma_1 = \dots = \gamma_I = \gamma$ (say), so Λ_ψ is asymptotically follows a chi-squared distribution with degrees of freedom $(I - 1)M$.

3.4 Results

3.4.1 Simulation Design

To study the behavior of the procedure, we conducted a simulation study. We generated the log-mean dose response curve as; $\log(\mu_{ij}) = \zeta_i(D_{ij}^{(s)}) + \beta X_{ij}$, where $D_{i,j} = \{0, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30\}$ and $D_{ij}^{(s)} = \frac{\log(D_{i,j} + 1)}{\log(30)}$. The dose response function $\zeta_i(D_{ij}^{(s)})$ is then assumed to follow four different patterns

- Case 1:

$$\begin{aligned} \zeta_{1,1}(D_{1,j}^{(s)}) &= 1 + 0.1D_{1,j}^{(s)} + 0.1D_{1,j}^{(s)2} + 3(D_{1,j}^{(s)} - 0.5)_+^2 \\ \zeta_{1,2}(D_{2,j}^{(s)}) &= 1 + 0.1D_{1,j}^{(s)} + 0.8D_{1,j}^{(s)2} + 0.8(D_{1,j}^{(s)} - 0.5)_+^2 \\ \zeta_{1,3}(D_{3,j}^{(s)}) &= 0.5 + 0.2D_{1,j}^{(s)} \end{aligned}$$

- Case 2:

$$\begin{aligned}\zeta_{2,1}(D_{1,j}^{(s)}) &= 0.15 \exp\{2 \times D_{1,j}^{(s)}\} \\ \zeta_{2,2}(D_{2,j}^{(s)}) &= 0.15 \exp[\{2 + (0.5/300)\} \times D_{2,j}^{(s)}] \\ \zeta_{2,3}(D_{3,j}^{(s)}) &= 0.15 \exp[\{2 + (0.6/300)\} \times D_{3,j}^{(s)}]\end{aligned}$$

- Case 3:

$$\zeta_i(D_{ij}^{(s)}) = \frac{1.5 \sin(D_{ij}^{(s)})}{1 + 1.5 \sin(D_{ij}^{(s)})} + 2, \quad i = 1, 2, 3$$

- Case 4:

$$\zeta_i(D_{ij}^{(s)}) = 3, \quad i = 1, 2, 3$$

We generated a single covariate $X_{i,j} \sim N(0, 1)$ and set $\beta = 0.5$. The B-spline basis matrix $B_m(D_{i,j})$ is then generated using three knots at 0.008, 0.075 and 0.4. We also fixed the degree of the B-spline basis as $d = 3$. We chose $\psi_0 = 0.5$, $\psi_1 = 0.1$ and computed $\omega_{i,j}$ using Equation 3.7. The binary random variable $R_{i,j}$ is then generated from a Bernoulli distribution with mean $\omega_{i,j}$. Finally expression data Y_{ij} is generated as $Y_{i,j} = Y'_{i,j} R_{i,j}$ where, $Y'_{i,j}$ is generated from a negative binomial distribution with mean $s_{i,j} \mu_{i,j}$, and dispersion $\exp(\phi)$ where $\phi \sim N(0.5, 0.1)$. Datasets with three cell types are generated. The analysis is repeated with cell-type specific sample sizes $n_i = 100, 300$.

3.4.2 Analysis

For each dose response pattern and for $n_i = (100, 300)$ we simulated 500 data sets. Tables 3.1, 3.2, 3.3 and 3.4 summarizes the estimated mean, SE, bias, Mean square error and coverage probabilities of parameters β, ϕ, ψ_0 and ψ_1 for cases 1-4. The results are compared with an intercept model having $\psi_1 = 0$. It can be seen in general that the performance of the full model is better than the intercept model in all aspects, thus indicating a dependence of the zero-inflation parameter on the log-mean. Figures 3.1, 3.3, 3.5, 3.6, 3.9, 4.0, 4.3 and 4.4

demonstrate the fits of the estimated curve along the 95% confidence interval. From the figures, it can be seen that the fitted values for both sample sizes are close to the true values, and the 95% confidence bands in general covers the true curves. As expected, the confidence intervals are wider when sample size decreases.

To further demonstrate the performance of our proposed ZINB spline model (ZINB-SPL) we benchmarked it against three competing methods; (a) Zero inflated negative binomial generalized additive model (ZINB-GAM-Dose) with the zero inflation parameter dependent on the dose as $\text{logit}(\omega_{ij}) = \gamma_0 + \gamma_1 * \text{dose}$, (b) Zero inflated negative binomial generalized additive model (ZINB-GAM-Int) with the zero inflation parameter having an intercept model; $\text{logit}(\omega_{ij}) = \gamma_0$, and (c) Negative Binomial GAM. We use R packages `mgcv` [105] for fitting NB GAM and `zigam` [106] for fitting models (a) and (b). The benchmarking curves for the simulation scenarios are reported in Figures 3.3, 3.7, 3.11 and 3.15. The performance of the four models are very close in case of the linear and constant models demonstrated in simulation scenarios 3 and 4. However ZINB-SPL leads to almost perfect curve estimates in non-linear curves shapes illustrated in simulation scenarios 1 and 2. This is further demonstrated by the RMSE boxplots in Figures 3.4, 3.8, 3.12 and 3.16. The RMSE boxplots in Figure 3.4 and 3.8 for the ZINB-SPL lies significantly lower than the other 3 models. Also note that the NB-GAM model has the worst performance in scenarios 1 and 2, hence justifying the need of a zero-inflation structure to model the excess zeroes in the data generating process.

Under Case 1 and Case 2, we had 500 positive control datasets where true mean curve varies for the three cell-types. Therefore Case 1 and Case 2 were used to study the power of Test 2 in detecting true positive instances, (Section 3.3). In this scenario, it is expected that the the LRT test statistic will have large values, with the expected p-value close to 0. Therefore the expected true positive rate is computed as $\text{TPR} = \frac{1}{500} \sum_{l=1}^{500} I[P(\Lambda_{\zeta,l} > \chi_{M(I-1)}^2 | H_a) < \alpha]$. Case 2 embodies a local alternative testing scenario where the slope parameter of the DR functions for cell type 2 and 3 are written as $\alpha_2 = \alpha_1 + g_1/n_k$ and

$\alpha_3 = \alpha_1 + g_2/n_k$, where α_1 is the slope parameter of the DR function for cell-type 1 and $g_1 = 0.5$ and $g_2 = 0.6$ are sufficiently close. Therefore it is expected that the power of test 2 in detecting true positives for Case 1 will be much smaller than the expected TPR in case 1, where there is larger discrepancy in the shape of the true mean functions. In Case1, out of 500 simulations, in 490 instances the null hypothesis was rejected with an expected TPR of 0.98 for $n_i = 300$. For $n_i = 300$, 476 instances were rejected the null hypothesis with an expected TPR of 0.952. For Case 2, $n_i = 300$, 298 out of 500 instances yielded a rejection of the null hypothesis with an expected TPR of 0.578 and for $n_i = 100$, 196 out of 500 simulations yielded a rejection of the null hypothesis with an expected TPR of 0.392. The decrease in TPR for smaller sample sizes and local alternatives is an expected result.

Next, under Case 3 and 4, the true log-mean response $\eta_i(D_{ij}^{(s)})$ followed the same DR model for all the cell types. The only difference is that in Case 3, all celltypes has a monotonically increasing DR function, whereas in Case 4, the DR function is a constant. Therefore the datasets generated from these cases serve as negative control datasets, where the true dose response curves does not change with the cell-types. We conducted Test 2 on these datasets to study the ability of the test in controlling for false positives. Since all the datasets in this scenario are generated from a null model, the average p-value should be close to 1 and we determine false positive rate as $FPR = \frac{1}{500} \sum_{l=1}^{500} I[P(\Lambda_{\eta,l} > \chi_{(M)(I-1)}^2 | H_0) < \alpha]$, where $\alpha = 0.05$. For Case 3, $n_i = 300$, 39 out of 500 simulations yielded a rejection of the null hypothesis with expected FPR = 0.076 . For $n_i = 100$, 54 instances rejected the null hypothesis with an expected FPR of 0.108. In Case 4, $n_i = 300$, 22 out of 500 instances are rejected, thus having an expected FPR of 0.044. For $n_i = 100$, 76 out of 500 instances are rejected at a level of 0.05, with an expected FPR of 0.152. This clearly indicates that a higher sample size is preferable for controlling the false positive rates.

Scenario	Model	$n_i = 300$					$n_i = 100$				
		$\hat{\beta}$	SE	Bias	MSE	Cov	$\hat{\beta}$	SE	Bias	MSE	Cov
Case 1	Full	0.507	0.049	0.007	0.003	0.95	0.491	0.083	-0.008	0.008	0.92
	Intercept	0.515	0.049	0.015	0.002	0.94	0.500	0.0824	0.007	0.007	0.93
Case 2	Full	0.498	0.112	-0.002	0.004	0.96	0.502	0.123	0.02	0.015	0.97
	Intercept	0.512	0.045	0.012	0.003	0.998	0.5079	0.0795	-0.007	0.011	0.992
Case 3	Full	0.500	0.037	-0.0015	0.0014	0.942	0.499	0.066	-0.001	0.004	0.938
	Intercept	0.500	0.037	0.001	0.001	0.938	0.498	0.067	0.002	0.004	0.965
Case 4	Full	0.495	0.031	-0.004	0.001	0.952	0.504	0.060	0.004	0.003	0.964
	Intercept	0.501	0.033	0.001	0.001	0.931	0.503	0.060	0.003	0.003	0.964

Table 3.1 Parameter estimate, asymptotic standard error, bias and mean squared error(MSE)of parameter β under full model and the intercept model. The results are averaged over 500 replicates and reported for $n_i = 100, 300$

Scenario	Model	$n_i = 300$					$n_i = 100$				
		$\hat{\phi}$	SE	Bias	MSE	Cov	$\hat{\phi}$	SE	Bias	MSE	Cov
Case 1	Full	0.618	0.168	0.118	0.043	0.902	0.891	0.303	-0.008	0.008	0.74
	Intercept	0.601	0.171	0.101	0.041	0.904	0.888	0.308	0.389	0.263	0.75
Case 2	Full	0.621	0.243	0.221	0.127	0.856	1.283	0.325	0.7836	1.136	0.532
	Intercept	0.687	0.264	0.187	0.119	0.776	1.317	0.314	0.817	1.158	0.525
Case 3	Full	0.554	0.089	0.054	0.011	0.922	0.684	0.155	0.185	0.080	0.764
	Intercept	0.555	0.0907	0.055	0.011	0.901	0.682	0.153	0.182	0.159	0.752
Case 4	Full	0.529	0.070	0.029	0.006	0.851	0.625	0.120	0.126	0.032	0.796
	Intercept	0.540	0.069	0.040	0.007	0.892	0.635	0.120	0.135	0.034	0.762

Table 3.2 Parameter estimate, asymptotic standard error, bias and mean squared error(MSE)of parameter ϕ under full model and the intercept model. The results are averaged over 500 replicates and reported for $n_i = 100, 300$

Scenario	Model	$n_i = 300$					$n_i = 100$				
		$\hat{\psi}_0$	SE	Bias	MSE	Cov	$\hat{\psi}_0$	SE	Bias	MSE	Cov
Case 1	Full	0.512	0.166	0.012	0.050	0.97	0.488	0.427	-0.012	0.173	0.978
	Intercept	0.619	0.268	0.119	0.029	1	0.609	0.480	0.109	0.064	1
Case 2	Full	0.455	0.223	-0.045	0.069	0.96	0.044	0.592	0.101	0.229	0.91
	Intercept	0.552	0.206	0.052	0.048	0.988	0.5218	0.383	0.021	0.146	0.976
Case 3	Full	0.4744	0.353	-0.025	0.0267	1	0.431	0.598	-0.068	0.190	0.972
	Intercept	0.725	0.352	0.0554	0.011	1	0.728	0.602	0.227	0.072	1
Case 4	Full	0.482	0.462	-0.018	0.016	0.961	0.468	0.795	-0.031	0.120	1
	Intercept	0.796	0.502	0.296	0.093	1	0.797	0.788	0.136	0.034	1

Table 3.3 Parameter estimate, asymptotic standard error, bias and mean squared error(MSE)of parameter ψ_0 under full model and the intercept model. The results are averaged over 500 replicates and reported for $n_i = 100, 300$

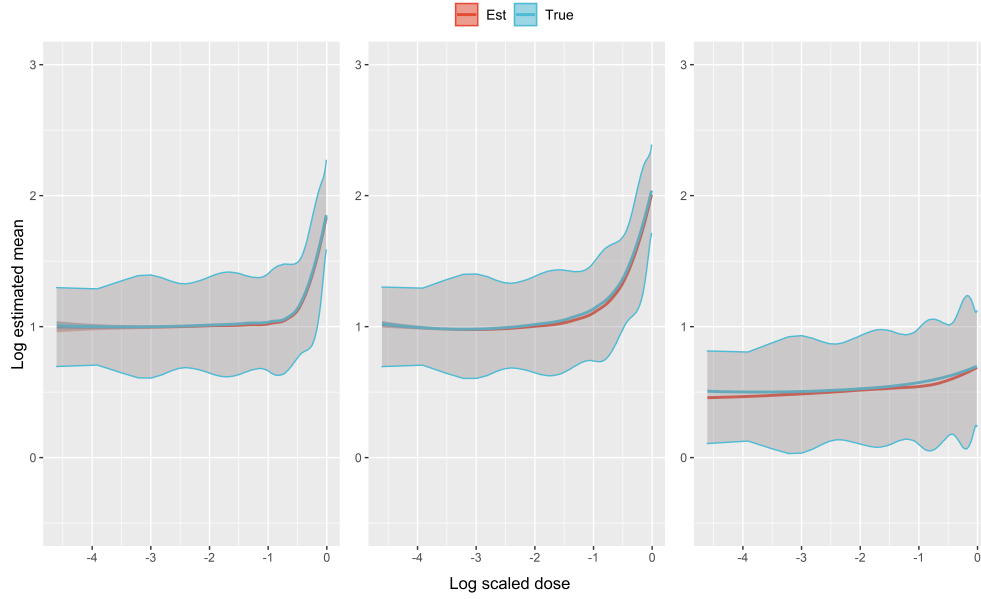


Figure 3.1 Results of the simulation study, illustrating the performance of Model 1 (see Simulation Design) in 500 replicates with 300 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.

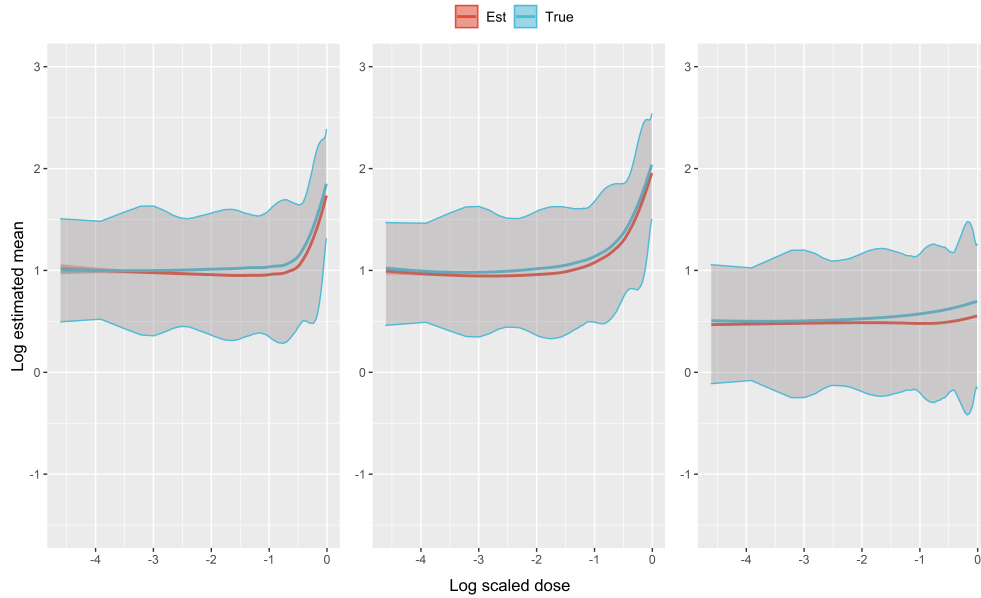


Figure 3.2 Results of the simulation study, illustrating the performance of Model 1 (see Simulation Design) in 500 replicates with 100 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.

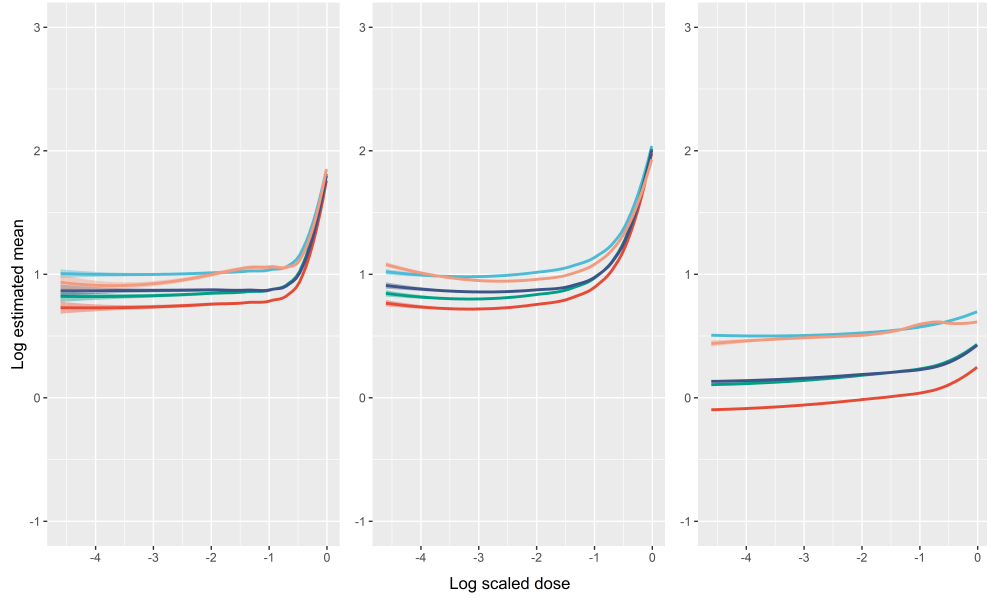


Figure 3.3 Results of the simulation study, illustrating the performance of Model 1 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to the three different cell-types. Continuous cyan, peach, darkblue, darkgreen and red lines represent the true curve and the estimated log-mean curves averaged across 500 random replicates for ZINB-SPL, ZINB-GAM-Dose, ZINB-GAM-Int and NB-GAM models respectively

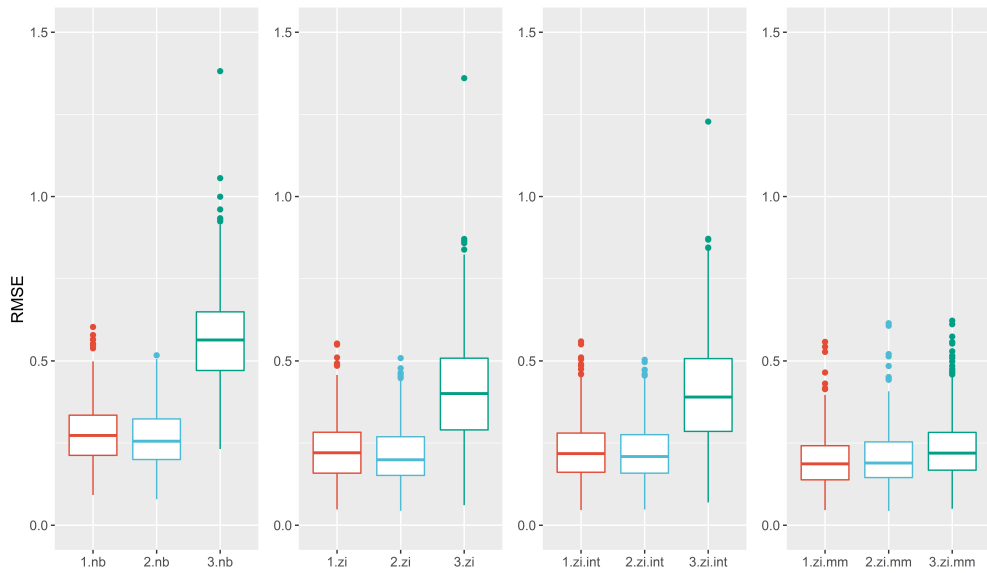


Figure 3.4 Results of the simulation study, illustrating the estimated RMSE of Model 1 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to RMSE boxplots of the NB-GAM, ZINB-GAM-Dose, ZINB-GAM-Int and ZINB-SPL models respectively for the three different cell-types plotted over 500 replicates.

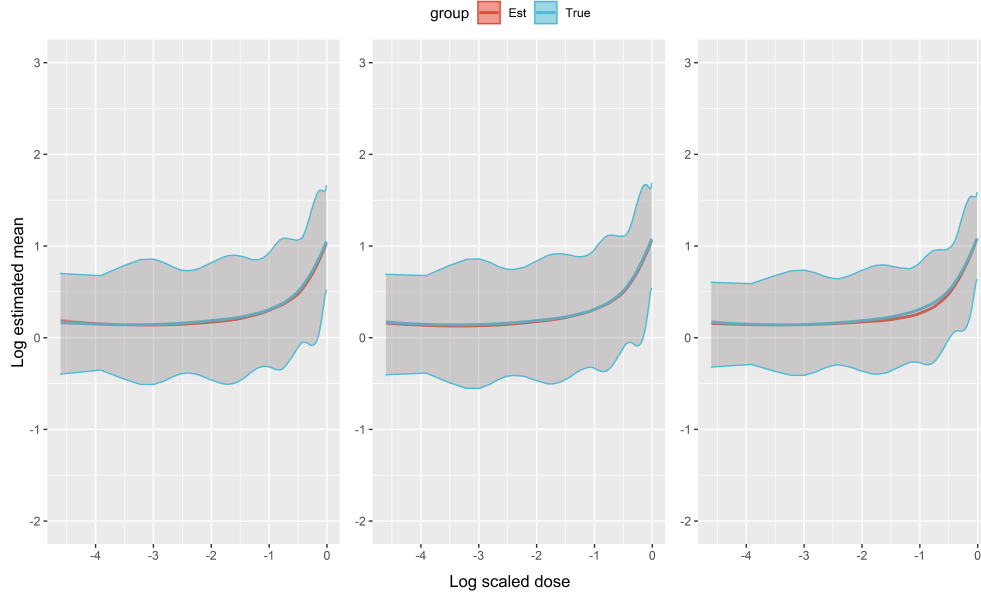


Figure 3.5 Results of the simulation study, illustrating the performance of Model 2 (see Simulation Design) in 500 replicates for 300 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.

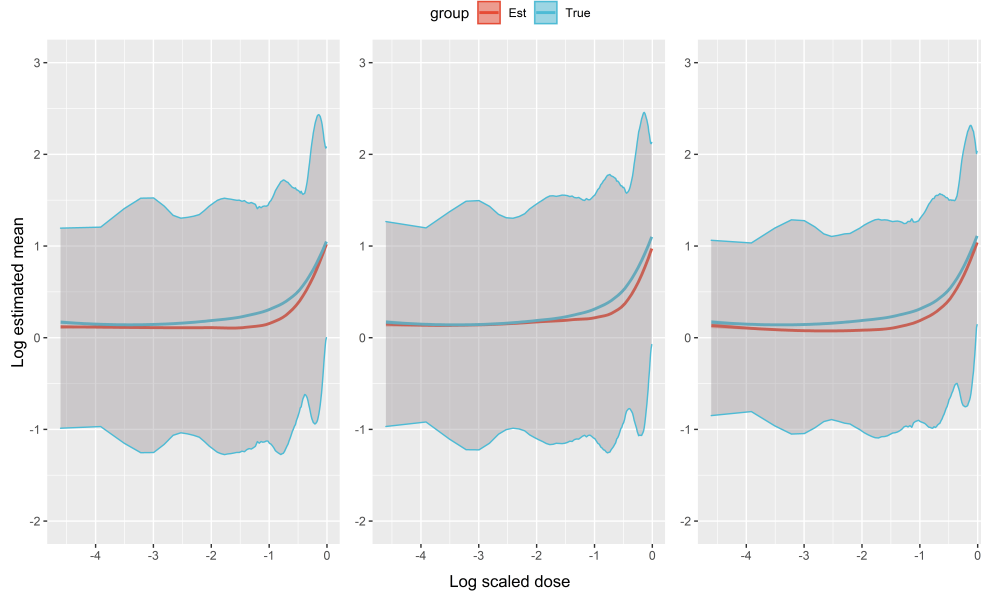


Figure 3.6 Results of the simulation study, illustrating the performance of Model 2 (see Simulation Design) in 500 replicates for 100 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.

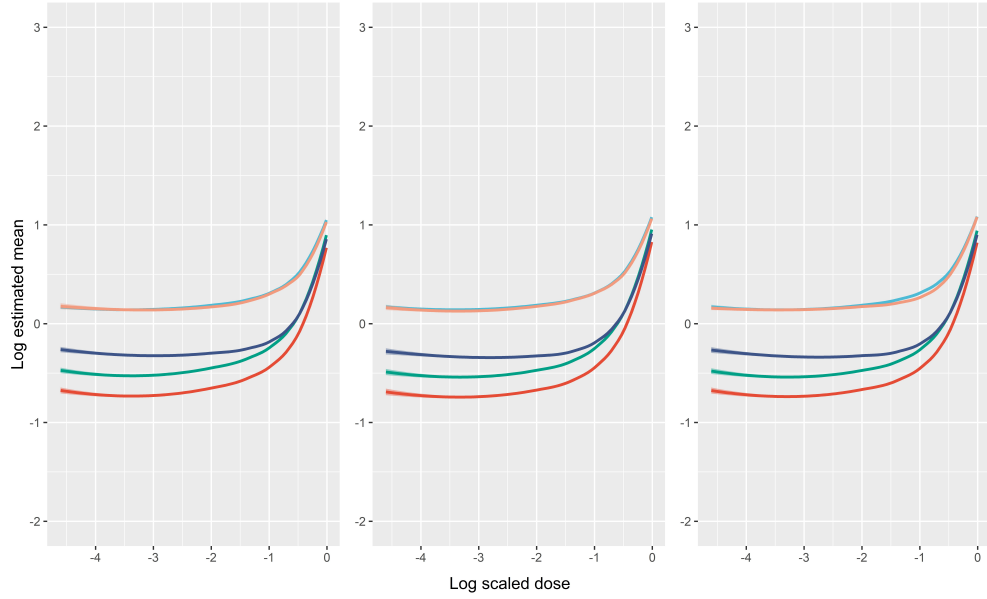


Figure 3.7 Results of the simulation study, illustrating the performance of Model 2 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to the three different cell-types. Continuous cyan, peach, darkblue, darkgreen and red lines represent the true simulated curves and the estimated log-mean curve averaged across 500 random replicates for ZINB-SPL, ZINB-GAM-Dose, ZINB-GAM-Int and NB-GAM models respectively

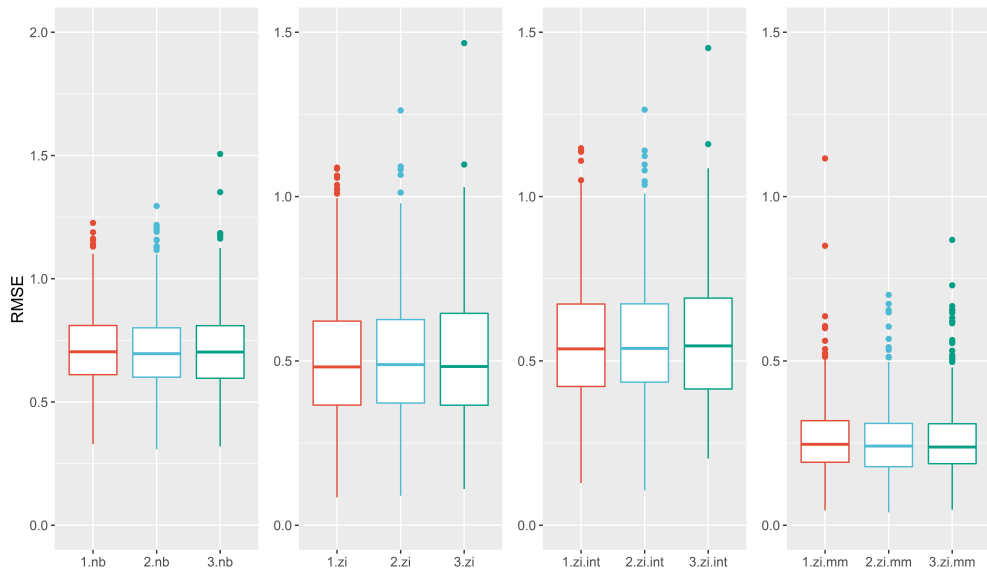


Figure 3.8 Results of the simulation study, illustrating the estimated RMSE of Model 2 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to RMSE boxplots of the NB-GAM, ZINB-GAM-Dose, ZINB-GAM-Int and ZINB-SPL models respectively for the three different cell-types plotted over 500 replicates.

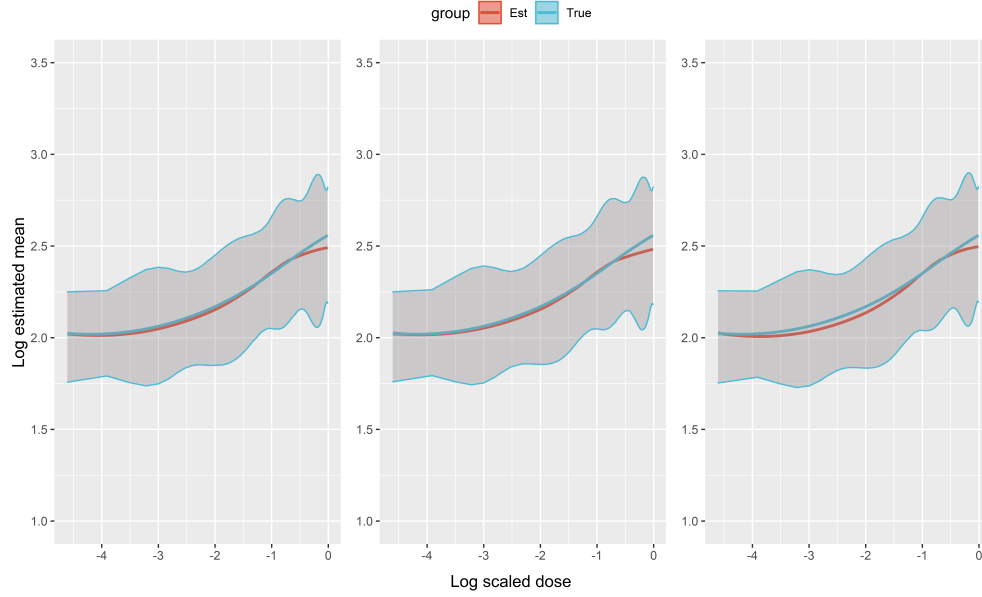


Figure 3.9 Results of the simulation study, illustrating the performance of Model 3 (see Simulation Design) in 500 replicates for 300 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.

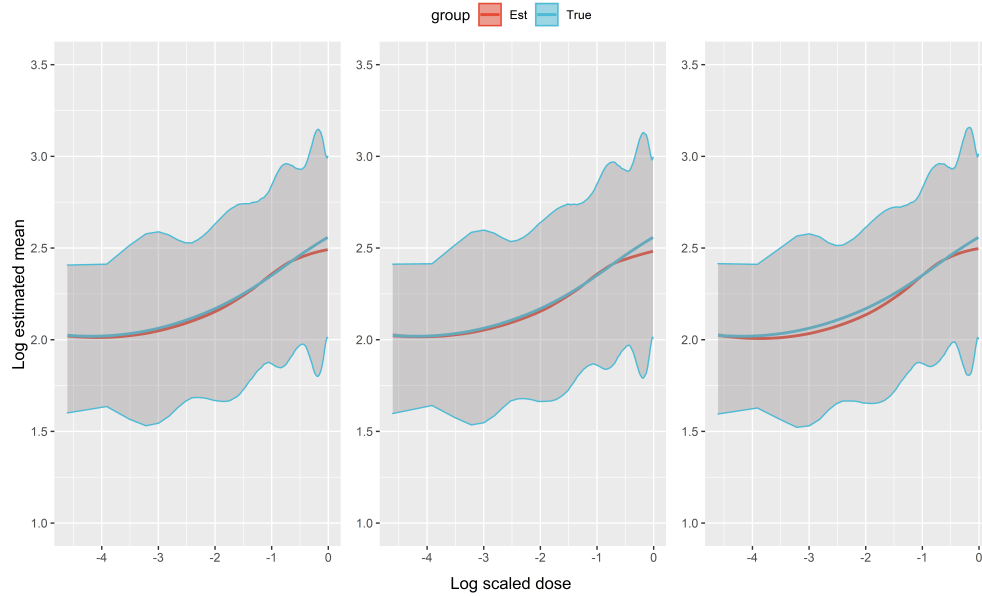


Figure 3.10 Results of the simulation study, illustrating the performance of Model 3 (see Simulation Design) in 500 replicates for 100 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.

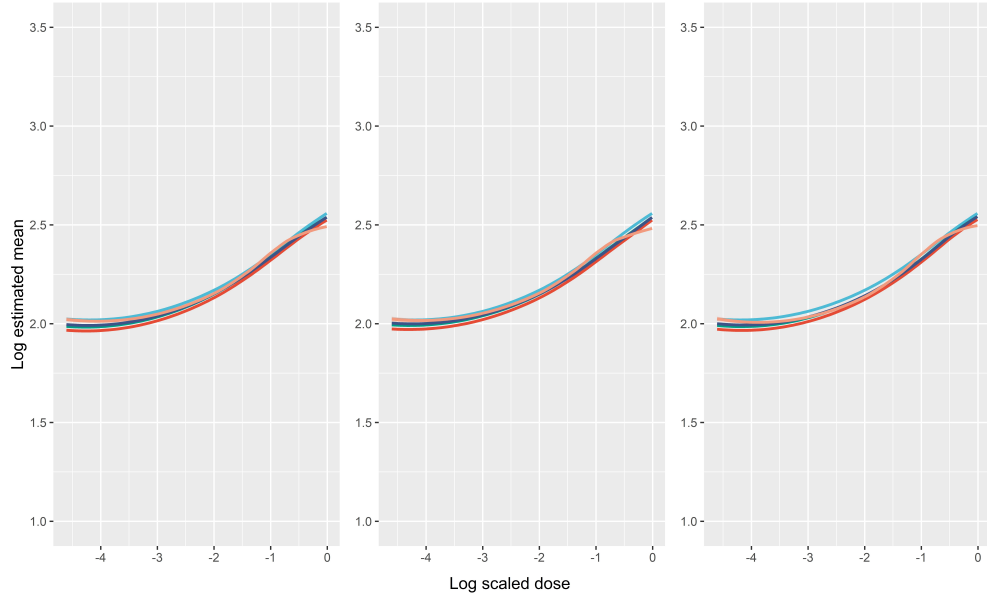


Figure 3.11 Results of the simulation study, illustrating the performance of Model 3 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to the three different cell-types. Continuous cyan, peach, darkblue, darkgreen and red lines represent the true simulated curves and the estimated log-mean curve averaged across 500 random replicates for ZINB-SPL, ZINB-GAM-Dose, ZINB-GAM-Int and NB-GAM models respectively

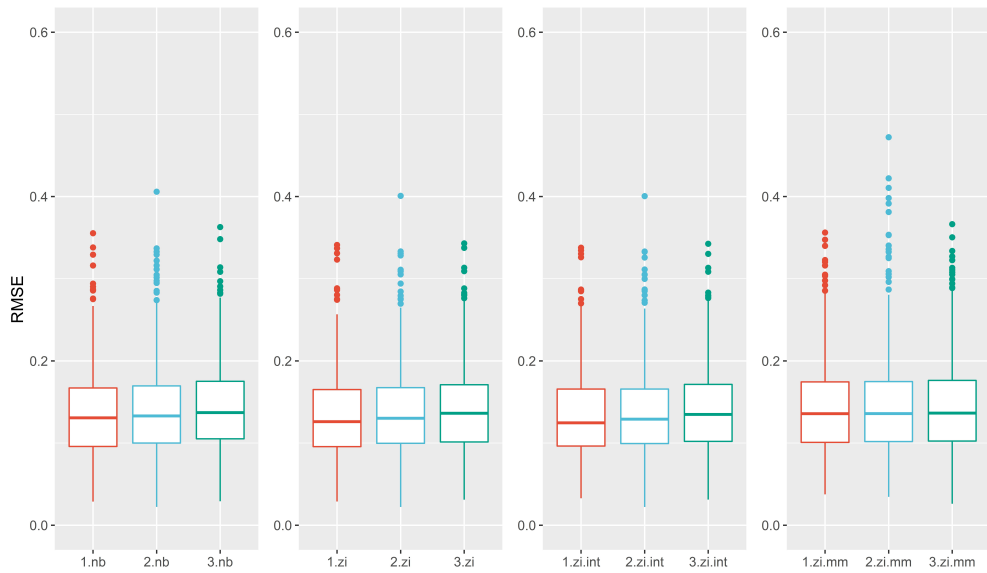


Figure 3.12 Results of the simulation study, illustrating the estimated RMSE of Model 3 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to RMSE boxplots of the NB-GAM, ZINB-GAM-Dose, ZINB-GAM-Int and ZINB-SPL models respectively for the three different cell-types plotted over 500 replicates.

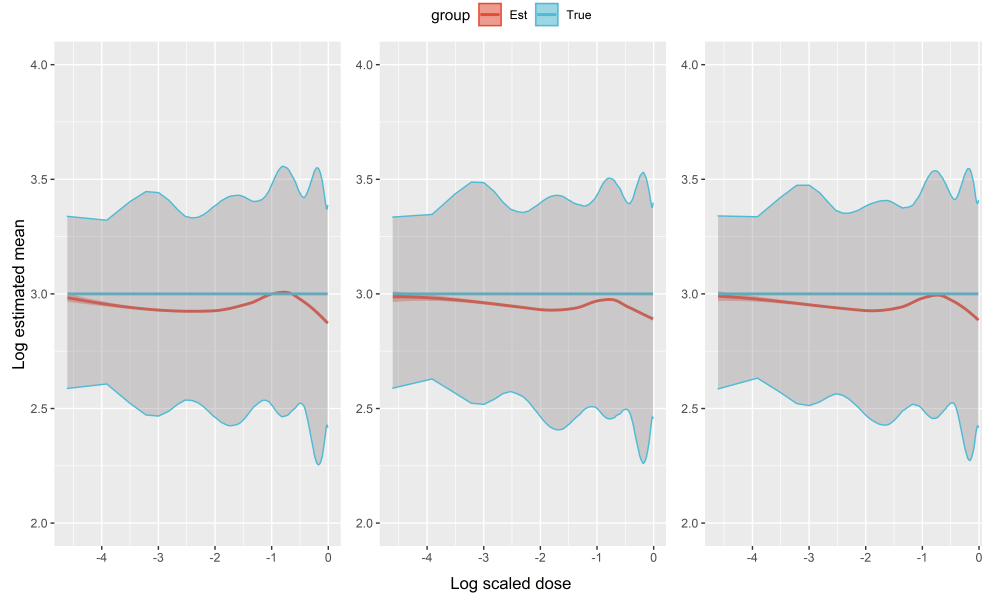


Figure 3.13 Results of the simulation study, illustrating the performance of Model 4 (see Simulation Design) in 500 replicates for 100 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.

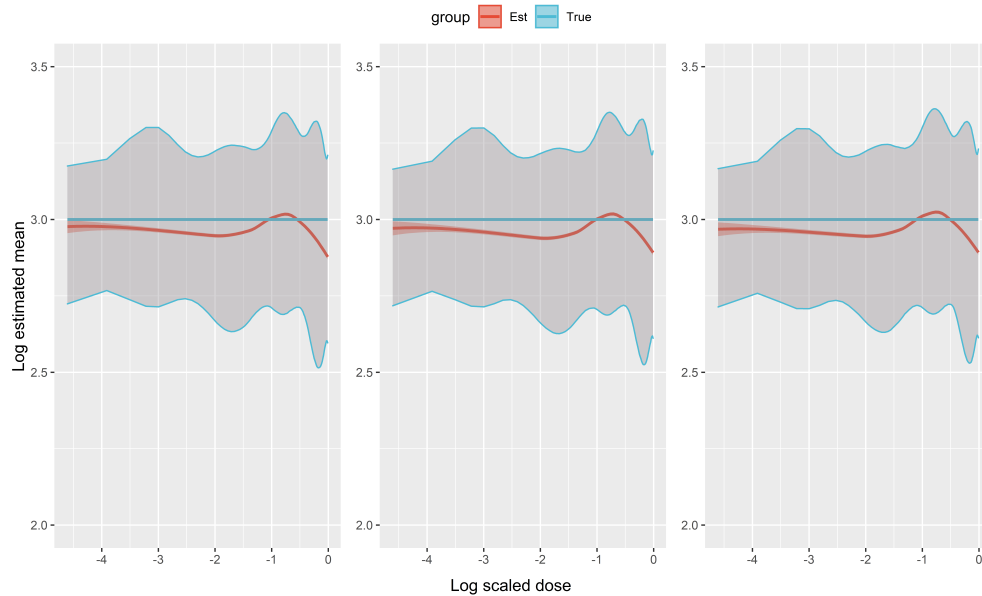


Figure 3.14 Results of the simulation study, illustrating the performance of Model 4 (see Simulation Design) in 500 replicates for 300 sample size. The columns correspond to the three different cell-types. Continuous red and blue lines and the shaded grey region represent the log-mean curve averaged across 500 random replicates, the true simulated curves, and the 95% pointwise confidence intervals. respectively.

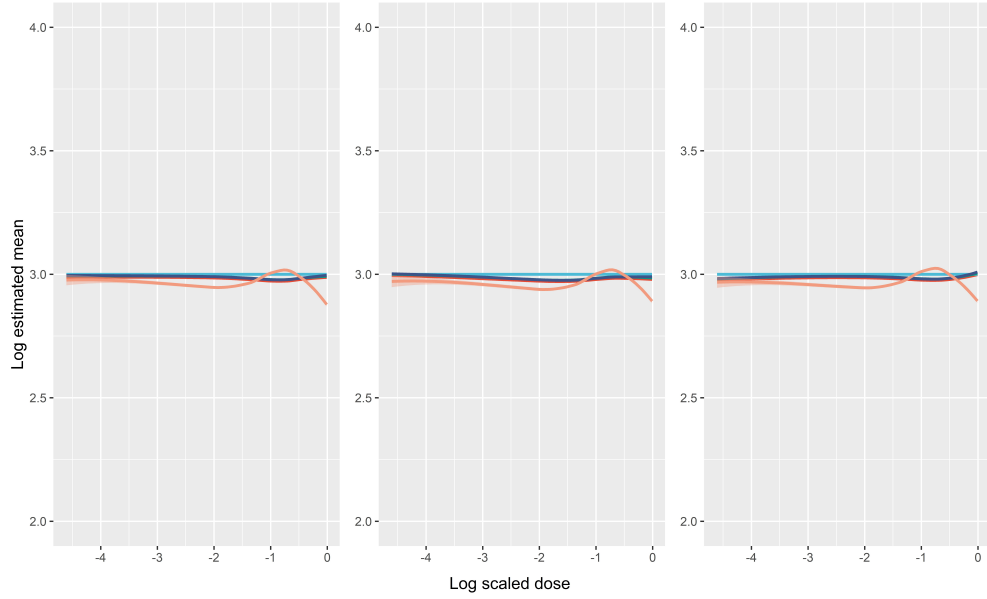


Figure 3.15 Results of the simulation study, illustrating the performance of Model 4 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to the three different cell-types. Continuous cyan, peach, darkblue, darkgreen and red lines represent the true simulated curves and the estimated log-mean curve averaged across 500 random replicates for ZINB-SPL, ZINB-GAM-Dose, ZINB-GAM-Int and NB-GAM models respectively

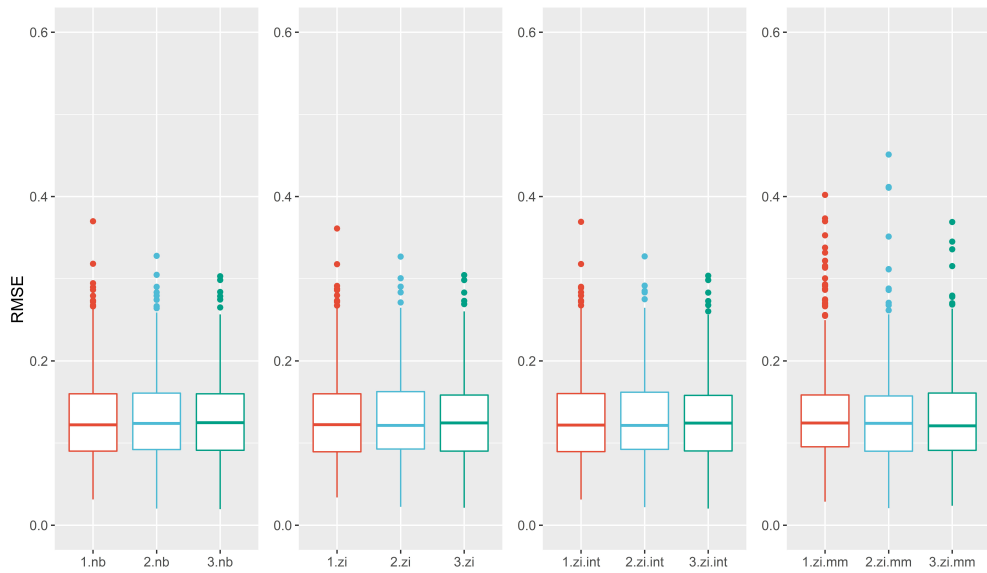


Figure 3.16 Results of the simulation study, illustrating the estimated RMSE of Model 4 (see Simulation Design) in 500 replicates for sample size 300. The columns correspond to RMSE boxplots of the NB-GAM, ZINB-GAM-Dose, ZINB-GAM-Int and ZINB-SPL models respectively for the three different cell-types plotted over 500 replicates.

Scenario	Model	$n_i = 300$		Bias	MSE	Cov	$n_i = 100$		Bias	MSE	Cov
		$\hat{\psi}_1$	SE				$\hat{\psi}_1$	SE			
Case 1	Full	0.081	0.166	-0.019	0.025	0.956	0.101	0.294	0.099	0.009	0.964
	Intercept	0.000	0.1658	-0.099	0.094	1	0.000	0.291	-0.09	0.009	1
Case 2	Full	0.109	0.2607	0.0089	0.010	0.96	0.0438	0.572	-0.056	0.324	0.892
	Intercept	0.000	0.148	-0.100	0.057	0.998	0.000	0.556	-0.100	0.0100	0.992
Case 3	Full	0.113	0.152	0.013	0.006	0.956	0.135	0.262	0.034	0.038	0.988
	Intercept	0.000	0.150	-0.100	0.010	1	0.000	0.318	-0.100	0.010	1
Case 4	Full	0.107	0.154	0.007	0.002	0.952	0.111	0.264	0.012	0.014	0.964
	Intercept	0.000	0.164	-0.100	0.010	1	0.000	0.288	-0.100	0.0100	1

Table 3.4 Parameter estimate, asymptotic standard error, bias and mean squared error(MSE)of parameter ψ_1 under full model and the intercept model. The results are averaged over 500 replicates and reported for $n_i = 100, 300$

3.5 Discussion

In this chapter we present the first statistically rigorous approach for modelling zero-inflated single cell dose response data. Our method proposes a semi-parametric regression framework for the estimation of several non-linear cell-type specific curves arising naturally in single cell dose response experiments. We discuss an extremely flexible penalized regression approach along with further extensions to estimation under monotonicity constraints. Our novel technique flexibly models cell-type specific heterogeneity and its use of a ZINB-Bspline model allows it to effectively capture diverse DR functions and accommodate undesirable zero inflation in data. Further, to better deal with a large number of parameters, arising due to the joint modelling of multiple cell-type specific DR curves, we propose a highly scalable and computationally efficient MM algorithm. Through our MM estimator we are able to significantly reduce the computational complexity of our proposed optimization problem. Finally we propose two highly relevant hypothesis testing approaches that test of presence of DR signal along with differences in the shape of the functional form for different cell-types.

Comprehensive studies on simulated datasets confirm that our proposed model ZINB-SPL yields better RMSE than the competing methods (ZINB-GAM-Dose, ZINB-GAM-Int and NB-GAM). Further, ZINB-SPL displayed greater estimation accuracy for constant, linear and non-linear curve shapes. In case of the highly non-linear shapes, the performance of

ZINB-SPL showed a clear improvement over the competing methods. The consistent performance of ZINB-SPL in four realistic simulation scenarios highlighted the robustness of our proposed approach in modeling dose response relationships under varying characteristics of scRNA-seq datasets. Additionally, the hypothesis testing results for the simulation scenarios demonstrated effective false positive rate control while maintaining considerable power for testing against local alternatives.

An advantage of our proposed MM framework is that it is highly scalable and can provide computationally efficient estimation for a large number of parameters in the model, which could be the case if many cell-types are modelled jointly. In addition, it is straightforward to generalize the proposed framework to monotonicity constrained estimation which may be of interest in several biomedical problems. Thus we provide the first framework for statistically modelling single cell dose response datasets along with appropriate machinery for testing differences between cell-type specific toxicological responses.

CHAPTER 4

KERNELIZED SIGNED GRAPH LEARNING FOR SINGLE CELL GENE REGULATORY NETWORK INFERENCE

4.1 Single Cell Gene Regulatory Networks

Gene regulatory networks (GRNs) represent fundamental molecular regulatory interactions among genes that establish and maintain all required biological functions characterizing a certain physiological state of a cell in an organism [107]. Cell type identity in an organism is determined by how active transcription factors interact with a set of cis-regulatory regions in the genome and controls the activity of genes by either activation or repression of transcription [108]. Usually, the relationship between these active transcription factors and their target genes characterize GRNs. Due to the inherent causality captured by these meaningful biological interactions in GRNs, genome-wide inference of these networks holds great promise in enhancing the understanding of normal cell physiology, and also in characterizing the molecular compositions of complex diseases [109, 110].

GRNs can be mathematically characterized as graphs where nodes represent genes and the edges quantify the regulatory relations. GRN reconstruction attempts to infer this regulatory network from high-throughput data using statistical and computational approaches. Multiple methods encompassing varying mathematical concepts have been proposed during the last decade to infer GRNs using gene expression data from bulk population sequencing technologies, which accumulate expression profile from all cells in a tissue. These methods can be broadly classified into two groups: the first group infers a static GRN, considering steady state of gene expression, while the second group uses temporal measurements to capture the expression profile of the genes in a dynamic process. A thorough evaluation of the static and dynamic models used in bulk GRN reconstruction can be found in [47, 48].

Recent advances in RNA-sequencing technologies have enabled the measurement of gene

expression in single cells. This has led to the development of several computational approaches aimed at quantifying the expression of individual cells for cell-type labelling and estimation of cellular lineages. Several algorithms have been developed to arrange cells in a projected temporal order (pseudotime trajectory) based on similarities in their transcriptional states. In parallel, several dynamic models for single cell GRN reconstruction have also been developed taking into account the estimated pseudotimes. Since single cell network reconstruction algorithms try to establish functional relationships between genes taking into account the entire population of cells, it is debatable as to whether additional knowledge regarding cell state transitions may provide any added benefits [111, 3]. In summary, direct application of bulk GRN reconstruction methods may not be adequate for single cell network inference.

The complex nature of single-cell transcriptomics data pose unique challenges in GRN inference. Changes in gene expression due to cell-cell stochastic variation, cell-cycle heterogeneity and high sparsity due to insufficient sequencing depths and capture inefficiency for genes with low expression form some of the unique characteristics of these datasets [112, 113]. Most importantly the high sparsity/high zero values feature in single cell datasets has garnered a lot of attention and several statistical methods have been designed to particularly model this phenomenon [10, 16, 114]. Recent research has indicated that these zero values referred to as "dropouts" most likely result from biological variation and may be indicative of heterogeneity in gene expression for varying cell types [21, 22].

To account for these unique challenges a variety of algorithms for network reconstruction in scRNAseq data have been recently proposed, but most of these methods fail to outperform network estimation methods developed for bulk data or microarrays. [3, 111]. To that end, we propose a network reconstruction algorithm that learns the co-expression between genes by borrowing ideas from graph signal processing (GSP) literature. GSP provides a framework for analyzing signals defined on graphs by extending classical signal processing tools and concepts [115]. In many applications of GSP, the graph topology is not always available,

thus it must be inferred or learned from the observed data. The major approaches to graph learning (GL) include smoothness based methods [116], where the graph is learned with the assumption that graph signals vary smoothly with respect to graph structure; and diffusion process based models, where the graph is learned from signals that are assumed to be graph filtered versions of random processes [117]. In this work, we focus on learning graphs with the smoothness assumption for the following reasons. First, smooth signals admit low-pass and sparse representations in the graph Fourier domain. Thus, the GL problem is equivalent to finding efficient information processing transforms for graph signals. Second, many graph-based machine learning tasks, such as spectral clustering, graph regularized learning etc., are developed based on the smoothness of the graph signals. Finally, smooth graph signals are observed ubiquitously in real-world applications [117].

Smoothness based GL is first considered in [118] by modeling graph signals using factor analysis, where the transformation from factors to observed signals exploits the graph topology. By imposing a suitable prior on factors, the graph signals are modelled to have low-frequency representation in the graph Fourier domain. This analysis results in an optimization problem where a graph is learned such that variation of signals over the learned graph is minimized. Different variations of this framework with constraints on the learned topology and for handling noisy graph signals were considered in [119, 120, 121, 122, 123]. All of the previous works learn unsigned graphs with the exception of [124], where a signed graph is learned by employing signed graph Laplacian defined by [125]. By using signed Laplacian, [124] aim to learn positive edges between nodes whose signal values are similar and negative edges between nodes whose signal values have opposite signs with similar absolute values. However, this approach is not suitable when graph signals are either all positive- or negative-valued, as in the case of gene expression data.

Considering the advantages of GL approaches in learning graph topologies that are consistent with the observed signals, in this paper, we propose a novel GL algorithm for the reconstruction of GRNs. In particular, we assume gene expression data obtained from cells are

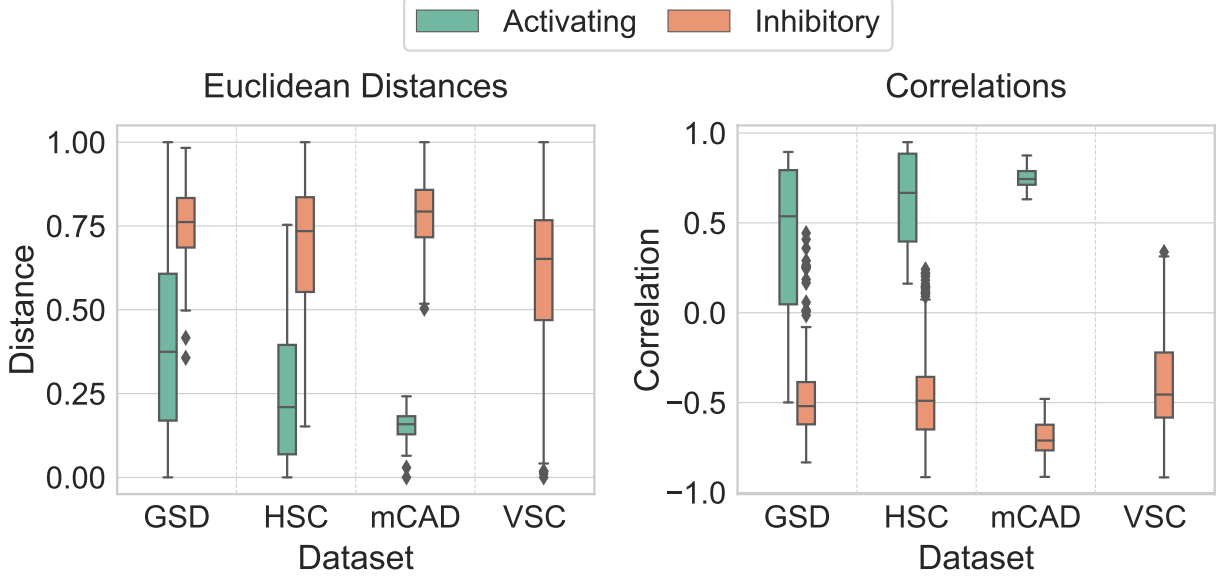


Figure 4.1 Euclidean distances (left, normalized to $[0, 1]$) and correlations (right) between expressions of gene pairs in curated datasets studied in Section 4.4. Values are calculated only for gene pairs that are connected in the ground truth GRNs and they are reported separately for activating and inhibitory edges. Only inhibitory edges are reported for VSC, since its GRN includes only inhibitory edges.

graph signals residing on an unknown graph structure, which corresponds to the GRN. One important characteristic of GRNs is that they are signed graphs, where positive and negative edges correspond to activating and inhibitory regulations between genes. To this end, we propose a novel and computationally efficient signed GL approach, scSGL, that reconstructs the GRN under the assumption that graph signals admit low-frequency representation over activating edges, while admitting high-frequency representation over inhibitory edges. Biologically, this modelling implies that two genes that are connected with an activating edge have similar expressions, while two genes connected with an inhibitory edge have dissimilar expressions. In Figure 4.1, we show how these assumptions hold for curated datasets studied in Section 4.4. The figure shows that Euclidean distances between expressions are smaller for gene pairs connected by activating edges than for those connected by inhibitory edges. The figure also reports correlations between expressions, which indicates that expressions of gene pairs connected with activating and inhibitory edges are positively correlated, i.e., similar,

and negatively correlated, i.e., dissimilar, respectively. We also performed a Wilcoxon Rank Sum test to determine whether the calculated associations for the positive ground truth connections were significantly lower than the associations for the negative ground truth connections for Euclidean distances. We test the null hypothesis, H_0 : the distributions of both populations are equal versus the alternative hypothesis H_a :, the distribution of the negative associations are stochastically greater than the distribution of positive associations. In case of the correlation distances we want to test H_a :the distribution of the positive associations are stochastically greater than the distribution of negative associations. The calculated p-values were all less than 0.01, hence justifying our assumptions for all curated datasets except VSC, which only has negative associations. Another important characteristic of scRNAseq is high proportion of dropouts. We address this issue by employing kernel functions to map graph signals to a higher dimensional space and assuming low- and high-frequency representation for these high dimensional graph signals. This mapping allows us to use kernels that are appropriate for modelling single cell data structures.

4.2 Background

4.2.1 Notations

Define an undirected graph as $G = (V, E)$ where V is the set of nodes with a cardinality of $|V| = n$ and $E \subseteq V \times V$ denotes the edge set with cardinality $|E| = m$. Edges between nodes i and j are denoted by $e_{i,j}$ having a weight $w_{i,j}$. If the weights $w_{i,j}$ are strictly positive then the graph is unsigned, whereas if $w_{i,j}$ allows both positive and negative values then the graph is referred to as signed. Define W to be the $n \times n$ symmetric adjacency matrix of graph G , where $W_{ij} = W_{ji} = w_{ij}$ if $e_{ij} \in E$ and 0, otherwise. In the case of signed graphs, W can be decomposed as $W^+ - W^-$ where $W_{ij}^+ = w_{ij}$ for $w_{ij} > 0$ and $W_{ij}^- = |w_{ij}|$ for $w_{ij} < 0$. Further the Laplacian matrix L is defined as $L = D - W$ where D is the degree matrix of G . Similar to the decomposition of the adjacency matrix, the Laplacian can be decomposed for estimation of signed Laplacians. All one and all zero vectors and matrices are represented

by $\mathbf{1}$ and $\mathbf{0}$, respectively. Finally, i th row and column of a matrix X are represented by X_i and $X_{\cdot i}$, respectively.

4.2.2 Low- and High-frequency Signals on Unsigned Graphs

A graph signal $\mathbf{x} \in \mathbb{R}^n$ is a vector whose entries reside on the nodes of an unsigned graph G . Graph Fourier transform (GFT) of $\mathbf{x}$ is defined as the expansion of $\mathbf{x}$ in terms of the eigenbasis of the graph Laplacian [126]. This representation allows us to characterize $\mathbf{x}$ in terms of its graph spectral content as either low- or high-frequency, where low(high)-frequency graph signals have small (large) variation with respect to the graph [127].

Let $\mathbf{L} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ be the eigendecomposition of $\mathbf{L}$ where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues with $\Lambda_{ii} = \lambda_i$ and $\mathbf{V}_{\cdot i}$ is the eigenvector corresponding to λ_i . GFT of $\mathbf{x}$ is then $\hat{\mathbf{x}} = \mathbf{V}^\top \mathbf{x}$ and inverse GFT is [126]:

$$\mathbf{x} = \mathbf{V}\hat{\mathbf{x}} = \sum_{i=1}^n \hat{x}_i \mathbf{V}_{\cdot i}. \quad (4.1)$$

Thus, $\mathbf{x}$ is the linear combination of eigenvectors of $\mathbf{L}$ with the coefficients equal to the entries of $\hat{\mathbf{x}}$. Eigenvectors of $\mathbf{L}$ corresponding to small eigenvalues have small variation over the graph. Thus, if most of the energy of $\hat{x}$ lies in $\hat{x}_i$ s corresponding to the small eigenvalues, then $\mathbf{x}$ varies little over G . On the other hand, if most of the energy of $\hat{x}$ lies in $\hat{x}_i$ s corresponding to the large eigenvalues, $\mathbf{x}$ has high variation over G . The total variation of $\mathbf{x}$ over G is then quantified as [126]:

$$\text{trace}(\hat{\mathbf{x}}^\top \mathbf{\Lambda} \hat{\mathbf{x}}) = \text{trace}(\mathbf{x}^\top \mathbf{V} \mathbf{L} \mathbf{V}^\top \mathbf{x}) = \text{trace}(\mathbf{x}^\top \mathbf{L} \mathbf{x}), \quad (4.2)$$

which is small for low-frequency graph signals and large for high-frequency ones.

4.2.3 Unsigned Graph Learning

An unknown unsigned graph G can be learned from a set of observed graph signals defined over it with the assumption that graph signals have low-frequency representation in graph

spectral domain, i.e., total variation is small. Using this assumption, [118] proposes to learn G by minimizing (4.2) with respect to $\mathbf{L}$ given a set of graph signals $\{\mathbf{x}_i\}_{i=1}^p$ as follows:

$$\min_{\mathbf{L} \in \mathbb{L}} \text{trace}(\mathbf{X}^\top \mathbf{L} \mathbf{X}) + \alpha \|\mathbf{L}\|_F^2 \quad \text{s.t.} \quad \text{trace}(\mathbf{L}) = 2n, \quad (4.3)$$

where $\mathbb{L} = \{\mathbf{L} : L_{ij} = L_{ji} \leq 0 \, \forall i \neq j, \mathbf{L}\mathbf{1} = \mathbf{0}\}$ is the set of Laplacian matrices and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the data matrix whose columns are graph signals. The constraint $\text{trace}(\mathbf{L}) = 2n$ ensures that trivial solutions are avoided. $\|\mathbf{L}\|_F$ is employed to control the sparsity of the learned graph as discussed in [119]. In particular, the Frobenius norm penalizes $\mathbf{L}$ from having entries with large absolute values, and the non-positivity constraint ($L_{ij} = L_{ji} \leq 0$) and degree constraint ($\text{trace}(\mathbf{L}) = 2n$) threshold edges with small weights. Thus, as $\alpha \rightarrow \infty$ edges of the learned graph start having very similar weights and no thresholding can be performed to sparsify the learned graph. On the other hand, small values of α leads $\mathbf{L}$ to have entries with varying values and the imposed constraints threshold the small values yielding a sparse graph.

4.2.4 Kernels

Traditional machine learning and signal processing applications are mostly developed based on linear modelling due to their simplicity. However, real world problems require nonlinear estimation that can detect more complex patterns in the data. For this purpose, kernels are introduced to capture the nonlinearity by mapping signals to a high-dimensional space [128]. Kernels correspond to dot products in a higher dimensional feature space and overcome explicit construction of the feature space; thus providing simplicity of linear methods in nonlinear estimation. Given data from input space $\mathcal{X}$, and a mapping function $\phi : \mathcal{X} \rightarrow \mathcal{H}$ where $\mathcal{H}$ is an Hilbert space, a kernel function can be expressed as an inner product in the corresponding feature space, i.e., $\kappa(\mathbf{x}_i, \mathbf{x}'_i) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}'_i) \rangle$, where $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a finitely positive semi-definite kernel function [129]. An explicit representation of the feature map ϕ is not necessary and the dimension of mapped feature vectors could be high and even

infinite. By using different kernels, learning algorithm can be augmented to exploit various (nonlinear) associations between input data. For example, the first term in (4.3) can be rewritten as $\text{trace}(\mathbf{X}\mathbf{X}^\top\mathbf{L}) = \sum_{i,j} \langle \mathbf{X}_i, \mathbf{X}_j \rangle L_{ij}$, where the dot product between the rows of $\mathbf{X}$ can be replaced by $\kappa(\mathbf{X}_i, \mathbf{X}_j)$. In the next section, this observation is used to develop a graph learning framework that is able to capture nonlinear relations between graph signals.

4.3 Methods

4.3.1 Signed Graph Learning

In (4.3), an unsigned graph is learned with the assumption that the observed graph signals have low-frequency representation in graph spectral domain. In order to learn a signed graph G , one needs to make some additional assumptions about the graph signals $\mathbf{X}$. In this work, we make the following assumptions:

1. Signal values on nodes connected by positive edge values are similar to each other, i.e., variation over positive edges is small.
2. Signal values on nodes connected by negative edge values are dissimilar to each other, i.e., variation over negative edges is large.

From GSP perspective, these assumptions correspond to graph signals being low- and high-frequency over positive and negative edges, respectively. Let G^+ be the graph corresponding to the positive edges of G and let G^- be the graph corresponding to the negative edges of G with the edge weights equal to the absolute value of the original edge values. Assumption 1 implies that the graph signals have low-frequency representation in the graph Fourier domain of G^+ . On the other hand, assumption 2 implies that the graph signals have high-frequency representation in graph Fourier domain of G^- . We use (4.2) to quantify how well the graph signals fit these assumptions. Thus, to learn an unknown signed graph, we minimize $\text{trace}(\mathbf{X}^\top\mathbf{L}^+\mathbf{X})$ with respect to $\mathbf{L}^+$ while maximizing $\text{trace}(\mathbf{X}^\top\mathbf{L}^-\mathbf{X})$ with respect

to $\mathbf{L}^-$:

$$\begin{aligned}
& \min_{\mathbf{L}^+, \mathbf{L}^- \in \mathbb{L}} \text{trace}(\mathbf{X}^\top \mathbf{L}^+ \mathbf{X}) - \text{trace}(\mathbf{X}^\top \mathbf{L}^- \mathbf{X}) + \alpha_1 \|\mathbf{L}^+\|_F^2 + \alpha_2 \|\mathbf{L}^-\|_F^2 \\
& \text{s.t.} \quad \text{trace}(\mathbf{L}^+) = 2n, \text{trace}(\mathbf{L}^-) = 2n \\
& \quad L_{ij}^+ = 0 \text{ if } L_{ij}^- \neq 0 \text{ and } L_{ij}^- = 0 \text{ if } L_{ij}^+ \neq 0 \quad \forall i \neq j,
\end{aligned} \tag{4.4}$$

where Frobenius norms and the first two constraints are similar to (4.3) and the last constraint ensures that L^+ and L^- are not non-zero for the same indices.

4.3.2 Kernelized Signed Graph Learning

As mentioned in Section 4.2.4, kernels are used in learning algorithms to exploit various (nonlinear) relations between input data. This is especially crucial in GRN inference as shown in [130], where 17 different association measures between gene expressions are compared in terms of their performance in GRN inference and various other tasks on single-cell transcriptomic datasets. In this paper, we consider three kernels: correlation coefficient, r , measure of proportionality, ρ [131] and a modification of Kendall’s tau (τ_{zi}) for zero inflated non-negative continuous data [132]. These kernels are selected because r is a commonly used measure for network inference, ρ performs the best in [130] and τ_{zi} can handle high ratio of dropouts in scRNAseq.

As mentioned in Section 4.2.4, kernels provide an efficient way of capturing the non-linear associations between input data samples. This is especially crucial in building single cell GRN learning algorithms due to the unique statistical properties exhibited by single cell data. To determine optimal measures of association in single cell network learning, [130] evaluated 17 association measures in terms of their ability to reconstruct gene and cellular networks from single cell transcriptomic datasets. Two measures, ρ [131], a measure of association for compositional data and τ_{zi} , a measure of association for zero inflated non-negative continuous data [132] are shown to perform consistently better in all learning scenarios investigated in [130]. The strong performance of ρ can be explained on the basis

that scRNA-seq captures only a small proportion of messenger RNA in each cell and therefore gene expression measurements can be viewed as relative measures of abundance (as seen in compositional data). On the other hand, τ_{zi} , a modification of Kendall’s rank correlation coefficient, is expected to provide less biased estimates of association in the setting of zero-inflated continuous data, a characteristic of single cell transcriptomic datasets [132]. To compare and contrast these two measures, the correlation kernel r is additionally investigated since it’s widely used in GRN reconstruction algorithms.

In its current form, (4.3) cannot be used directly for different associations. Thus, the optimization problem in (4.4) is extended using kernels. The first term in (4.4) can be written as $\text{trace}(\mathbf{X}\mathbf{X}^\top \mathbf{L}^+) = \sum_{i,j} \langle \mathbf{X}_i, \mathbf{X}_j \rangle L_{ij}^+$ and the second term can be written similarly. By replacing dot products with a given kernel function, i.e., $\kappa(\mathbf{X}_i, \mathbf{X}_j)$, the problem in (4.4) can be extended to incorporate the different associations as:

$$\begin{aligned}
& \min_{\mathbf{L}^+, \mathbf{L}^- \in \mathbb{L}} \quad \text{trace}(\mathbf{K}\mathbf{L}^+) - \text{trace}(\mathbf{K}\mathbf{L}^-) + \alpha_1 \|\mathbf{L}^+\|_F^2 + \alpha_2 \|\mathbf{L}^-\|_F^2 \\
& \text{s.t.} \quad \text{trace}(\mathbf{L}^+) = 2n, \text{trace}(\mathbf{L}^-) = 2n \\
& \quad L_{ij}^+ = 0 \text{ if } L_{ij}^- \neq 0 \text{ and } L_{ij}^- = 0 \text{ if } L_{ij}^+ \neq 0 \quad \forall i \neq j,
\end{aligned} \tag{4.5}$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the kernel matrix with $K_{ij} = \kappa(\mathbf{X}_i, \mathbf{X}_j)$. From GSP perspective, this modification implies that graph signals on each node, i.e., $\mathbf{X}_i$, are first mapped to a (higher dimensional) Hilbert space and the signed graph is learned in this new space. Namely, let $\Phi \in \mathbb{R}^{n \times \hat{p}}$ be the matrix constructed from mapping $\mathbf{X}_i$ ’s to the Hilbert space $\mathcal{H}$ with dimension $\hat{p}$ where rows of Φ are $\phi(\mathbf{X}_i)$. When learning unknown signed graph G with a kernel, each column of Φ is a graph signal over G and they are assumed to have low- and high-frequency representation with respect to G^+ and G^- , respectively. Extending signed graph learning problem in (4.4) using kernels brings flexibility and any association metric in [130] can be implemented in this framework if it is a positive semi-definite kernel. The optimization procedure for 4.5 is given in the Supplementary Material.

4.3.3 Hyperparameter Selection

The optimization problem in 4.5 requires the selection of two regularization parameters α_1 and α_2 , which determine the density of the learnt graph, i.e., large values of α_1 (α_2) result in denser $\mathbf{L}^+$ ($\mathbf{L}^-$). Their values can be set to obtain a graph with desired positive and negative edge densities. Next we illustrate the algorithm used for generating realistic single cell simulations.

1. Given a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ whose columns are graph signals, we randomly shuffle each column of the matrix k times creating k surrogate data matrices.
2. Association between rows of the surrogate data matrices are calculated by the kernel employed in (4.5).
3. Thresholds λ_1 and λ_2 are selected as the p th and $(100 - p)$ th percentiles of the values in the association matrix calculated in Step 2.
4. Steps (1-3) are repeated k times to construct the empirical distribution of the thresholds λ_1 and λ_2 .
5. Finally, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are selected to be the medians of the empirical distributions constructed in Step 4.
6. The association matrix for the original data $\mathbf{X}$ is constructed.
7. The number of entries in the association matrix that are smaller than $\hat{\lambda}_1$ are determined and normalized by the total number of entries in the association matrix to obtain the density of $\mathbf{L}^-$. Similarly, number of entries in the association matrix greater than $\hat{\lambda}_2$ is used to determine the density of $\mathbf{L}^+$.
8. Values of α_1 and α_2 are then selected to learn graphs with the estimated graph densities found in Step 7. Since the density of the positive (negative) graph increases monoton-

ically with the value of α_1 (α_2), bisection search is used to determine the values of α_1 and α_2 that give the desired densities.

4.3.4 Generation of simulated datasets from zero-inflated negative binomial distribution

In this section, we outline the algorithm that was used to generate the simulated datasets for parameter sensitivity analysis.

1. For each simulation setting, we first generated a binary association graph $G_{j_1 j_2}$, $\forall j_1, j_2 \in \{1, \dots, p\}$ using either of the three graph topologies random, hub and cluster.
2. A binary indicator $I_{j_1 j_2}$ was next sampled for each entry of the association graph with $I_{j_1 j_2} \sim \text{Bernoulli}(0.5)$.
3. Given the binary association graph a weight matrix was generated as:

$$W_{j_1 j_2} = \begin{cases} G_{j_1 j_2} \text{Unif}(0.3, 0.7) & \text{if } I_{j_1 j_2} = 0, \\ G_{j_1 j_2} \text{Unif}(-0.7, -0.3) & \text{if } I_{j_1 j_2} = 1. \end{cases}$$

4. Random samples of n multivariate Gaussian random variables were then generated, with known weight matrix $W_{j_1 j_2}$. The random sample was denoted as $(X_1, \dots, X_p)$, where each variable (gene vector) $X_j = (X_{j1}, \dots, X_{jn})^T$ consisted of n realizations.
5. To mimic the dropout phenomenon present in real scRNAseq datasets, we next introduced additional zeros to the gene expression matrix. Following [13], the dropout probability for each row (gene vector) in the gene expression matrix $\mathbf{X}$ was calculated as: $p_{ji} = \exp(-\rho X_{ji}^2)$, where ρ represents the exponential decay parameter that controls the dependence between the dropout probability and gene expression.

6. A binary indicator was next sampled for each entry: $\eta_{ji} \sim \text{Bernoulli}(p_{ji})$, with $\eta_{ji} = 1$ indicating that the corresponding entry of X_{ji} would be replaced by 0. The dropout probability for each gene vector was calculated as $\omega_j = \sum_{i=1}^n \eta_{ji}$.
7. Using a modification of the NORTA (Normal to Anything) method [133] we generated samples from a multivariate zero inflated negative binomial distribution based on the multivariate normal samples generated in Step 4 using mean, dispersion and zero-inflation parameters λ, k and ω_j s.
8. To mirror real scRNA-seq gene expression data behaviour, the gene expression mean λ and standard deviation k were estimated from a real scRNA-seq dataset, Peripheral Blood Mononuclear Cells (PBMC) freely available from 10X Genomics.

4.3.5 Performance Metrics

4.3.5.1 AUPRC and AUROC:

Area under precision and recall curve and area under the receiver operating characteristic are calculated by comparing inferred graphs to ground truth gene regulations. During this calculation, signs of the learned edges are ignored as the AUPRC and AUROC are performance metrics restricted to binary classification. In particular, we first take the absolute value of edge weights and then compare them to ground truth edges. Thus, these metrics indicate how well methods detect edges without considering the signs of the inferred edges. Ground truth networks are considered as undirected and self-loops are ignored. Following [3], we also defined *AUPRC ratio* and *AUROC ratio* as the ratio of AUPRC (AUROC) value of the methods to AUPRC (AUROC) of the random estimator.

4.3.5.2 AUPRC Activating/Inhibitory:

One of our goals is to learn whether the edges are activating or inhibitory. AUPRC as defined above cannot evaluate the sign information. Thus, for curated datasets, whose ground truth

gene regulations include signed edge information, we calculate AUPRC for activating and inhibitory edges separately. In particular, for methods that learn signed graphs we compare the learned positive edges to activating edges in ground truth and learned negative edges to inhibitory edges in the ground truth. For methods that do not learn signed edges, we evaluate the inferred edges with respect to the ground truth activating and inhibitory edges separately to calculate two AUPRC values.

4.3.5.3 EPR

Early precision ratio is the fraction of true positives in the top- k edges in the inferred graphs where k is the number of edges in the ground truth network [3]. For methods that return signed edges, we found top- k edges after taking absolute value of the edge weights, thus this metric is used for edge detection performance rather than the detection of edge signs. Ground truth networks are considered to be directed graphs and self-loops are ignored. Finally, *EPR ratio* is defined as the ratio of the EPR values of the methods to the EPR value of the random estimator.

4.4 Results

In this section, performance of scSGL is evaluated and compared to state-of-the-art GRN inference methods on various simulated and experimental scRNAseq datasets. We selected GENIE3 [52], GRNBOOST2 [53], PIDC [134] and PPCOR [50] for comparison as they are the top performing methods in [3]. GENIE3, GRNBOOST2 and PPCOR were originally developed for bulk analysis, while PIDC is developed for single cell gene expression data. Among these methods, GENIE3 and GRNBOOST2 return fully connected directed networks, while the remaining two infer undirected networks. Finally, only PPCOR algorithm returns signed graphs. Given the inherent sparsity of gene networks, we used the area under the precision-recall curves (AUPRC) ratio as the primary evaluation metric. Supplementary Material also includes results using area under the receiver operating characteristic curves

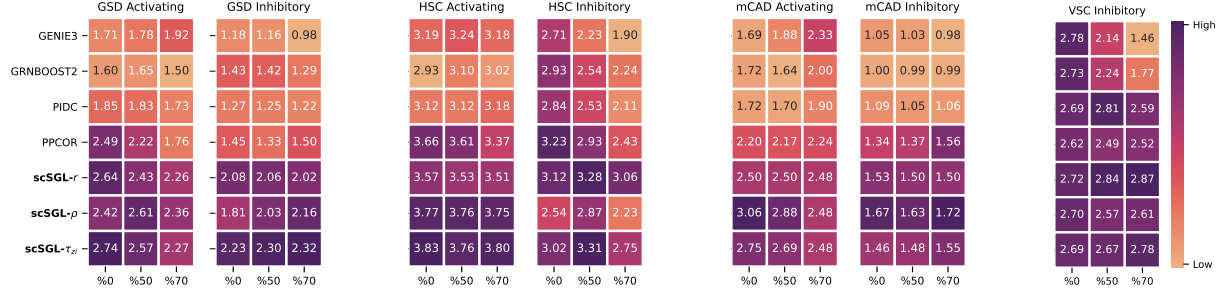


Figure 4.2 Performance of scSGL and state-of-the-art methods on curated datasets as measured by AUPRC for activating and inhibitory edges. x-axis indicates dropout ratio in the dataset.

(AUROC) ratio and early precision ratio (EPR) ratio, whose overall results are consistent with the following observations made with AUPRC ratio.

4.4.1 Synthetic Datasets

Curated Datasets From BEELINE: The first simulation datasets we consider are curated from "published Boolean models of GRNs" [3]. These datasets were generated using the recently proposed single cell GRN simulator BoolODE [3]. BoolODE converts boolean functions specifying a GRN directly to ODE equations using

GeneNetWeaver [135, 136], a widely used method to simulate bulk transcriptomic data from GRNs. These datasets are generated from four literature-curated Boolean models: mammalian cortical area development (mCAD), ventral spinal cord (VSC) development, hematopoietic stem cell (HSC) differentiation and gonadal sex determination (GSD). These models represent different types of graph structures, with varying numbers of positive and negative edges; thus serving as good examples for illustrating the robustness of the proposed method in modelling signed graph topologies. BoolODE is used to create ten random simulations of the synthetic gene expression datasets with 2,000 cells for each model. For each dataset, one version with a dropout rate of 50% and another with a rate of 70% are also considered to evaluate the performance of the methods under missing values.

AUPRC ratios are calculated separately for activating and inhibitory edges and their

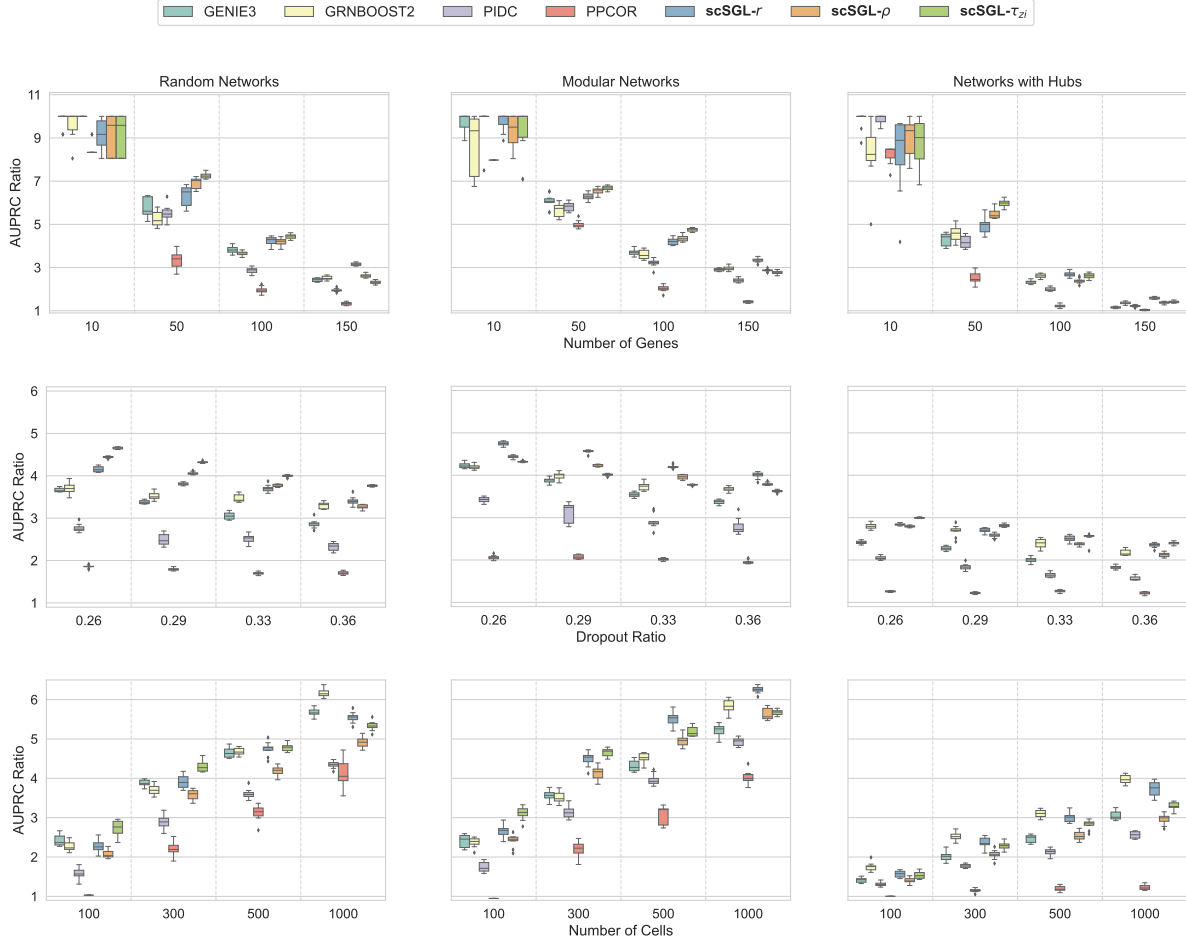


Figure 4.3 Performance of various methods for synthetic datasets with varying number of genes (top row), dropout ratio (middle row) and number of cells (bottom row).

average over realizations are reported in Figure 4.2. For most of the datasets, scSGL performs better than other benchmarking methods in inferring both activating and inhibitory edges. Although there is a difference between the performances of different kernels, scSGL generally performs better than state-of-the-art methods irrespective of the selected kernel. Comparing the performances of different kernels, it is observed that τ_{zi} results in higher AUPRC ratios in GSD, HSC and VSC while ρ performs better in mCAD datasets. It is also observed that AUPRC ratios are higher for activating edges than inhibitory edges. Increasing the dropout ratio causes a drop in the performance of all methods for inferring the activating edges but not for learning the inhibitory edges. Overall, the best performing kernel is τ_{zi} , which might be because of its robustness to increasing dropout ratio compared to other kernels.

Parameter Sensitivity Analysis: To mimic the zero inflated and overly dispersed nature of most scRNAseq datasets, we simulated gene expression data from a multivariate zero-inflated negative binomial (ZINB) distribution for our second simulation. These datasets were then used to conduct parameter sensitivity analysis for the proposed methods. Given a known graph structure, synthetic datasets are generated from a ZINB distribution by adapting an algorithm developed by [133]. The three parameters of the ZINB distribution; λ , k and ω , which control its mean, dispersion and degree of zero-inflation, respectively were determined from a real scRNAseq dataset to make the simulations mirror the properties of real datasets.

The ZINB simulator is then used to generate expression data from three different network structures: random networks, networks with a given community structure and networks with hubs. Random networks are generated using Erdős–Rényi model with desired edge density. Since Erdős–Rényi model is not realistic due to its binomial degree distribution, we also consider networks with hubs. These networks are generated using a Barabási–Albert model whose degree distribution follows a power-law function. Finally, networks with community structure, also known as modular networks, are generated using a disjoint union of random graphs. The accuracy of the scSGL inferred graphs were then evaluated for all three graph structures. To investigate the robustness of scSGL, we simulated datasets from the aforementioned network topologies by varying the following parameters: *(i)* number of genes (10, 50, 100 and 250), *(ii)* number of cells (100, 300, 500 and 1000) and *(iii)* dropout probabilities (0.26-0.36). To account for the inherent randomness of the simulations, 10 independent data replicates were generated for every parameter combination and the mean AUPRC ratios obtained by averaging over the replicates are reported in Figure 4.3.

Recent investigations of scRNAseq datasets have revealed that dropout rates are primarily driven by a combination of technical and biological factors [108]. Consequently, while mean gene expression and proportion of zeros are linked, this may vary based on cell type, sex, and other biological and technical factors. While investigating the impact of dropout rates

on network estimation accuracy, we found a steady decline in AUPRC ratios for all methods with an increase in the number of zeroes. scSGL irrespective of the kernel choice maintained the highest AUPRC ratios across all network topologies. Gene expression in scRNAseq datasets can be interpreted as relative measures of abundance owing to the datasets being a combination of gene expression derived from several cell-types. This could be the reason why proportionality measures perform well [130]. The strong performance of τ_{zi} can be explained on the basis that it explicitly models the dropouts present in scRNAseq datasets. Despite the poor performance of regularized correlation networks (PPCOR), we see a strong performance of scSGL when using the correlation kernel. This proves that gene-gene relationships are in fact non-linear in nature. This belief is also strengthened by the above average performance of tree-based machine learning algorithms like GENIE3 and GRNBOOST2. It is to be noted that PIDC, the only other method capable of modelling excess zeroes, while accounting for non-linear relationships fails to achieve a top-ranking AUPRC ratios.

Next, we evaluated the impact of cell sizes on network reconstruction. Figure 2 demonstrates a clear rise in AUPRC ratios when the number of cells are increased. PIDC, the only other single cell network estimation technique, achieves a below average performance at the lowest sample size of 100. This could be due to the fact that PIDC requires large sample sizes for accurate estimation of pairwise joint probability distributions for calculating mutual information. In general, PPCOR has the worst performance among all methods. It should also be noted that the performance of GRNBOOST2 was equivalent to scSGL for all the network topologies when the sample size was 10 times the number of genes. These results indicate the importance of sample size in accurate network estimation for all of the methods and network topologies is considered.

Finally, the performance of each of the methods was evaluated by varying the number of genes. All methods had high AUPRC ratios across network topologies when the number of genes was small. While the AUPRC ratios of all the methods declined with an increase in the number of genes, scSGL performed significantly better than most of the benchmarking

methods. This dip in performance could be attributed to the fact that all methods learn very dense networks. With an increase in the number of nodes, there is an increasing number of false edges detected by every algorithm. The performance of scSGL could further be improved with a more biologically informed framework for hyperparameter selection.

Computational Complexity: Methods are compared in terms of their scalability to datasets with large number of genes. For this purpose, synthetic data generation process used in parameter sensitivity analysis is employed to create three datasets with 500, 1000 and 2000 genes. Each dataset is generated from Barabási–Albert model, includes 1000 cells, and has a dropout ratio of 0.26. Average run time and AUPRC ratios over 10 replicates are reported in Figure 4.4. We reported results only for the correlation kernel, as other kernels have similar performances and run times. It is observed that scSGL runs significantly faster than GENIE3, GRNBOOST2 and PIDC while having superior performance in terms of AUPRC ratio. Although PPCOR runs faster than scSGL, it shows poor performance. Further discussion on the computational and storage complexity of scSGL is provided in [2]

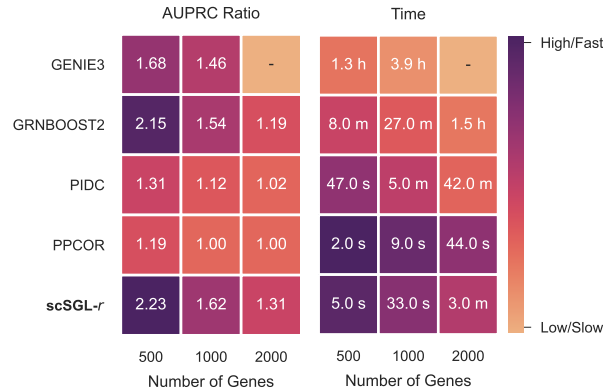


Figure 4.4 Scalability analysis of different methods. Run time of benchmarking methods are calculated using BEELINE pipeline [3]. Run time of scSGL includes kernel construction and optimization procedure. All methods are run on the same computer. Results of GENIE3 for 2000 genes are not reported due to its high run time.

4.4.2 Real Datasets

For real datasets, we consider scRNAseq expressions of human embryonic stem cells (hESC) and mouse embryonic stem cells (mESC) which include 758 and 451 cells, respectively. We inferred GRNs between 500 highly varying genes along with highly varying TFs [3]. Inferred GRNs are compared to three different databases of gene regulations: STRING [137], cell-type specific [138] and nonspecific [139, 140, 141]. AUPRC ratios are reported in Figure 4.5. All methods have performance values close to random estimator. Except PPCOR, which has random performance in both datasets and for all databases, methods have comparable performances, with scSGL showing slightly better performance in hESC and while benchmarking methods working slightly better in mESC.

To add biological meaning to the estimated networks we compared them to the reference networks in the STRING database. The STRING database is a compendium of protein-protein interactions created by gathering information from varying sources like experimental studies, text mining etc. The edges in the STRING network are classified as high confidence (minimum score of 0.700), medium confidence (minimum score of 0.400) and low confidence (minimum score of 0.150). In hESC dataset, scSGL- ρ identified the maximum number of high confidence associations present in the STRING reference network. scSGL- ρ , scSGL- r and scSGL- τ_{zi} each identified 60, 56 and 24 high confidence STRING interactions, respectively, with an edge confidence greater than 0.5. The interactions identified by scSGL- r form a network of 56 unique genes including genes Nanog, Sox2, Sox4, Pou5f1, Ctnnb1, Gata2, Gata3 and many others. Lineage-specific marker genes, Cdk6, Col5a1, Vim, and Itg5, which are known to have regulatory roles in cell differentiation were also detected by scSGL- r but with edge confidence less than 0.5 (0.1-0.3) [142, 143]. scSGL- ρ and scSGL- r identified 20 common genes including Sox2, Sox4, Gata6, Ctnnb1 and Bmp4. scSGL- τ_{zi} identified the least number of genes but successfully retrieved lineage markers Nanog, Sox2, Sox4, Pou5f1, Ctnnb1, Gata2, Gata3. All three kernel methods identified genes Sox4, Ctnnb1, Bmp4 and Gata6. According to the STRING database, the 56 genes identified by



Figure 4.5 Performance of methods for two real-world scRNAseq datasets. Inferred graphs are compared to three different gene regulatory databases.

scSGL- r are associated with 839 significantly enriched biological process gene ontology (GO) terms that include cell differentiation, chromosome separation, specification of animal organ position, mitotic nuclear division and organ formation. Genes identified by scSGL- ρ and scSGL- τ_{zi} had similar functional enrichments for biological processes. To demonstrate some of the learned associations in hESC, we plotted the subnetwork of 24 lineage specific marker genes using scSGL [143]. Figure 6 shows the presence of activating relationships between key definitive endoderm (DE) markers like Gata6, Gata4, and Eomes and joint inhibition of pluripotency markers Pou5f1, Nanog, and Sox2. Gata4 and Gata6 have been reported as necessary for the development and function of a number of endoderm-derived tissues and cells [144, 145] and onset of Gata4 and Gata6 expression has been reported to be coincident with the beginning of endoderm gene expression [146]. In addition, inhibition of pluripotency markers by the key DE markers indicates progression of the cells towards a DE state. We also learned day specific scSGL graphs for the 24 marker genes by first clustering the dataset over days (0,12,24,36,72 and 96 hrs). Day specific graphs showed that scSGL can effectively recover gene network changes from data clustered over time points (Section 7; Supplementary Data) .

In mESC dataset, scSGL- ρ , scSGL- r and scSGL- τ_{zi} each identified 67, 103 and 55 high confidence STRING interactions, respectively, with an edge confidence greater than 0.5. The

three estimated networks capture interactions regulated by known transcription factors Sox2, Nanog, Klf4, Myc and Sall4 [147]. scSGL- r identified known relationships between Sox2 and Nanog; Esrrb with Sox2 and Rybp among many others. scSGL- ρ identified known relationships between Esrrb and Etv5 and indirect interactions between Sall4 and Rybp regulated by TF Oct4. scSGL- τ_{zi} identified most of the important relationships identified by scSGL- r along with additional relationships between Sox2, Nanog, and Rif1. According to the STRING database, the 103 genes identified by scSGL- r are associated with 908 significantly enriched biological process GO terms that include cell fate determination, specification and commitment, mitotic DNA replication and regulation of nodal signalling pathway. Similar to hESC analysis, scSGL, irrespective of the chosen kernel, identified genes with similar functional enrichments for biological processes. To demonstrate some of the learned associations in mESC, we plotted the subnetwork of 19 well known marker genes+TF in mESC differentiation, estimated using scSGL. As can be seen in Figure 6B, Nanog, Gata4, Sox2, Sox17, Zfp42 and Lefty1 emerge as some of the hub nodes with high degrees of associations. The learned network also captures vital signed associations between Sox2, Nanog, Sox17,

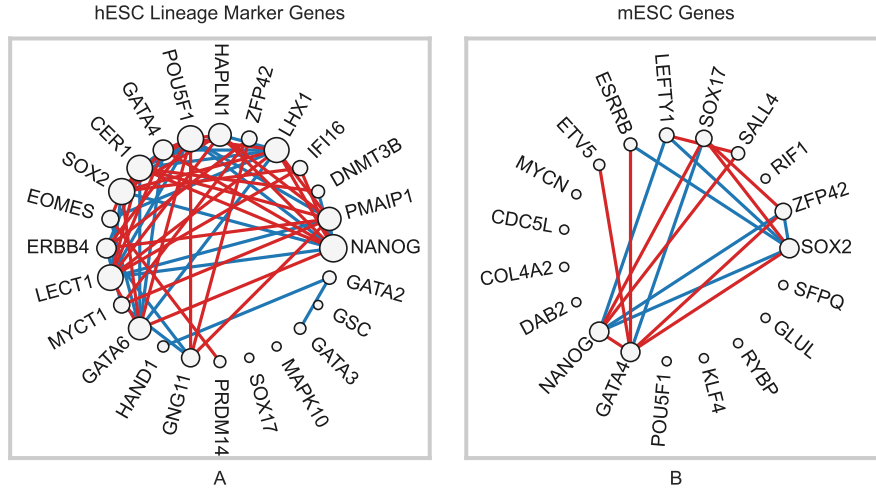


Figure 4.6 The subnetworks of 24 lineage specific genes in hESC (A) and 19 well known marker genes in mESC (B). We report results of scSGL- r as it has the highest AUPRC ratio in Figure 4.5. For clarity, only those edges whose absolute edge weight fall into the top 1 percentile are shown. Node sizes are proportional to their degrees.

Zfp42 and Gata4. It is well known that Sox2 and Nanog form the core of a transcription factor network that promotes embryonic stem cell pluripotency and self-renewal. Zfp42 is also known to be a direct target of Nanog, which is augmented by Sox2 [148]. In addition, Sox17 together with Gata4 expression reinforce a transcriptional network that antagonizes Nanog expression to initiate differentiation [149].

Finally, to analyze the relation between edges identified by scSGL and benchmarking methods, the intersection between the top 1000 edges is reported as an UpSet plot [150] in Figure 4.7. In both datasets, PPCOR does not have any intersection with other methods probably because of its poor performance reported in Figure 4.5. The remaining 6 methods have an intersection set with cardinality around 40 edges. The same number of common edges is found in the intersection of PIDC, GENIE3, GRNBOOST2, scSGL- τ_{zi} , scSGL- r and in the intersection of PIDC, GENIE3, scSGL- τ_{zi} , scSGL- r , scSGL- ρ . These observations hold for both datasets, indicating the reproducibility of the proposed approach across different datasets. Edges identified by τ_{zi} and r have more intersecting edges with benchmarking methods and with each other than those identified by ρ , which indicates that the benchmarking methods have more common edges with correlation based association metrics than with proportionality measures. scSGL methods have more common edges with PIDC than with GENIE3 and GRNBOOST2, which may be due to the fact that PIDC learns co-expression GRN similar to scSGL, while GENIE3 and GRNBOOST2 learn directed interactions between genes.

4.5 Discussion

In this paper, we have introduced a novel network inference algorithm based on GSP. Our proposed algorithm scSGL identifies functional relationships between genes by learning the signed adjacency matrix from the gene expression data under the assumption that graph signals are similar over positive edges and dissimilar over negative edges. This novel technique also takes into account the nonlinearity of the gene interactions by employing kernel

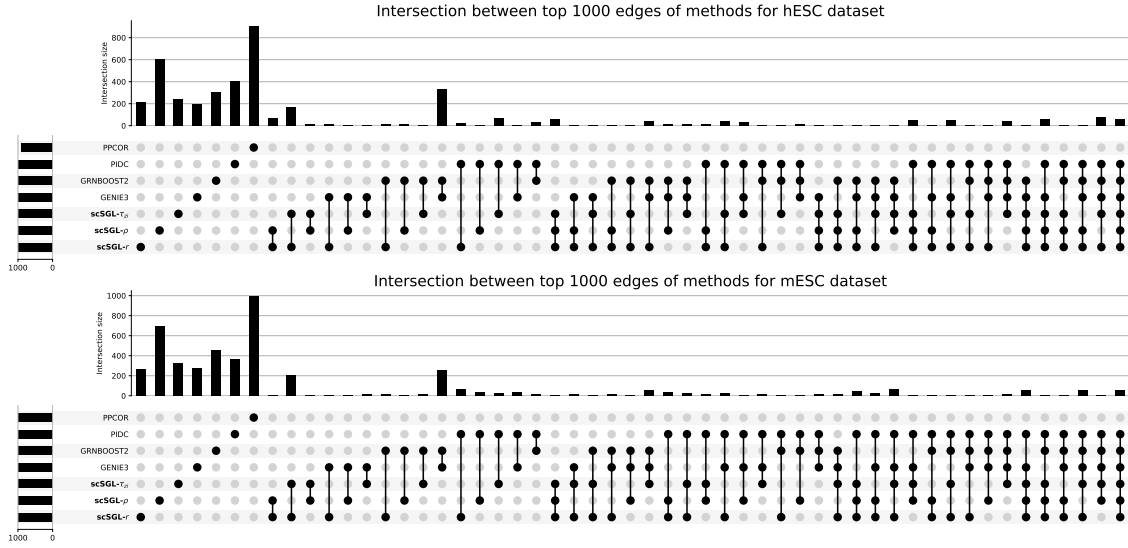


Figure 4.7 UpSet plot that shows intersection between the top 1000 edges by scSGL with 3 kernels and benchmarking methods in hESC and mESC datasets.

mappings. We applied scSGL to four curated datasets derived from "published Boolean models of GRN" and two real experimental scRNAseq datasets during differentiation. To conduct an in-depth analysis of gene co-expression network reconstruction from scRNAseq datasets, we generated simulations from zero inflated negative binomial distributions. These simulations, generated using different parameter combinations, were used to investigate the robustness of our proposed methods to changing cell sizes, gene numbers and dropout rates.

For the curated datasets, scSGL consistently obtained higher AUPRC ratios in comparison to the benchmarking methods, despite each dataset having a different number of stable cell states. Parameter sensitivity analysis reflected the superior performance of scSGL in estimating networks under varying network topologies. The performance remained consistent even when the gene numbers increased, the dropout rates were high and the sample sizes were low. This indicated the robustness of scSGL in modelling networks under varying characteristics of scRNA-seq datasets.

The networks estimated from real data using scSGL identified important functional relationships between target genes and transcription factors and exhibited enrichment for ap-

appropriate functional processes. In addition, day specific analysis of the hESC dataset showed how scSGL can be used to infer gene networks for clustered datasets (Supplementary Material). This may be particularly important in single cell network inference, if cell-type specific networks are of interest. We also demonstrated that scSGL attained performance comparable to state-of-the-art-methods in real data experiments, with the performance of all the GRN reconstruction methods being close to random. Accuracy evaluation of the predicted networks for the real datasets were done using cell-type specific, non-specific and functional networks described in [3]. However, most of the information in these ground truth datasets have been accumulated based on tissue level data and hence it's not completely appropriate to calculate precision and recall rates from these databases.

Although scRNAseq techniques provide significant advantages over bulk data such as increased sample size with higher depth coverage and presence of highly distinct cell clusters, it also comes laced with multiple sources of technical and biological noise. Moreover, the inability to differentiate between technical and biological noise, and the absence of adequate noise modelling techniques further exacerbate the problem [151, 8]. scSGL aims to capture the node similarities and dissimilarities based on distances between graph signals. These graph signals exhibit smoothness, which implies that within a given node cluster, genes tend to be homogeneous, while varying across clusters. This leads to densely connected graphs where the heterogeneity induced by distinct cell sub-populations can be simultaneously curbed. Using single cell data with cell cluster labels, easily obtained from single cell clustering algorithms [152], in conjunction with scSGL can aid in identifying functional modules that are associated with a cell type [153]. Integrating pseudotemporal ordering with scSGL can further help in identifying the functional modules associated with differential pathways [154].

Despite the availability of a large number of computational methods, accurate GRN reconstruction still remains an open problem. Most reconstruction methods are based on the assumption that presence of an edge implies regulatory relationships. They also have

the tendency to establish links between genes regulated by the same regulator. These issues can generate a lot of false positives and therefore additional sources of data such as ChIP-seq measurements that help in identifying direct interactions between TFs and target genes, can provide a way to filter out the spurious interactions [155]. Finally, gene regulation has multiple layers beyond direct TF-target interaction, but functional relationships can only be established if these relationships induce persistent changes in transcriptional state. As single cell data sources over multiple modalities continue to become available, it will be interesting to see how integration of these data types aids GRN reconstruction using scSGL. [156].

4.6 Acknowledgements

This chapter is based on my published paper [2]. I would like to thank the joint first author Abdullah Karaslaanli, and co-authors Dr Selin Aviyente and Dr Tapabrata Maiti for their support and advice. This work was supported by the National Science Foundation grant: CCF 2006800.

APPENDICES

APPENDIX A

BAYESIAN SINGLE CELL RNASEQ DIFFERENTIAL GENE EXPRESSION TEST FOR DOSE RESPONSE STUDY DESIGNS

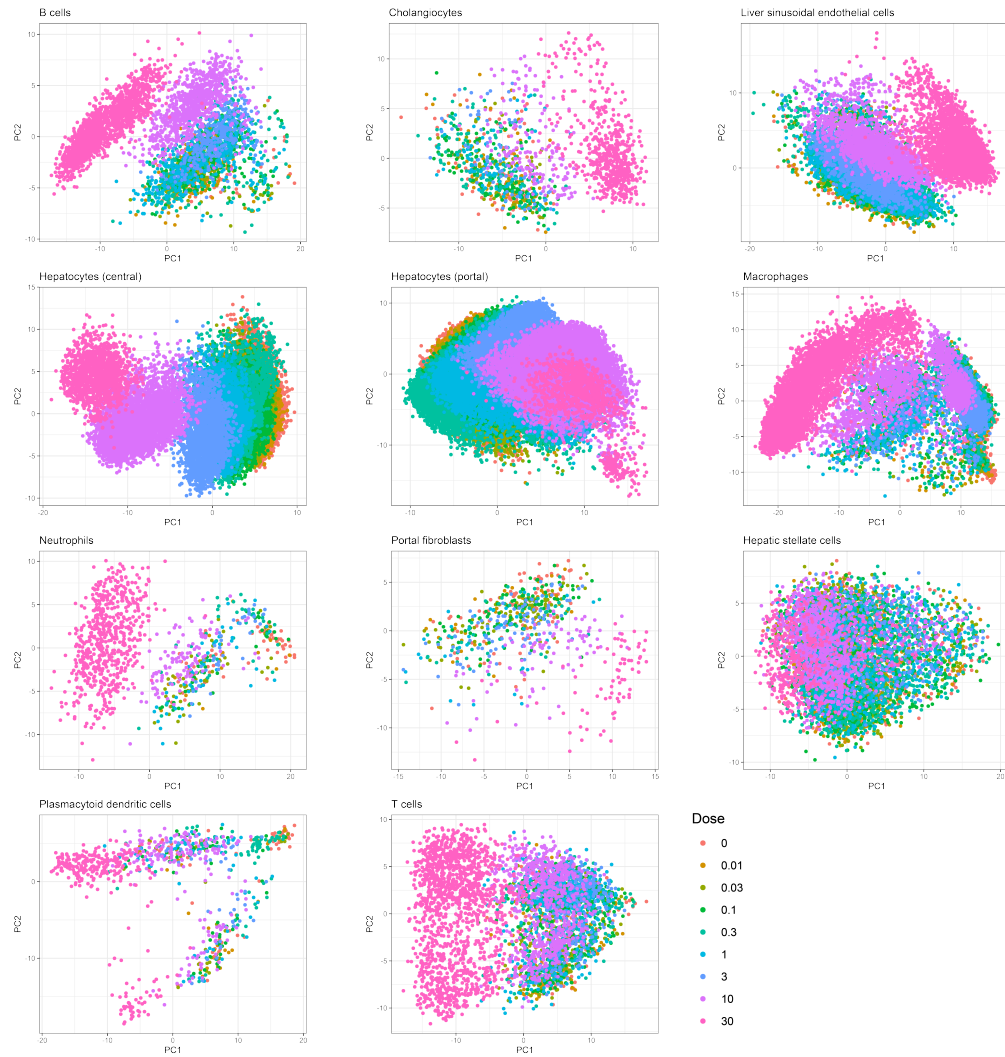


Figure A.1 Principal components analysis of cell types identified in a real hepatic dose-response snRNAseq dataset. Points represent a distinct cell and colors reflect the dose group.

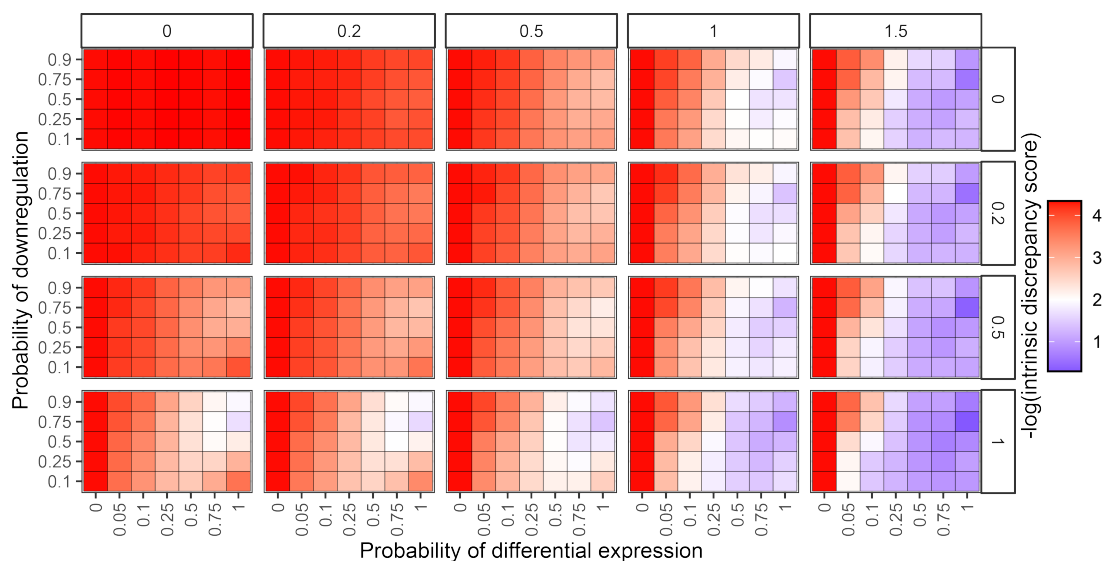


Figure A.2 Comparison of fold-change distribution in simulated and real dose-response snRNAseq data where the log-normal mean (facLoc) and standard deviation (facScale) were varied as well as the percentage of differentially expressed genes and proportion of downregulated DE genes. A total of 5000 genes were simulated or sampled from real data and the fold-change for the highest dose group was calculated. The Kullback-Leibler Divergence (KLD) intrinsic discrepancy (ID) was used to evaluate the similarity in distributions.

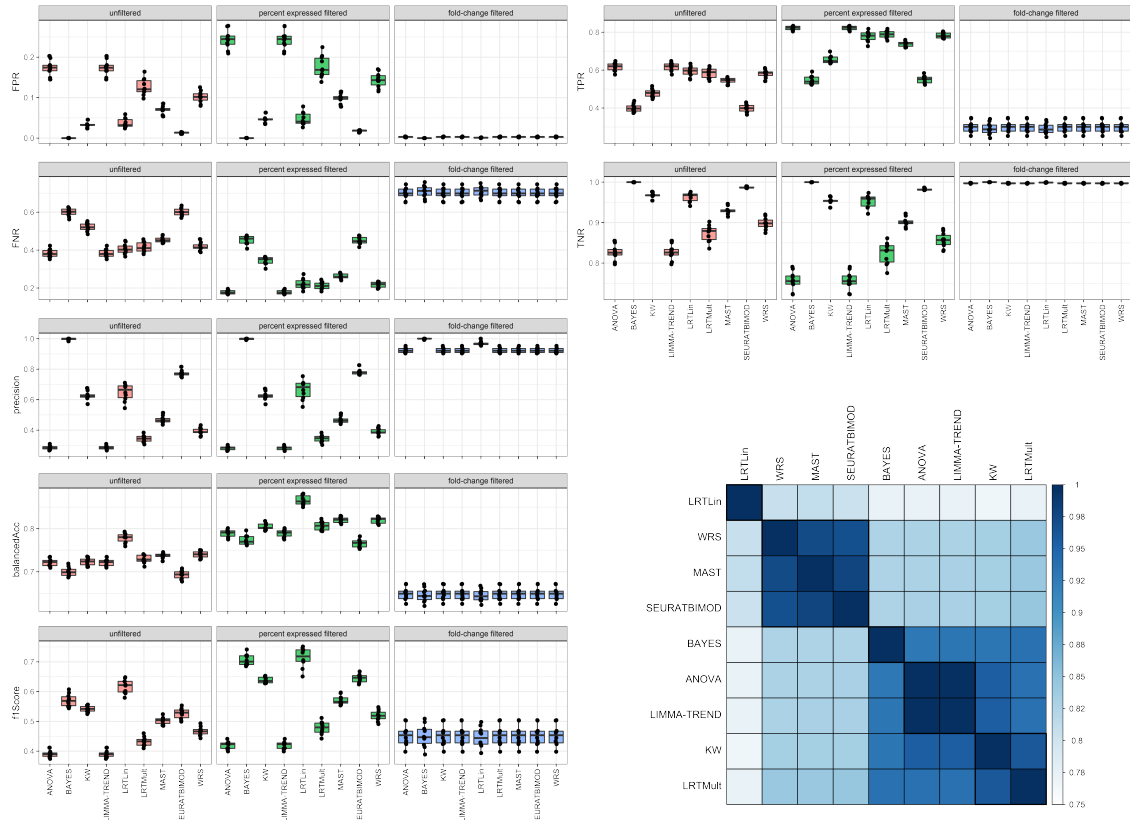


Figure A.3 Benchmarking scores for data simulated using the Splatter [4] wrapper Splattdr with default initial parameters. A total of 4500 cells and 5000 genes were simulated across 9 dose groups with a probability of being differentially expressed of 10%, 50% of which were downregulated. (A) Ground truth was used to estimated the False Positive Rate (FPR), True Positive Rate (TPR), False Negative Rate (FNR), True Negative Rate (TNR), precision, balanced accuracy, and F1 score. Boxplots and whiskers represent values for 10 replicate simulations. (B) The area-under the concordance curve (AUCC) was calculated as previously described (ref) for the 100 most significant genes ($K = 100$). Heatmap represents the pairwise AUCC for each DE analysis grouped by similarity.

APPENDIX B

SEMIPARAMETRIC DOSE RESPONSE CURVE ESTIMATION FOR SINGLE CELL DOSE RESPONSE EXPERIMENTS

B.1 Proof of Theorem 1

Define $\Theta^{(0)} = (\phi^{(0)}, \beta^{(0)}, \gamma_{1,1}^{(0)}, \dots, \gamma_{1,\delta+K+1}^{(0)}, \dots, \gamma_{J,1}^{(0)}, \dots, \gamma_{J,\delta+K+1}^{(0)}, \psi_0^{(0)}, \psi_1^{(0)})^\top$. Then,

$$\begin{aligned}
\ell(\Theta) &= \ell(\Theta^{(0)}) + \ell(\Theta) - \ell(\Theta^{(0)}) \\
&= \ell(\Theta^{(0)}) \\
&\quad + \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(Y_{i,j} \log \left(\frac{\mu_{i,j}}{\mu_{i,j}^{(0)}} \right) - Y_{i,j} \log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \right. \\
&\quad - \exp(\phi) \log \{s_i \mu_{i,j} + \exp(\phi)\} + \exp(\phi^{(0)}) \log \{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} \\
&\quad + \phi \exp(\phi) - \phi^{(0)} \exp(\phi^{(0)}) + \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] \\
&\quad - \log[\Gamma\{Y_{i,j} + \exp(\phi^{(0)})\}] \\
&\quad - \log[\Gamma\{\exp(\phi)\}] + \log[\Gamma\{\exp(\phi^{(0)})\}] \\
&\quad \left. \left. + \log(\omega_{i,j}) - \log(\omega_{i,j}^{(0)}) \right) \right. \\
&\quad + I(Y_{i,j} = 0) \log \left[(1 - \omega_{i,j}) + \omega_{i,j} \left\{ \frac{\exp(\phi)}{s_i \mu_{i,j} + \exp(\phi)} \right\}^{e^\phi} \right] \\
&\quad \left. - I(Y_{i,j} = 0) \log \left[(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} \left\{ \frac{\exp(\phi^{(0)})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\}^{e^{\phi^{(0)}}} \right] \right\}
\end{aligned}$$

For the sake of brevity, define $G(\phi, \mu_{i,j}) = [\exp(\phi) / \{s_j \mu_{i,j} + \exp(\phi)\}]^{e^\phi}$. Then,

$$\begin{aligned}
\ell(\Theta) &= \ell(\Theta^{(0)}) \\
&\quad + \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(Y_{i,j} \log \left(\frac{\mu_{i,j}}{\mu_{i,j}^{(0)}} \right) - Y_{i,j} \log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \right. \\
&\quad - \exp(\phi) \log \{s_i \mu_{i,j} + \exp(\phi)\} + \exp(\phi^{(0)}) \log \{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} \\
&\quad \left. + \phi \exp(\phi) - \phi^{(0)} \exp(\phi^{(0)}) + \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] \right.
\end{aligned}$$

$$\begin{aligned}
& -\log[\Gamma\{Y_{i,j} + \exp(\phi^{(0)})\}] \\
& -\log[\Gamma\{\exp(\phi)\}] + \log[\Gamma\{\exp(\phi^{(0)})\}] \\
& +\log(\omega_{i,j}) - \log(\omega_{i,j}^{(0)}) \Big) \\
& +I(Y_{i,j} = 0)\log \left\{ \frac{(1 - \omega_{i,j}) + \omega_{i,j}G(\phi, \mu_{i,j})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \Big\} \\
= & \ell(\Theta^{(0)}) \\
& + \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(Y_{i,j} \log \left(\frac{\mu_{i,j}}{\mu_{i,j}^{(0)}} \right) - Y_{i,j} \log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \right. \\
& - \exp(\phi) \log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \\
& - \exp(\phi) \log\{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} + \exp(\phi^{(0)}) \log\{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} \\
& + \phi \exp(\phi) - \phi^{(0)} \exp(\phi^{(0)}) + \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] \\
& - \log[\Gamma\{Y_{i,j} + \exp(\phi^{(0)})\}] - \log[\Gamma\{\exp(\phi)\}] + \log[\Gamma\{\exp(\phi^{(0)})\}] \\
& +\log(\omega_{i,j}) - \log(\omega_{i,j}^{(0)}) \Big) \\
& +I(Y_{i,j} = 0) \\
& \times \left\{ \log \left[(1 - \omega_{i,j}^{(0)}) \{ (1 - \omega_{i,j}) / (1 - \omega_{i,j}^{(0)}) \} \right] \right. \\
& + \log \left[1 + \frac{\omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)}) \{ \omega_{i,j} G(\phi, \mu_{i,j}) / \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)}) \}}{(1 - \omega_{i,j}^{(0)}) \{ (1 - \omega_{i,j}) / (1 - \omega_{i,j}^{(0)}) \}} \right] \\
& \left. - \log[(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})] \right\} \\
\geq & \ell(\Theta^{(0)}) \\
& + \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(Y_{i,j} \log \left(\frac{\mu_{i,j}}{\mu_{i,j}^{(0)}} \right) - Y_{i,j} \log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \right. \\
& - \exp(\phi) \log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \\
& - \exp(\phi) \log\{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} + \exp(\phi^{(0)}) \log\{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} \\
& + \phi \exp(\phi) - \phi^{(0)} \exp(\phi^{(0)}) + \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] \\
& - \log[\Gamma\{Y_{i,j} + \exp(\phi^{(0)})\}] - \log[\Gamma\{\exp(\phi)\}] + \log[\Gamma\{\exp(\phi^{(0)})\}]
\end{aligned}$$

$$\begin{aligned}
& +\log(\omega_{i,j}) - \log(\omega_{i,j}^{(0)}) \Big) \\
& + I(Y_{i,j} = 0) \times \left[\frac{(1 - \omega_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \log \left\{ \frac{(1 - \omega_{i,j})}{(1 - \omega_{i,j}^{(0)})} \right\} \right. \\
& \left. + \frac{\omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \log \left\{ \frac{\omega_{i,j} G(\phi, \mu_{i,j})}{\omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right] \Big\} \quad (B.1)
\end{aligned}$$

The last inequality follows by applying the Jensen's inequality to the term involved with $I(Y_{i,j} = 0)$. Now, applying the inequality $-\log(x) \geq 1 - x$ for any generic $x > 0$ we obtain

$$\begin{aligned}
-\log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} & \geq 1 - \frac{s_i \mu_{i,j} + \exp(\phi)}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& = \frac{s_j (\mu_{i,j}^{(0)} - \mu_{i,j}) + \exp(\phi^{(0)}) - \exp(\phi)}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})}. \quad (B.2)
\end{aligned}$$

Let us now consider the very last term of (B.1).

$$\begin{aligned}
\log \left\{ \frac{\omega_{i,j} G(\phi, \mu_{i,j})}{\omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} & = \log \left\{ \frac{\omega_{i,j}}{\omega_{i,j}^{(0)}} \right\} + \log \left\{ \frac{G(\phi, \mu_{i,j})}{G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& = \log \left\{ \frac{\omega_{i,j}}{\omega_{i,j}^{(0)}} \right\} - \log \{G(\phi^{(0)}, \mu_{i,j}^{(0)})\} + \log \{G(\phi, \mu_{i,j})\} \\
& = \log \left\{ \frac{\omega_{i,j}}{\omega_{i,j}^{(0)}} \right\} - \log \{G(\phi^{(0)}, \mu_{i,j}^{(0)})\} + \exp(\phi) \phi \\
& \quad - \exp(\phi) \left[\log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \\
& \quad \left. + \log \left\{ s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \right] \\
& = \log \left\{ \frac{\omega_{i,j}}{\omega_{i,j}^{(0)}} \right\} - \log \{G(\phi^{(0)}, \mu_{i,j}^{(0)})\} + \exp(\phi) \phi \\
& \quad - \exp(\phi) \log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \\
& \quad - \exp(\phi) \log \left\{ s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \\
& \geq \log \left\{ \frac{\omega_{i,j}}{\omega_{i,j}^{(0)}} \right\} - \log \{G(\phi^{(0)}, \mu_{i,j}^{(0)})\} + \exp(\phi) \phi \\
& \quad + \exp(\phi) \left\{ \frac{s_j (\mu_{i,j}^{(0)} - \mu_{i,j}) + \exp(\phi^{(0)}) - \exp(\phi)}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\}
\end{aligned}$$

$$\begin{aligned}
& -\exp(\phi)\log\left\{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})\right\} \\
= & \log\left\{\frac{\omega_{i,j}}{\omega_{i,j}^{(0)}}\right\}-\log\{G(\phi^{(0)},\mu_{i,j}^{(0)})\}+\exp(\phi)\phi \\
& -\exp(\phi)\log\left\{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})\right\} \\
& +\exp(\phi)\left\{\frac{\exp(\phi^{(0)})-\exp(\phi)}{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})}\right\} \\
& +\frac{s_j\exp(\phi)(\mu_{i,j}^{(0)}-\mu_{i,j})}{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})} \\
= & \log\left\{\frac{\omega_{i,j}}{\omega_{i,j}^{(0)}}\right\}-\log\{G(\phi^{(0)},\mu_{i,j}^{(0)})\}+\exp(\phi)\phi \\
& -\exp(\phi)\log\left\{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})\right\} \\
& +\exp(\phi)\left\{\frac{\exp(\phi^{(0)})-\exp(\phi)}{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})}\right\} \\
& +\frac{s_j\{\exp(\phi)-\exp(\phi^{(0)})\}(\mu_{i,j}^{(0)}-\mu_{i,j})}{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})} \\
& +\frac{s_j\exp(\phi^{(0)})(\mu_{i,j}^{(0)}-\mu_{i,j})}{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})} \\
\geq & \log\left\{\frac{\omega_{i,j}}{\omega_{i,j}^{(0)}}\right\}-\log\{G(\phi^{(0)},\mu_{i,j}^{(0)})\}+\exp(\phi)\phi \\
& -\exp(\phi)\log\left\{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})\right\} \\
& +\exp(\phi)\left\{\frac{\exp(\phi^{(0)})-\exp(\phi)}{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})}\right\} \\
& -\frac{0.5s_j}{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})} \\
& \times\left[\{\exp(\phi)-\exp(\phi^{(0)})\}^2+(\mu_{i,j}^{(0)}-\mu_{i,j})^2\right] \\
& +\frac{s_j\exp(\phi^{(0)})(\mu_{i,j}^{(0)}-\mu_{i,j})}{s_j\mu_{i,j}^{(0)}+\exp(\phi^{(0)})}
\end{aligned} \tag{B.3}$$

The first inequality in the above derivation is obtained by using (B.2). The second inequality is obtained by applying the inequality $ab \geq -0.5(a^2 + b^2)$. Once again using (B.2) we obtain

$$\begin{aligned}
-\exp(\phi) \log \left\{ \frac{s_i \mu_{i,j} + \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} &\geq \exp(\phi) \frac{s_j (\mu_{i,j}^{(0)} - \mu_{i,j}) + \exp(\phi^{(0)}) - \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
&= \frac{s_j \exp(\phi) (\mu_{i,j}^{(0)} - \mu_{i,j})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
&\quad + \frac{\exp(\phi) \{ \exp(\phi^{(0)}) - \exp(\phi) \}}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
&= \frac{s_j \{ \exp(\phi) - \exp(\phi^{(0)}) \} (\mu_{i,j}^{(0)} - \mu_{i,j})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
&\quad + \frac{s_j \exp(\phi^{(0)}) (\mu_{i,j}^{(0)} - \mu_{i,j})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
&\quad + \frac{\exp(\phi) \{ \exp(\phi^{(0)}) - \exp(\phi) \}}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
&\geq - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
&\quad \times \left[\{ \exp(\phi) - \exp(\phi^{(0)}) \}^2 + (\mu_{i,j}^{(0)} - \mu_{i,j})^2 \right] \\
&\quad + \frac{s_j \exp(\phi^{(0)}) (\mu_{i,j}^{(0)} - \mu_{i,j})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
&\quad + \frac{\exp(\phi) \{ \exp(\phi^{(0)}) - \exp(\phi) \}}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})}. \tag{B.4}
\end{aligned}$$

Now, we apply inequalities (B.2), (B.3), and (B.4) to (B.1) and obtain,

$$\begin{aligned}
\ell(\Theta) &\geq \ell(\Theta^{(0)}) \\
&\quad + \sum_{i=1}^I \sum_{j=1}^{n_i} \left[I(Y_{i,j} > 0) \left(Y_{i,j} \log \left(\frac{\mu_{i,j}}{\mu_{i,j}^{(0)}} \right) \right. \right. \\
&\quad \left. \left. + Y_{i,j} \left\{ \frac{s_j (\mu_{i,j}^{(0)} - \mu_{i,j}) + \exp(\phi^{(0)}) - \exp(\phi)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \right. \\
&\quad \left. \left. - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left[\{ \exp(\phi) - \exp(\phi^{(0)}) \}^2 + (\mu_{i,j}^{(0)} - \mu_{i,j})^2 \right] \right. \right. \\
&\quad \left. \left. + \frac{s_j \exp(\phi^{(0)}) (\mu_{i,j}^{(0)} - \mu_{i,j})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right) \right]
\end{aligned}$$

$$\begin{aligned}
& + \frac{\exp(\phi)\{\exp(\phi^{(0)}) - \exp(\phi)\}}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& - \exp(\phi)\log\{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} + \exp(\phi^{(0)})\log\{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} \\
& + \phi\exp(\phi) - \phi^{(0)}\exp(\phi^{(0)}) + \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] \\
& - \log[\Gamma\{Y_{i,j} + \exp(\phi^{(0)})\}] - \log[\Gamma\{\exp(\phi)\}] \\
& + \log[\Gamma\{\exp(\phi^{(0)})\}] + \log(\omega_{i,j}) - \log(\omega_{i,j}^{(0)}) \Big) \\
& + I(Y_{i,j} = 0) \times \left\{ \frac{(1 - \omega_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \log \left\{ \frac{(1 - \omega_{i,j})}{(1 - \omega_{i,j}^{(0)})} \right\} \right. \\
& + \frac{\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \left(\log \left\{ \frac{\omega_{i,j}}{\omega_{i,j}^{(0)}} \right\} - \log\{G(\phi^{(0)}, \mu_{i,j}^{(0)})\} \right. \\
& + \exp(\phi)\phi - \exp(\phi)\log\{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} \\
& + \exp(\phi) \left\{ \frac{\exp(\phi^{(0)}) - \exp(\phi)}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \\
& - \frac{0.5s_j}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left[\{\exp(\phi) - \exp(\phi^{(0)})\}^2 + (\mu_{i,j}^{(0)} - \mu_{i,j})^2 \right] \\
& \left. + \frac{s_j\exp(\phi^{(0)})(\mu_{i,j}^{(0)} - \mu_{i,j})}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right) \Big] \\
= & \ell_1^\dagger(\phi|\Theta^{(0)}) + \ell_2^*(\psi_0, \psi_1, \gamma_1, \dots, \gamma_I, \beta|\Theta^{(0)}) \\
& + \ell_3^*(\gamma_1, \dots, \gamma_I, \beta|\Theta^{(0)}) + \ell_4^*(\Theta^{(0)})
\end{aligned} \tag{B.5}$$

where

$$\begin{aligned}
\ell_1^\dagger(\phi|\Theta^{(0)}) = & \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(Y_{i,j} \left\{ \frac{\exp(\phi^{(0)}) - \exp(\phi)}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \right. \\
& - \frac{0.5s_j}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{\exp(\phi) - \exp(\phi^{(0)})\}^2 \\
& + \frac{\exp(\phi)\{\exp(\phi^{(0)}) - \exp(\phi)\}}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& - \exp(\phi)\log\{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})\} \\
& + \phi\exp(\phi) + \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] \\
& \left. \left. - \log[\Gamma\{\exp(\phi)\}] \right) \right\}
\end{aligned}$$

$$\begin{aligned}
& + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \left[\exp(\phi)\phi \right. \\
& - \exp(\phi)\log \left\{ s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \\
& + \exp(\phi) \left\{ \frac{\exp(\phi^{(0)}) - \exp(\phi)}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \\
& \left. - \frac{0.5s_j}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{\exp(\phi) - \exp(\phi^{(0)})\}^2 \right] \Bigg\},
\end{aligned}$$

$$\begin{aligned}
\ell_2^*(\psi_0, \psi_1, \gamma_1, \dots, \gamma_I, \beta | \Theta^{(0)}) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[I(Y_{i,j} > 0) \log(\omega_{i,j}) \right. \\
& + I(Y_{i,j} = 0) \\
& \times \left\{ \frac{(1 - \omega_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \log(1 - \omega_{i,j}) \right. \\
& + \left. \frac{\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \log(\omega_{i,j}) \right\} \Bigg] \\
&= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[I(Y_{i,j} > 0) \left(\psi_0 + \psi_1 \log(s_i\mu_{i,j}) \right. \right. \\
& \quad \left. \left. - \log[1 + \exp\{\psi_0 + \psi_1 \log(s_i\mu_{i,j})\}] \right) \right. \\
& \quad + I(Y_{i,j} = 0) \times \left\{ - \frac{(1 - \omega_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right. \\
& \quad \times \log[1 + \exp\{\psi_0 + \psi_1 \log(s_i\mu_{i,j})\}] \\
& \quad + \frac{\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \\
& \quad \times \left(\psi_0 + \psi_1 \log(s_i\mu_{i,j}) \right. \\
& \quad \left. \left. - \log[1 + \exp\{\psi_0 + \psi_1 \log(s_i\mu_{i,j})\}] \right) \right] \Bigg] \\
&= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) \right. \right. \\
& \quad + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \Bigg\} \\
& \quad \times \{ \psi_0 + \psi_1 \log(s_i\mu_{i,j}) \}
\end{aligned}$$

$$\begin{aligned}
& -\log [1 + \exp\{\psi_0 + \psi_1 \log(s_i \mu_{i,j})\}] \Big) \\
= & \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) \right. \right. \\
& \left. \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
& \times \{\psi_0 + \psi_1 \log(s_i \mu_{i,j})\} \\
& -\log \left[\frac{1 + \exp\{\psi_0 + \psi_1 \log(s_i \mu_{i,j})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \right] \\
& \left. -\log[1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}] \right) \\
\geq & \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) \right. \right. \\
& \left. \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
& \times \{\psi_0 + \psi_1 \log(s_i \mu_{i,j})\} \\
& + 1 - \frac{1 + \exp\{\psi_0 + \psi_1 \log(s_i \mu_{i,j})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
& \left. -\log[1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}] \right),
\end{aligned}$$

where the last inequality is obtained using the result $\log(x) \leq x - 1$, so $-\log(x) \geq 1 - x$.

Now,

$$\begin{aligned}
\ell_2^*(\psi_0, \psi_1, \gamma_1, \dots, \gamma_I, \beta | \Theta^{(0)}) \geq & \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) \right. \right. \\
& \left. \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
& \times \{\psi_0 + \psi_1 \log(s_i \mu_{i,j})\} \\
& + \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
& - \frac{\exp\{\psi_0 + \psi_1 \log(s_i \mu_{i,j})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
& \left. -\log[1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}] \right)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) \right. \right. \\
&\quad \left. \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left[\psi_0 + \psi_1 \log(s_i) \right. \\
&\quad \left. + \psi_1 \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m} B_m(D_{i,j}) + X_{i,j}^\top \beta \right\} \right. \\
&\quad \left. + \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \right. \\
&\quad \times \left(1 - \exp\{\psi_0 - \psi_0^{(0)} + \psi_1 \log(s_i \mu_{i,j}) \right. \\
&\quad \left. \left. - \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\} \right) \right. \\
&\quad \left. \left. - \log[1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}] \right) \right].
\end{aligned}$$

Let us work on two specific terms of the above expression. First,

$$\begin{aligned}
\psi_1 \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m} B_m(D_{i,j}) + X_{i,j}^\top \beta \right\} &= (\psi_1 - \psi_1^{(0)}) \left\{ \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right. \\
&\quad \left. + X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
&\quad + \psi_1^{(0)} \left\{ \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right. \\
&\quad \left. + X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
&\quad + \psi_1 \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \\
&\geq -0.5 \left[(d+K+2)(\psi_1 - \psi^{(0)})^2 \right. \\
&\quad \left. + \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)})^2 B_m^2(D_{i,j}) \right. \\
&\quad \left. + \{X_{i,j}^\top (\beta - \beta^{(0)})\}^2 \right] \\
&\quad + \psi_1^{(0)} \left\{ \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right. \\
&\quad \left. + X_{i,j}^\top (\beta - \beta^{(0)}) \right\}
\end{aligned}$$

$$+ \psi_1 \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\}.$$

The above inequality is obtained by applying the result, $ab \geq -0.5(a^2 + b^2)$ repeatedly on the product terms. The second term,

$$\begin{aligned}
& - \exp\{\psi_0 - \psi_0^{(0)} + \psi_1 \log(s_i \mu_{i,j}) - \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\} \\
= & - \exp \left[\psi_0 - \psi_0^{(0)} + \psi_1 \log(s_i) + \psi_1 \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m} B_m(D_{i,j}) + X_{i,j}^\top \beta \right\} \right. \\
& \left. - \psi_1^{(0)} \log(s_i) - \psi_1^{(0)} \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \right] \\
= & - \exp \left[\psi_0 - \psi_0^{(0)} + \log(s_i)(\psi_1 - \psi_1^{(0)}) + (\psi_1 - \psi_1^{(0)}) \right. \\
& \times \left\{ \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) + X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
& + \psi_1^{(0)} \left\{ \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) + X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
& \left. + (\psi_1 - \psi_1^{(0)}) \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \right] \\
= & - \exp \left[\psi_0 - \psi_0^{(0)} + \log(s_i)(\psi_1 - \psi_1^{(0)}) + (\psi_1 - \psi_1^{(0)}) \right. \\
& \times \left\{ \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) + X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
& + \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) \psi_1^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \psi_1^{(0)} (\beta - \beta^{(0)}) \\
& \left. + (\psi_1 - \psi_1^{(0)}) \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \right] \\
\geq & - \frac{1}{2d + 2K + 7} \left(\exp\{(\psi_0 - \psi_0^{(0)})(2d + 2K + 7)\} \right. \\
& + \exp\{\log(s_i)(\psi_1 - \psi_1^{(0)})(2d + 2K + 7)\} \\
& + \sum_{m=1}^{d+K+1} \exp\{(\psi_1 - \psi_1^{(0)})(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})(2d + 2K + 7)\} \\
& + \exp\{(\psi_1 - \psi_1^{(0)}) X_{i,j}^\top (\beta - \beta^{(0)})(2d + 2K + 7)\} \\
& \left. + \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) \psi_1^{(0)} B_m(D_{i,j})(2d + 2K + 7)\} \right)
\end{aligned}$$

$$\begin{aligned}
& + \exp\{X_{i,j}^\top \psi_1^{(0)}(\beta - \beta^{(0)})(2d + 2K + 7)\} \\
& + \exp\left[(\psi_1 - \psi_1^{(0)})\left\{\sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)}\right\}(2d + 2K + 7)\right] \Bigg) \\
\geq & -\frac{1}{2d + 2K + 7} \left(\exp\{(\psi_0 - \psi_0^{(0)})(2d + 2K + 7)\} \right. \\
& + \exp\{\log(s_i)(\psi_1 - \psi_1^{(0)})(2d + 2K + 7)\} \\
& + \sum_{m=1}^{d+K+1} \exp\left[0.5(\psi_1 - \psi_1^{(0)})^2 + 0.5\{(\gamma_{i,m} - \gamma_{i,m}^{(0)})B_m(D_{i,j})(2d + 2K + 7)\}^2\right] \\
& + \exp\left[0.5(\psi_1 - \psi_1^{(0)})^2 + 0.5\{X_{i,j}^\top(\beta - \beta^{(0)})(2d + 2K + 7)\}^2\right] \\
& + \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)})\psi_1^{(0)} B_m(D_{i,j})(2d + 2K + 7)\} \\
& + \exp\{X_{i,j}^\top \psi_1^{(0)}(\beta - \beta^{(0)})(2d + 2K + 7)\} \\
& + \exp\left[(\psi_1 - \psi_1^{(0)})\left\{\sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)}\right\}(2d + 2K + 7)\right] \Bigg) \\
\geq & -\frac{1}{2d + 2K + 7} \left(\exp\{(\psi_0 - \psi_0^{(0)})(2d + 2K + 7)\} \right. \\
& + \exp\{\log(s_i)(\psi_1 - \psi_1^{(0)})(2d + 2K + 7)\} \\
& + 0.5(d + K + 1) \exp\{(\psi_1 - \psi_1^{(0)})^2\} \\
& + 0.5 \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)})B_m(D_{i,j})(2d + 2K + 7)\}^2 \\
& + 0.5 \exp\{(\psi_1 - \psi_1^{(0)})^2\} + 0.5 \exp\{X_{i,j}^\top(\beta - \beta^{(0)})(2d + 2K + 7)\}^2 \\
& + \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)})\psi_1^{(0)} B_m(D_{i,j})(2d + 2K + 7)\} \\
& + \exp\{X_{i,j}^\top \psi_1^{(0)}(\beta - \beta^{(0)})(2d + 2K + 7)\} \\
& + \exp\left[(\psi_1 - \psi_1^{(0)})\left\{\sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)}\right\}(2d + 2K + 7)\right] \Bigg).
\end{aligned}$$

The first and third inequalities follow from the AM-GM inequality and the second inequality follows from the result that $ab \leq (a^2 + b^2)/2$. Now,

$$\begin{aligned}
\ell_2^*(\psi_0, \psi_1, \gamma_1, \dots, \gamma_I, \beta | \Theta^{(0)}) \geq & \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) \right. \right. \\
& \left. \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right)
\end{aligned}$$

$$\begin{aligned}
& \times \left[\psi_0 + \psi_1 \log(s_i) \right. \\
& - 0.5(d + K + 2)(\psi_1 - \psi^{(0)})^2 \\
& - 0.5 \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)})^2 B_m^2(D_{i,j}) \\
& - 0.5 \{X_{i,j}^\top (\beta - \beta^{(0)})\}^2 \\
& + \psi_1^{(0)} \left\{ \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right. \\
& + X_{i,j}^\top (\beta - \beta^{(0)}) \} \\
& + \psi_1 \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \Big] \\
& + \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
& \times \left\{ 1 - \frac{1}{2d + 2K + 7} \right. \\
& \times \left(\exp\{(\psi_0 - \psi_0^{(0)})(2d + 2K + 7)\} \right. \\
& + \exp\{\log(s_i)(\psi_1 - \psi_1^{(0)})(2d + 2K + 7)\} \\
& + 0.5(d + K + 1) \exp\{(\psi_1 - \psi_1^{(0)})^2\} \\
& + 0.5 \sum_{m=1}^{d+K+1} \exp\{[(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \\
& \times (2d + 2K + 7)]^2\} + 0.5 \exp\{(\psi_1 - \psi_1^{(0)})^2\} \\
& + 0.5 \exp\{[X_{i,j}^\top (\beta - \beta^{(0)})(2d + 2K + 7)]^2\} \\
& + \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) \psi_1^{(0)} B_m(D_{i,j}) \\
& \times (2d + 2K + 7)\} \\
& + \exp\{X_{i,j}^\top \psi_1^{(0)} (\beta - \beta^{(0)})(2d + 2K + 7)\} \\
& + \exp \left[(\psi_1 - \psi_1^{(0)}) \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \right. \\
& \times (2d + 2K + 7) \Big] \Big\} \\
& - \log[1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}] \Big)
\end{aligned}$$

$$\begin{aligned}
&= g_1(\psi_0|\Theta_0) + g_2(\psi_1|\Theta_0) + \sum_{i=1}^I g_{3,i}(\gamma_i|\Theta_0) \\
&\quad + g_4(\beta|\Theta_0) + g_5(\Theta_0),
\end{aligned}$$

$$\begin{aligned}
g_1(\psi_0|\Theta_0) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[\left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \psi_0 \right. \\
&\quad \left. - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}} \right. \\
&\quad \left. \times \frac{1}{2d + 2K + 7} \exp\{(\psi_0 - \psi_0^{(0)})(2d + 2K + 7)\} \right],
\end{aligned}$$

$$\begin{aligned}
g_2(\psi_1|\Theta_0) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left[\psi_1 \log(s_i) - 0.5(d + K + 2)(\psi_1 - \psi^{(0)})^2 \right. \\
&\quad \left. + \psi_1 \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \right] \\
&\quad \left. - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}} \times \frac{1}{2d + 2K + 7} \left(\right. \right. \\
&\quad \exp\{\log(s_i)(\psi_1 - \psi_1^{(0)})(2d + 2K + 7)\} \\
&\quad + 0.5(d + K + 1) \exp\{(\psi_1 - \psi_1^{(0)})^2\} \\
&\quad + 0.5 \exp\{(\psi_1 - \psi_1^{(0)})^2\} \\
&\quad \left. \left. + \exp \left[(\psi_1 - \psi_1^{(0)}) \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} (2d + 2K + 7) \right] \right) \right\},
\end{aligned}$$

$$\begin{aligned}
g_{3,i}(\gamma_i|\Theta^{(0)}) &= \sum_{j=1}^{n_i} \left\{ \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left[-0.5 \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)})^2 B_m^2(D_{i,j}) \right. \\
&\quad \left. + \psi_1^{(0)} \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right]
\end{aligned}$$

$$\begin{aligned}
& - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
& \times \frac{1}{2d + 2K + 7} \left(0.5 \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})\} \right. \\
& \times (2d + 2K + 7)^2 \Big] \\
& + \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) \psi_1^{(0)} B_m(D_{i,j}) (2d + 2K + 7)\} \Big) \Big\},
\end{aligned}$$

and

$$\begin{aligned}
g_4(\beta | \Theta^{(0)}) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
& \times \left[-0.5 \{X_{i,j}^\top (\beta - \beta^{(0)})\}^2 + \psi_1^{(0)} X_{i,j}^\top (\beta - \beta^{(0)}) \right] \\
& - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
& \times \frac{1}{2d + 2K + 7} \left[0.5 \exp\{X_{i,j}^\top (\beta - \beta^{(0)}) (2d + 2K + 7)\}^2 \right. \\
& \left. \left. + \exp\{X_{i,j}^\top \psi_1^{(0)} (\beta - \beta^{(0)}) (2d + 2K + 7)\} \right] \right).
\end{aligned}$$

$$\begin{aligned}
\ell_3^*(\gamma_1, \dots, \gamma_I, \beta | \Theta^{(0)}) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[I(Y_{i,j} > 0) \left\{ Y_{i,j} \log(\mu_{i,j}) - \frac{Y_{i,j} s_j \mu_{i,j}}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \right. \\
& - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} (\mu_{i,j} - \mu_{i,j}^{(0)})^2 \\
& \left. \left. - \frac{s_j \exp(\phi^{(0)}) \mu_{i,j}}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \\
& - \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \\
& \times \left\{ \frac{0.5 s_j}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} (\mu_{i,j} - \mu_{i,j}^{(0)})^2 + \frac{s_j \exp(\phi^{(0)}) \mu_{i,j}}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \Big] \\
&= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[I(Y_{i,j} > 0) Y_{i,j} \log(\mu_{i,j}) \right. \\
& \left. - \left\{ \frac{I(Y_{i,j} > 0) Y_{i,j} s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} + \frac{I(Y_{i,j} > 0) s_j \exp(\phi^{(0)})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right]
\end{aligned}$$

$$\begin{aligned}
& + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \frac{s_j \exp(\phi^{(0)})}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \Big\} \mu_{i,j} \\
& - \left\{ \frac{0.5s_j I(Y_{i,j} > 0)}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right. \\
& \quad \times \frac{0.5s_j}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \Big\} \times (\mu_{i,j} - \mu_{i,j}^{(0)})^2 \Big] \\
& = \sum_{i=1}^I \sum_{j=1}^{n_i} \left[I(Y_{i,j} > 0) Y_{i,j} \log(\mu_{i,j}) \right. \\
& \quad - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& \quad + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \Big\} \mu_{i,j} \\
& \quad - \frac{0.5s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) \right. \\
& \quad + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \Big\} (\mu_{i,j} - \mu_{i,j}^{(0)})^2 \Big]
\end{aligned}$$

Now, we obtain two inequalities for $-\mu_{i,j}$ and $-(\mu_{i,j} - \mu_{i,j}^{(0)})^2$ and apply them in the above expression. First,

$$\begin{aligned}
-\mu_{i,j} & = -\exp\left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m} B_m(D_{i,j}) + X_{i,j}^\top \beta \right\} \\
& = -\mu_{i,j}^{(0)} \exp\left\{ \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) + X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
& \geq -\frac{\mu_{i,j}^{(0)}}{d+K+2} \left[\sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})(d+K+2)\} \right. \\
& \quad \left. + \exp\{X_{i,j}^\top (\beta - \beta^{(0)})(d+K+2)\} \right],
\end{aligned}$$

where the inequality is obtained by using the AM-GM inequality. Second,

$$\begin{aligned}
-(\mu_{i,j} - \mu_{i,j}^{(0)})^2 & = -(\mu_{i,j}^{(0)})^2 + 2\mu_{i,j}^{(0)}\mu_{i,j} - (\mu_{i,j})^2 \\
& = -(\mu_{i,j}^{(0)})^2 + 2(\mu_{i,j}^{(0)})^2 \left(\frac{\mu_{i,j}}{\mu_{i,j}^{(0)}} \right) - (\mu_{i,j}^{(0)})^2 \left(\frac{\mu_{i,j}}{\mu_{i,j}^{(0)}} \right)^2 \\
& = -(\mu_{i,j}^{(0)})^2
\end{aligned}$$

$$\begin{aligned}
& + 2(\mu_{i,j}^{(0)})^2 \exp\left\{ \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) + X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
& - (\mu_{i,j}^{(0)})^2 \exp\left\{ 2 \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) + 2X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
& \geq - (\mu_{i,j}^{(0)})^2 \\
& + 2(\mu_{i,j}^{(0)})^2 \left\{ 1 + \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) + X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
& - \frac{(\mu_{i,j}^{(0)})^2}{d+K+2} \left[\sum_{m=1}^{d+K+1} \exp\{2(d+K+2)(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})\} \right. \\
& \left. + \exp\{2(d+K+2)X_{i,j}^\top (\beta - \beta^{(0)})\} \right],
\end{aligned}$$

where the inequality is obtained by applying $\exp(x) \geq 1 + x$ for any generic x to the middle term of the above expression and using the AM-GM inequality to the third term. Now, applying these two inequalities we obtain,

$$\begin{aligned}
\ell_3^*(\gamma_1, \dots, \gamma_I, \beta | \Theta^{(0)}) & \geq \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) Y_{i,j} \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m} B_m(D_{i,j}) + X_{i,j}^\top \beta \right\} \right. \\
& - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
& \times \frac{\mu_{i,j}^{(0)}}{d+K+2} \left[\sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})\} \right. \\
& \left. \times (d+K+2) \right\} + \exp\{X_{i,j}^\top (\beta - \beta^{(0)}) (d+K+2)\} \left. \right] \\
& - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) \right. \\
& \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \times \left((\mu_{i,j}^{(0)})^2 \right)
\end{aligned}$$

$$\begin{aligned}
& -2(\mu_{i,j}^{(0)})^2 \left\{ 1 + \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right. \\
& \left. + X_{i,j}^\top (\beta - \beta^{(0)}) \right\} \\
& + \frac{(\mu_{i,j}^{(0)})^2}{d+K+2} \left[\sum_{m=1}^{d+K+1} \exp\{2(d+K+2) \right. \\
& \times (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})\} \\
& \left. + \exp\{2(d+K+2) X_{i,j}^\top (\beta - \beta^{(0)})\} \right] \Bigg\} \\
& = \sum_{i=1}^K \ell_{3,i}^\dagger(\gamma_i | \Theta_0) + \ell_4^\dagger(\beta | \Theta_0),
\end{aligned}$$

where

$$\begin{aligned}
\ell_{3,i}^\dagger(\gamma_i | \Theta_0) &= \sum_{j=1}^{n_i} \left(I(Y_{i,j} > 0) Y_{i,j} \sum_{m=1}^{d+K+1} \gamma_{i,m} B_m(D_{i,j}) \right. \\
& - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
& \times \frac{\mu_{i,j}^{(0)}}{d+K+2} \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})(d+K+2)\} \\
& - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \times \left[-2(\mu_{i,j}^{(0)})^2 \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right. \\
& \left. + \frac{(\mu_{i,j}^{(0)})^2}{d+K+2} \sum_{m=1}^{d+K+1} \exp\{2(d+K+2)(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})\} \right] \Bigg),
\end{aligned}$$

$$\begin{aligned}
\ell_4^\dagger(\beta | \Theta_0) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(I(Y_{i,j} > 0) Y_{i,j} X_{i,j}^\top \beta \right. \\
& - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
& \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\}
\end{aligned}$$

$$\begin{aligned}
& \times \frac{\mu_{i,j}^{(0)}}{d+K+2} \exp\{X_{i,j}^\top(\beta - \beta^{(0)})(d+K+2)\} \\
& - \frac{0.5s_j}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \times \left[-2(\mu_{i,j}^{(0)})^2 X_{i,j}^\top(\beta - \beta^{(0)}) + \frac{(\mu_{i,j}^{(0)})^2}{d+K+2} \exp\{2(d+K+2) \right. \\
& \left. \times X_{i,j}^\top(\beta - \beta^{(0)})\} \right].
\end{aligned}$$

Combining all results we can write

$$\begin{aligned}
\ell(\Theta) & \geq \ell_1^\dagger(\phi|\Theta^{(0)}) + g_1(\psi_0|\Theta^{(0)}) + g_2(\psi_1|\Theta^{(0)}) + \sum_{i=1}^I \ell_{3,i}^\dagger(\gamma_i|\Theta^{(0)}) \\
& + \ell_4^\dagger(\beta|\Theta^{(0)}) + \ell_5^\dagger(\Theta^{(0)}),
\end{aligned}$$

where

$$\begin{aligned}
\ell_{3,i}^\dagger(\gamma_i|\Theta^{(0)}) & = g_{3,i}(\gamma_i|\Theta^{(0)}) + \ell_{3,i}^\dagger(\gamma_i|\Theta^{(0)}), \\
\ell_4^\dagger(\beta|\Theta^{(0)}) & = g_4(\beta|\Theta^{(0)}) + \ell_4^\dagger(\beta|\Theta^{(0)}).
\end{aligned}$$

Let us now consider the derivatives.

$$\begin{aligned}
\frac{\partial}{\partial \phi} \ell_1^\dagger(\phi|\Theta^{(0)}) & = \frac{\partial}{\partial \phi} \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(Y_{i,j} \left\{ \frac{\exp(\phi^{(0)}) - \exp(\phi)}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \right. \\
& - \frac{0.5s_j}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{ \exp(\phi) - \exp(\phi^{(0)}) \}^2 \\
& + \frac{\exp(\phi) \{ \exp(\phi^{(0)}) - \exp(\phi) \}}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& - \exp(\phi) \log \{ s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \} \\
& + \phi \exp(\phi) + \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] \\
& \left. \left. - \log[\Gamma\{\exp(\phi)\}] \right) \right. \\
& + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \left[\exp(\phi)\phi \right. \\
& - \exp(\phi) \log \{ s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \} \\
& \left. + \exp(\phi) \left\{ \frac{\exp(\phi^{(0)}) - \exp(\phi)}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right]
\end{aligned}$$

$$\begin{aligned}
& - \frac{0.5s_j}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{ \exp(\phi) - \exp(\phi^{(0)}) \}^2 \Big] \Big\}, \\
= & \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(-Y_{i,j} \left\{ \frac{\exp(\phi)}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right\} \right. \right. \\
& - \frac{s_j}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{ \exp(\phi) - \exp(\phi^{(0)}) \} \exp(\phi) \\
& + \frac{\exp(\phi) \{ \exp(\phi^{(0)}) - 2\exp(\phi) \}}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& - \exp(\phi) \log \{ s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \} \\
& + \exp(\phi) + \phi \exp(\phi) + \frac{\partial}{\partial \phi} \log [\Gamma \{ Y_{i,j} + \exp(\phi) \}] \\
& \left. - \frac{\partial}{\partial \phi} \log [\Gamma \{ \exp(\phi) \}] \right) \\
& + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \left[\exp(\phi) + \exp(\phi) \phi \right. \\
& - \exp(\phi) \log \left\{ s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \\
& + \frac{\exp(\phi) \{ \exp(\phi^{(0)}) - 2\exp(\phi) \}}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& \left. \left. - \frac{s_j}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{ \exp(\phi) - \exp(\phi^{(0)}) \} \exp(\phi) \right] \right\}, \\
= & \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(-\frac{Y_{i,j} \exp(\phi)}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \right. \\
& + \frac{\partial}{\partial \phi} \log [\Gamma \{ Y_{i,j} + \exp(\phi) \}] - \frac{\partial}{\partial \phi} \log [\Gamma \{ \exp(\phi) \}] \Big) \\
& + \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \left[\exp(\phi) + \exp(\phi) \phi \right. \\
& - \exp(\phi) \log \left\{ s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \\
& + \frac{\exp(\phi) \{ \exp(\phi^{(0)}) - 2\exp(\phi) \}}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& \left. \left. - \frac{s_j}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{ \exp(\phi) - \exp(\phi^{(0)}) \} \exp(\phi) \right] \right\},
\end{aligned}$$

Next,

$$\frac{\partial^2}{\partial \phi^2} \ell_1^\dagger(\phi | \Theta^{(0)}) = \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(-\frac{Y_{i,j} \exp(\phi)}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \right.$$

$$\begin{aligned}
& + \frac{\partial^2}{\partial \phi^2} \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] - \frac{\partial^2}{\partial \phi^2} \log[\Gamma\{\exp(\phi)\}] \Big) \\
& + \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \left[2 \exp(\phi) \right. \\
& + \exp(\phi) \phi - \exp(\phi) \log \left\{ s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \\
& + \frac{\exp(\phi) \{ \exp(\phi^{(0)}) - 4 \exp(\phi) \}}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& \left. - \frac{s_j}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{ 2 \exp(\phi) - \exp(\phi^{(0)}) \} \exp(\phi) \right] \Big\}.
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{\partial}{\partial \phi} \ell_1^\dagger(\phi | \Theta^{(0)}) |_{\Theta = \Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(- \frac{Y_{i,j} \exp(\phi^{(0)})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \right. \\
& + \frac{\partial}{\partial \phi} \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] |_{\phi = \phi^{(0)}} \\
& \left. - \frac{\partial}{\partial \phi} \log[\Gamma\{\exp(\phi)\}] |_{\phi = \phi^{(0)}} \right) \\
& + \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \left[\exp(\phi^{(0)}) + \exp(\phi^{(0)}) \phi^{(0)} \right. \\
& - \exp(\phi^{(0)}) \log \left\{ s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \\
& + \frac{\exp(\phi^{(0)}) \{ \exp(\phi^{(0)}) - 2 \exp(\phi^{(0)}) \}}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& - \frac{s_j}{s_j \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \{ \exp(\phi^{(0)}) - \exp(\phi^{(0)}) \} \\
& \left. \exp(\phi^{(0)}) \right] \Big\}, \\
& = \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(- \frac{Y_{i,j} \exp(\phi^{(0)})}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \right. \\
& + \frac{\partial}{\partial \phi} \log[\Gamma\{Y_{i,j} + \exp(\phi)\}] |_{\phi = \phi^{(0)}} \\
& \left. - \frac{\partial}{\partial \phi} \log[\Gamma\{\exp(\phi)\}] |_{\phi = \phi^{(0)}} \right)
\end{aligned}$$

$$\begin{aligned}
& + \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \times \left[\exp(\phi^{(0)}) + \exp(\phi^{(0)})\phi^{(0)} \right. \\
& - \exp(\phi^{(0)})\log \left\{ s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \\
& \left. - \frac{\exp(2\phi^{(0)})}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right] \Big\}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2}{\partial \phi^2} \ell_1^\dagger(\phi|\Theta^{(0)})|_{\Theta=\Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ I(Y_{i,j} > 0) \left(-\frac{Y_{i,j} \exp(\phi^{(0)})}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \right. \\
& + \frac{\partial^2}{\partial \phi^2} \log[\Gamma\{Y_{i,j} + \exp(\phi)\}]|_{\Theta=\Theta^{(0)}} \\
& \left. - \frac{\partial^2}{\partial \phi^2} \log[\Gamma\{\exp(\phi)\}]|_{\Theta=\Theta^{(0)}} \right) \\
& + \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \times \left[2 \exp(\phi^{(0)}) + \exp(\phi^{(0)})\phi^{(0)} \right. \\
& - \exp(\phi^{(0)})\log \left\{ s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)}) \right\} \\
& - \frac{3 \exp(2\phi^{(0)})}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \\
& \left. - \frac{s_j \exp(2\phi^{(0)})}{s_j\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right] \Big\}.
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \psi_0} g_1(\psi_0|\Theta_0) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[\left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
& - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}} \\
& \left. \times \exp\{(\psi_0 - \psi_0^{(0)})(2d + 2K + 7)\} \right],
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \psi_0} g_1(\psi_0|\Theta_0)|_{\Theta=\Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right. \\
& \left. - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}} \right],
\end{aligned}$$

$$\begin{aligned} \frac{\partial^2}{\partial \psi_0^2} g_1(\psi_0 | \Theta_0) &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[-\frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \right. \\ &\quad \left. \times (2d + 2K + 7) \exp\{(\psi_0 - \psi_0^{(0)})(2d + 2K + 7)\} \right], \end{aligned}$$

$$\frac{\partial^2}{\partial \psi_0^2} g_1(\psi_0 | \Theta_0) \big|_{\Theta=\Theta^{(0)}} = \sum_{i=1}^I \sum_{j=1}^{n_i} \left[-\frac{(2d + 2K + 7) \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \right],$$

$$\begin{aligned} \frac{\partial}{\partial \psi_1} g_2(\psi_1 | \Theta_0) &= \frac{\partial}{\partial \psi_1} \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\ &\quad \times \left[\psi_1 \log(s_i) - 0.5(d + K + 2)(\psi_1 - \psi^{(0)})^2 \right. \\ &\quad \left. + \psi_1 \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \right] \\ &\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \times \frac{1}{2d + 2K + 7} \left(\right. \\ &\quad \times \exp\{\log(s_i)(\psi_1 - \psi_1^{(0)})(2d + 2K + 7)\} \\ &\quad + 0.5(d + K + 1) \exp\{(\psi_1 - \psi_1^{(0)})^2\} \\ &\quad + 0.5 \exp\{(\psi_1 - \psi_1^{(0)})^2\} \\ &\quad + \exp \left[(\psi_1 - \psi_1^{(0)}) \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) \right. \right. \\ &\quad \left. \left. + X_{i,j}^\top \beta^{(0)} \right\} (2d + 2K + 7) \right] \Big) \Big\} \\ &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\ &\quad \times \left[\log(s_i) - (d + K + 2)(\psi_1 - \psi_1^{(0)}) \right. \\ &\quad \left. + \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \right] \\ &\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \times \frac{1}{2d + 2K + 7} \left(\right. \\ &\quad (2d + 2K + 7) \times \log(s_i) \times \exp\{\log(s_i)(\psi_1 - \psi_1^{(0)})(2d + 2K + 7)\} \\ &\quad \left. + (d + K + 1) \exp\{(\psi_1 - \psi_1^{(0)})^2\} \times (\psi_1 - \psi_1^{(0)}) \right. \end{aligned}$$

$$\begin{aligned}
& + \exp\{(\psi_1 - \psi_1^{(0)})^2\}(\psi_1 - \psi_1^{(0)}) \\
& + \exp\left[(\psi_1 - \psi_1^{(0)})\left\{\sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)}\right\}(2d + 2K + 7)\right] \\
& \times \left\{\sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)}\right\}(2d + 2K + 7)\bigg)\bigg\},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \psi_1} g_2(\psi_1 | \Theta_0) |_{\Theta = \Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left\{ \log(s_i) + \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \\
&\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \times \frac{1}{2d + 2K + 7} \left[\right. \\
&\quad (2d + 2K + 7) \times \log(s_i) \\
&\quad \left. \left. + \left\{ \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} (2d + 2K + 7) \right] \right) \\
&= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[\left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left\{ \log(s_i) + \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \\
&\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
&\quad \times \left\{ \log(s_i) + \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \bigg] \\
&= \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ \log(s_i) + \sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)} B_m(D_{i,j}) + X_{i,j}^\top \beta^{(0)} \right\} \\
&\quad \times \left[I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right. \\
&\quad \left. - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \right]
\end{aligned}$$

$$\frac{\partial^2}{\partial \psi_1^2} g_2(\psi_1 | \Theta_0) = \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ -(d + K + 2) \left\{ I(Y_{i,j} > 0) \right. \right.$$

$$\begin{aligned}
& + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \Big\} \\
& - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}} \times \frac{1}{2d + 2K + 7} \Big(\\
& (2d + 2K + 7)^2 \times \{\log(s_i)\}^2 \\
& \times \exp\{\log(s_i)(\psi_1 - \psi_1^{(0)})(2d + 2K + 7)\} \\
& + 2(d + K + 1) \exp\{(\psi_1 - \psi_1^{(0)})^2\} \times (\psi_1 - \psi_1^{(0)})^2 \\
& + (d + K + 1) \exp\{(\psi_1 - \psi_1^{(0)})^2\} \\
& + 2 \exp\{(\psi_1 - \psi_1^{(0)})^2\}(\psi_1 - \psi_1^{(0)})^2 \\
& + \exp\{(\psi_1 - \psi_1^{(0)})^2\} \\
& + \exp\left[(\psi_1 - \psi_1^{(0)})\left\{\sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)}B_m(D_{i,j}) + X_{i,j}^\top\beta^{(0)}\right\}(2d + 2K + 7)\right] \\
& \times \left\{\sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)}B_m(D_{i,j}) + X_{i,j}^\top\beta^{(0)}\right\}^2(2d + 2K + 7)^2 \Big) \Big\},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2}{\partial\psi_1^2}g_2(\psi_1|\Theta_0) \big|_{\Theta=\Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} \Big(-(d + K + 2) \Big\{ I(Y_{i,j} > 0) \\
& + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \Big\} \\
& - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)}\log(s_i\mu_{i,j}^{(0)})\}} \times \frac{1}{2d + 2K + 7} \Big[\\
& (2d + 2K + 7)^2 \times \{\log(s_i)\}^2 + d + K + 2 \\
& + \left\{\sum_{m=1}^{d+K+1} \gamma_{i,m}^{(0)}B_m(D_{i,j}) + X_{i,j}^\top\beta^{(0)}\right\}^2(2d + 2K + 7)^2 \Big] \Big)
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial}{\partial\gamma_{i,m}}g_{3,i}(\gamma_i|\Theta^{(0)}) \\
&= \frac{\partial}{\partial\gamma_{i,m}} \sum_{j=1}^{n_i} \Big\{ \Big\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \Big\} \\
& \times \left[-0.5 \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)})^2 B_m^2(D_{i,j}) + \psi_{\mathbf{1}}^{(0)} \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right] \Big\}
\end{aligned}$$

$$\begin{aligned}
& - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
& \times \frac{1}{2d + 2K + 7} \left(0.5 \sum_{m=1}^{d+K+1} \exp\{[(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})(2d + 2K + 7)]^2\} \right. \\
& \left. + \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) \psi_1^{(0)} B_m(D_{i,j})(2d + 2K + 7)\} \right) \Bigg\} \\
= & \sum_{j=1}^{n_i} \left\{ \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
& \times \left[-(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m^2(D_{i,j}) + \psi_1^{(0)} B_m(D_{i,j}) \right] \\
& - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
& \times \frac{1}{2d + 2K + 7} \left(\exp\{[(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})(2d + 2K + 7)]^2\} \right. \\
& \times (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \times B_m(D_{i,j}) \times (2d + 2K + 7)^2 \\
& + \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) \psi_1^{(0)} B_m(D_{i,j})(2d + 2K + 7)\} \\
& \left. \left. \times \psi_1^{(0)} B_m(D_{i,j})(2d + 2K + 7) \right) \right\}, \\
= & \sum_{j=1}^{n_i} B_m(D_{i,j}) \left\{ \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
& \times \left\{ -(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) + \psi_1^{(0)} \right\} \\
& - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
& \times \left(\exp\{[(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})(2d + 2K + 7)]^2\} \right. \\
& \times (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \times (2d + 2K + 7) \\
& + \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) \psi_1^{(0)} B_m(D_{i,j})(2d + 2K + 7)\} \\
& \left. \left. \times \psi_1^{(0)} \right) \right\},
\end{aligned}$$

$$\frac{\partial}{\partial \gamma_{i,m}} g_{3,i}(\gamma_i | \Theta^{(0)}) \Big|_{\Theta = \Theta^{(0)}}$$

$$\begin{aligned}
&= \psi_1^{(0)} \sum_{j=1}^{n_i} B_m(D_{i,j}) \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right. \\
&\quad \left. - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \right\} \\
&\quad \frac{\partial^2}{\partial \gamma_{i,m}^2} g_{3,i}(\gamma_i | \Theta^{(0)}) \\
&= \frac{\partial}{\partial \gamma_{i,m}} \sum_{j=1}^{n_i} B_m(D_{i,j}) \left\{ \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left\{ -(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) + \psi_1^{(0)} \right\} \\
&\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
&\quad \times \left(\exp[\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) (2d + 2K + 7)\}^2] \right. \\
&\quad \times (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \times (2d + 2K + 7) \\
&\quad + \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) \psi_1^{(0)} B_m(D_{i,j}) (2d + 2K + 7)\} \\
&\quad \left. \left. \times \psi_1^{(0)} \right) \right\} \\
&= \sum_{j=1}^{n_i} B_m(D_{i,j}) \left\{ \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left\{ -B_m(D_{i,j}) \right\} \\
&\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
&\quad \times \left(\exp[\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) (2d + 2K + 7)\}^2] \right. \\
&\quad \times 2(\gamma_{i,m} - \gamma_{i,m}^{(0)})^2 B_m^3(D_{i,j}) (2d + 2K + 7)^3 \\
&\quad + \exp[\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) (2d + 2K + 7)\}^2] \\
&\quad \times B_m(D_{i,j}) \times (2d + 2K + 7) \\
&\quad + \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) \psi_1^{(0)} B_m(D_{i,j}) (2d + 2K + 7)\} \\
&\quad \left. \left. \times \psi_1^{(0)} B_m(D_{i,j}) (2d + 2K + 7) \right) \right\}.
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial^2}{\partial \gamma_{i,m}^2} g_{3,i}(\gamma_i | \Theta^{(0)}) \big|_{\Theta = \Theta^{(0)}} \\
&= \sum_{j=1}^{n_i} B_m(D_{i,j}) \left\{ \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left. \left\{ -B_m(D_{i,j}) \right\} \right. \\
&\quad \times \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
&\quad \times \left. \left(B_m(D_{i,j}) \times (2d + 2K + 7) + \psi^{(0)}_1 B_m(D_{i,j}) (2d + 2K + 7) \right) \right\} \\
&= - \sum_{j=1}^{n_i} B_m^2(D_{i,j}) \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right. \\
&\quad \times \left. \left. \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \right\} \right. \\
&\quad \times \left. (2d + 2K + 7) \{1 + \psi^{(0)}_1\} \right\}.
\end{aligned}$$

Also,

$$\frac{\partial^2}{\partial \gamma_{i,m} \partial \gamma_{i,r}} g_{3,i}(\gamma_i | \Theta^{(0)}) \big|_{\Theta = \Theta^{(0)}} = 0.$$

$$\begin{aligned}
& \frac{\partial}{\partial \gamma_{i,r}} \ell_{3,i}^\dagger(\gamma_i | \Theta^{(0)}) \\
&= \frac{\partial}{\partial \gamma_{i,r}} \sum_{j=1}^{n_i} \left(I(Y_{i,j} > 0) Y_{i,j} \sum_{m=1}^{d+K+1} \gamma_{i,m} B_m(D_{i,j}) \right. \\
&\quad - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
&\quad \times \left. \left. \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \right. \\
&\quad \times \frac{\mu_{i,j}^{(0)}}{d + K + 2} \sum_{m=1}^{d+K+1} \exp\{(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) (d + K + 2)\} \\
&\quad - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
&\quad \times \left[-2(\mu_{i,j}^{(0)})^2 \sum_{m=1}^{d+K+1} (\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j}) \right. \\
&\quad \times \left. \left. \frac{(\mu_{i,j}^{(0)})^2}{d + K + 2} \sum_{m=1}^{d+K+1} \exp\{2(d + K + 2)(\gamma_{i,m} - \gamma_{i,m}^{(0)}) B_m(D_{i,j})\} \right] \right),
\end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^{n_i} \left(I(Y_{i,j} > 0) Y_{i,j} B_r(D_{i,j}) \right. \\
&\quad - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
&\quad \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
&\quad \times \frac{\mu_{i,j}^{(0)}}{d + K + 2} \exp\{(\gamma_{i,r} - \gamma_{i,r}^{(0)}) B_r(D_{i,j})(d + K + 2)\} B_r(D_{i,j})(d + K + 2) \\
&\quad - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
&\quad \times \left[-2(\mu_{i,j}^{(0)})^2 B_r(D_{i,j}) \right. \\
&\quad \left. + \frac{(\mu_{i,j}^{(0)})^2}{d + K + 2} \exp\{2(d + K + 2)(\gamma_{i,r} - \gamma_{i,r}^{(0)}) B_r(D_{i,j})\} 2(d + K + 2) B_r(D_{i,j}) \right] \Bigg), \\
&= \sum_{j=1}^{n_i} B_r(D_{i,j}) \left(I(Y_{i,j} > 0) Y_{i,j} \right. \\
&\quad - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
&\quad \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
&\quad \times \mu_{i,j}^{(0)} \exp\{(\gamma_{i,r} - \gamma_{i,r}^{(0)}) B_r(D_{i,j})(d + K + 2)\} \\
&\quad - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
&\quad \times \left[-(\mu_{i,j}^{(0)})^2 + (\mu_{i,j}^{(0)})^2 \exp\{2(d + K + 2)(\gamma_{i,r} - \gamma_{i,r}^{(0)}) B_r(D_{i,j})\} \right] \Bigg).
\end{aligned}$$

So,

$$\begin{aligned}
&\frac{\partial}{\partial \gamma_{i,r}} \ell_{3,i}^{\dagger}(\gamma_i | \Theta^{(0)}) \Big|_{\Theta = \Theta^{(0)}} \\
&= \sum_{j=1}^{n_i} B_r(D_{i,j}) \left[I(Y_{i,j} > 0) Y_{i,j} \right. \\
&\quad - \frac{s_j \mu_{i,j}^{(0)}}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
&\quad \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \Bigg].
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial^2}{\partial \gamma_{i,r}^2} \ell_{3,i}^\dagger(\gamma_i | \Theta^{(0)}) \\
= & \frac{\partial}{\partial \gamma_{i,r}} \sum_{j=1}^{n_i} B_r(D_{i,j}) \left(I(Y_{i,j} > 0) Y_{i,j} \right. \\
& - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& + \left. \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
& \times \mu_{i,j}^{(0)} \exp\{(\gamma_{i,r} - \gamma_{i,r}^{(0)}) B_r(D_{i,j})(d + K + 2)\} \\
& - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \times \left[-(\mu_{i,j}^{(0)})^2 + (\mu_{i,j}^{(0)})^2 \exp\{2(d + K + 2)(\gamma_{i,r} - \gamma_{i,r}^{(0)}) B_r(D_{i,j})\} \right] \Bigg) \\
= & \sum_{j=1}^{n_i} B_r(D_{i,j}) \left(-\frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \right. \\
& + \left. \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
& \times \mu_{i,j}^{(0)} B_r(D_{i,j})(d + K + 2) \exp\{(\gamma_{i,r} - \gamma_{i,r}^{(0)}) B_r(D_{i,j})(d + K + 2)\} \\
& - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \times 2(d + K + 2) B_r(D_{i,j})(\mu_{i,j}^{(0)})^2 \exp\{2(d + K + 2)(\gamma_{i,r} - \gamma_{i,r}^{(0)}) B_r(D_{i,j})\} \Bigg).
\end{aligned}$$

Now,

$$\begin{aligned}
& \frac{\partial^2}{\partial \gamma_{i,r}^2} \ell_{3,i}^\dagger(\gamma_i | \Theta^{(0)}) \Big|_{\Theta = \Theta^{(0)}} \\
= & -(d + K + 2) \sum_{j=1}^{n_i} B_r^2(D_{i,j}) \left[\frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} \right. \right. \\
& + I(Y_{i,j} > 0) \exp(\phi^{(0)}) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \Big\} \mu_{i,j}^{(0)} \\
& + \frac{2s_j (\mu_{i,j}^{(0)})^2}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \Bigg] \\
& \frac{\partial^2}{\partial \gamma_{i,r} \gamma_{i,m}} \ell_{3,i}^\dagger(\gamma_i | \Theta^{(0)}) \Big|_{\Theta = \Theta^{(0)}} = 0.
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta} g_4(\beta | \Theta^{(0)}) &= \frac{\partial}{\partial \beta} \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left[-0.5 \{X_{i,j}^\top (\beta - \beta^{(0)})\}^2 + \psi_1^{(0)} X_{i,j}^\top (\beta - \beta^{(0)}) \right] \\
&\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
&\quad \times \frac{1}{2d + 2K + 7} \left[0.5 \exp\{\{X_{i,j}^\top (\beta - \beta^{(0)}) (2d + 2K + 7)\}^2\} \right. \\
&\quad \left. \left. + \exp\{X_{i,j}^\top \psi_1^{(0)} (\beta - \beta^{(0)}) (2d + 2K + 7)\} \right] \right). \\
&= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left[-\{X_{i,j}^\top (\beta - \beta^{(0)})\} X_{i,j} + \psi_1^{(0)} X_{i,j} \right] \\
&\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
&\quad \times \frac{1}{2d + 2K + 7} \left[\exp\{\{X_{i,j}^\top (\beta - \beta^{(0)}) (2d + 2K + 7)\}^2\} \right. \\
&\quad \times X_{i,j}^\top (\beta - \beta^{(0)}) X_{i,j} (2d + 2K + 7)^2 \\
&\quad + \exp\{X_{i,j}^\top \psi_1^{(0)} (\beta - \beta^{(0)}) (2d + 2K + 7)\} \\
&\quad \left. \left. \times \psi_1^{(0)} X_{i,j} (2d + 2K + 7) \right] \right). \\
&= \sum_{i=1}^I \sum_{j=1}^{n_i} X_{i,j} \left(\left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\
&\quad \times \left[-X_{i,j}^\top (\beta - \beta^{(0)}) + \psi_1^{(0)} \right] \\
&\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\
&\quad \times \left[\exp\{\{X_{i,j}^\top (\beta - \beta^{(0)}) (2d + 2K + 7)\}^2\} X_{i,j}^\top (\beta - \beta^{(0)}) (2d + 2K + 7) \right. \\
&\quad \left. \left. + \exp\{X_{i,j}^\top \psi_1^{(0)} (\beta - \beta^{(0)}) (2d + 2K + 7)\} \times \psi_1^{(0)} \right] \right).
\end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \beta} g_4(\beta | \Theta^{(0)}) |_{\Theta = \Theta^{(0)}} &= \psi_1^{(0)} \sum_{i=1}^I \sum_{j=1}^{n_i} X_{i,j} \left\{ I(Y_{i,j} > 0) \right. \\ &\quad + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \\ &\quad \left. - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \right\}. \end{aligned}$$

$$\begin{aligned} \frac{\partial^2}{\partial \beta \partial \beta^\top} g_4(\beta | \Theta^{(0)}) &= \sum_{i=1}^I \sum_{j=1}^{n_i} X_{i,j} \left(\left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right. \\ &\quad \times (-X_{i,j}^\top) \\ &\quad - \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\ &\quad \times \left[\exp[\{X_{i,j}^\top (\beta - \beta^{(0)}) (2d + 2K + 7)\}^2] \right. \\ &\quad \times 2\{X_{i,j}^\top (\beta - \beta^{(0)})\}^2 (2d + 2K + 7)^3 X_{i,j}^\top \\ &\quad + \exp[\{X_{i,j}^\top (\beta - \beta^{(0)}) (2d + 2K + 7)\}^2] X_{i,j}^\top (2d + 2K + 7) \\ &\quad + \exp\{X_{i,j}^\top \psi_1^{(0)} (\beta - \beta^{(0)}) (2d + 2K + 7)\} (\psi_1^{(0)})^2 \\ &\quad \left. \left. \times X_{i,j}^\top (2d + 2K + 7) \right] \right) \\ &= - \sum_{i=1}^I \sum_{j=1}^{n_i} X_{i,j} X_{i,j}^\top \left(\left\{ I(Y_{i,j} > 0) \right. \right. \\ &\quad + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \left. \right\} \\ &\quad + \frac{(2d + 2K + 7) \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \\ &\quad \times \left[\exp[\{X_{i,j}^\top (\beta - \beta^{(0)}) (2d + 2K + 7)\}^2] \right. \\ &\quad \times 2\{X_{i,j}^\top (\beta - \beta^{(0)})\}^2 (2d + 2K + 7)^2 \\ &\quad + \exp[\{X_{i,j}^\top (\beta - \beta^{(0)}) (2d + 2K + 7)\}^2] \\ &\quad \left. \left. + \exp\{X_{i,j}^\top \psi_1^{(0)} (\beta - \beta^{(0)}) (2d + 2K + 7)\} \times (\psi_1^{(0)})^2 \right] \right). \end{aligned}$$

$$\begin{aligned}
\frac{\partial^2}{\partial \beta^2} g_4(\beta | \Theta^{(0)}) \big|_{\Theta = \Theta^{(0)}} &= - \sum_{i=1}^I \sum_{j=1}^{n_i} X_{i,j} X_{i,j}^\top \left\{ I(Y_{i,j} > 0) \right. \\
&\quad + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \\
&\quad + (2d + 2K + 7) \{1 + (\psi_1^{(0)})^2\} \\
&\quad \left. \times \frac{\exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}}{1 + \exp\{\psi_0^{(0)} + \psi_1^{(0)} \log(s_i \mu_{i,j}^{(0)})\}} \right\}.
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta} \ell_4^\dagger(\beta | \Theta_0) &= \frac{\partial}{\partial \beta} \sum_{i=1}^I \sum_{j=1}^{n_i} \left(I(Y_{i,j} > 0) Y_{i,j} X_{i,j}^\top \beta \right. \\
&\quad - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
&\quad \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
&\quad \times \frac{\mu_{i,j}^{(0)}}{d + K + 2} \exp\{X_{i,j}^\top (\beta - \beta^{(0)}) (d + K + 2)\} \\
&\quad - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
&\quad \times \left[-2(\mu_{i,j}^{(0)})^2 X_{i,j}^\top (\beta - \beta^{(0)}) \right. \\
&\quad \left. + \frac{(\mu_{i,j}^{(0)})^2}{d + K + 2} \exp\{2(d + K + 2) X_{i,j}^\top (\beta - \beta^{(0)})\} \right] \Bigg) \\
&= \sum_{i=1}^I \sum_{j=1}^{n_i} \left(I(Y_{i,j} > 0) Y_{i,j} X_{i,j} \right. \\
&\quad - \frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
&\quad \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
&\quad \times \frac{\mu_{i,j}^{(0)}}{d + K + 2} \exp\{X_{i,j}^\top (\beta - \beta^{(0)}) (d + K + 2)\} \times X_{i,j} (d + K + 2) \\
&\quad - \frac{0.5 s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
&\quad \times \left[-2(\mu_{i,j}^{(0)})^2 X_{i,j} + \frac{(\mu_{i,j}^{(0)})^2}{d + K + 2} \exp\{2(d + K + 2) X_{i,j}^\top (\beta - \beta^{(0)})\} \right] \Bigg)
\end{aligned}$$

$$\begin{aligned}
& \times 2(d + K + 2)X_{i,j} \Big] \Big) \\
& = \sum_{i=1}^I \sum_{j=1}^{n_i} X_{ij} \left(I(Y_{i,j} > 0)Y_{i,j} \right. \\
& \quad - \frac{s_j}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0)Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& \quad \left. + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
& \quad \times \mu_{i,j}^{(0)} \exp\{X_{i,j}^\top(\beta - \beta^{(0)})(d + K + 2)\} \\
& \quad - \frac{s_j(\mu_{i,j}^{(0)})^2}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \quad \left. \times \left[-1 + \exp\{2(d + K + 2)X_{i,j}^\top(\beta - \beta^{(0)})\} \right] \right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta} \ell_4^\dagger(\beta | \Theta_0) \big|_{\Theta = \Theta^{(0)}} &= \sum_{i=1}^I \sum_{j=1}^{n_i} X_{ij} \left[I(Y_{i,j} > 0)Y_{i,j} \right. \\
& \quad - \frac{s_j\mu_{i,j}^{(0)}}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0)Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& \quad \left. + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \Big].
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2}{\partial \beta \partial \beta^\top} \ell_4^\dagger(\beta | \Theta_0) &= \sum_{i=1}^I \sum_{j=1}^{n_i} X_{ij} \left(-\frac{s_j}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \\
& \quad \left\{ I(Y_{i,j} > 0)Y_{i,j} \right. \\
& \quad + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \\
& \quad + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \Big\} \\
& \quad \times \mu_{i,j}^{(0)} \exp\{X_{i,j}^\top(\beta - \beta^{(0)})(d + K + 2)\} X_{i,j}^\top(d + K + 2) \\
& \quad - \frac{s_j(\mu_{i,j}^{(0)})^2}{s_i\mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) \right. \\
& \quad \left. + \frac{I(Y_{i,j} = 0)\omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)}G(\phi^{(0)}, \mu_{i,j}^{(0)})} \Big\}
\end{aligned}$$

$$\begin{aligned}
& \times \left[\exp\{2(d+K+2)X_{i,j}^\top(\beta - \beta^{(0)})\} \times 2(d+K+2) \times X_{i,j}^\top \right] \Bigg) \\
& = -(d+K+2) \sum_{i=1}^I \sum_{j=1}^{n_i} X_{i,j} X_{i,j}^\top \left(\frac{s_j}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \\
& \quad \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& \quad \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
& \quad \times \mu_{i,j}^{(0)} \exp\{X_{i,j}^\top(\beta - \beta^{(0)})(d+K+2)\} \\
& \quad + \frac{s_j (\mu_{i,j}^{(0)})^2}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) \right. \\
& \quad \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \\
& \quad \left. \times \left[2 \exp\{2(d+K+2)X_{i,j}^\top(\beta - \beta^{(0)})\} \right] \right).
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial^2}{\partial \beta \partial \beta^\top} \ell_4^\dagger(\beta | \Theta_0) \Big|_{\Theta = \Theta^{(0)}} \\
& = -(d+K+2) \sum_{i=1}^I \sum_{j=1}^{n_i} X_{i,j} X_{i,j}^\top \left[\frac{s_j \mu_{i,j}^{(0)}}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \right. \\
& \quad \left\{ I(Y_{i,j} > 0) Y_{i,j} + I(Y_{i,j} > 0) \exp(\phi^{(0)}) \right. \\
& \quad \left. + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \times \exp(\phi^{(0)}) \right\} \\
& \quad \left. + \frac{2s_j (\mu_{i,j}^{(0)})^2}{s_i \mu_{i,j}^{(0)} + \exp(\phi^{(0)})} \left\{ I(Y_{i,j} > 0) + \frac{I(Y_{i,j} = 0) \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})}{(1 - \omega_{i,j}^{(0)}) + \omega_{i,j}^{(0)} G(\phi^{(0)}, \mu_{i,j}^{(0)})} \right\} \right].
\end{aligned}$$

APPENDIX C

KERNELIZED SIGNED GRAPH LEARNING FOR SINGLE CELL GENE REGULATORY NETWORK INFERENCE

C.1 Optimization Algorithm for Signed Graph Learning

In this section, we present an ADMM based algorithm to solve the optimization problem for signed graph learning. For convenience, we include the optimization problem below:

$$\begin{aligned}
& \min_{\mathbf{L}^+, \mathbf{L}^- \in \mathbb{L}} \text{trace}(\mathbf{K}\mathbf{L}^+) - \text{trace}(\mathbf{K}\mathbf{L}^-) + \alpha_1 \|\mathbf{L}^+\|_F^2 + \alpha_2 \|\mathbf{L}^-\|_F^2 \\
& \text{s.t.} \quad \text{trace}(\mathbf{L}^+) = 2n, \text{trace}(\mathbf{L}^-) = 2n \\
& \quad L_{ij}^+ = 0 \text{ if } L_{ij}^- \neq 0 \text{ and } L_{ij}^- = 0 \text{ if } L_{ij}^+ \neq 0 \quad \forall i \neq j.
\end{aligned} \tag{C.1}$$

This problem is non-convex due to the last two constraints, which are called complementarity constraints [157]. In [158], it is shown that alternating direction method of multipliers (ADMM) converges for problems with complementarity constraints under some assumptions. First, we rewrite the problem in vector form. Let $\text{upper}(\cdot)$ be an operator that takes an $n \times n$ matrix and returns a $n(n-1)/2$ -dimensional vector that corresponds to the upper triangular part of the input matrix. Define $\text{diag}(\mathbf{x})$ as an operator which returns a diagonal matrix with the diagonal elements equal to the input vector $\mathbf{x}$. Similarly, $\text{diag}(\mathbf{X})$ returns the diagonal of the input matrix $\mathbf{X}$ as a vector. The matrix $\mathbf{P} \in \mathbb{R}^{n \times n(n-1)/2}$ is defined such that $\mathbf{P}\text{upper}(\mathbf{A}) = \mathbf{A}\mathbf{1}$ where $\mathbf{A}$ is a symmetric matrix whose diagonal entries are equal to zero. Let $\mathbf{k} = \text{upper}(\mathbf{K})$, $\mathbf{d} = \text{diag}(\mathbf{K})$, $\boldsymbol{\ell}^+ = \text{upper}(\mathbf{L}^+)$, $\boldsymbol{\ell}^- = \text{upper}(\mathbf{L}^-)$. Thus, (C.1) can be rewritten as:

$$\begin{aligned}
& \min_{\boldsymbol{\ell}^+ \geq \mathbf{0}, \boldsymbol{\ell}^- \leq \mathbf{0}} \langle 2\mathbf{k} - \mathbf{P}^\top \mathbf{d}, \boldsymbol{\ell}^+ \rangle - \langle 2\mathbf{k} - \mathbf{P}^\top \mathbf{d}, \boldsymbol{\ell}^- \rangle + \alpha_1 \langle (2\mathbf{I} + \mathbf{P}^\top \mathbf{P}) \boldsymbol{\ell}^+, \boldsymbol{\ell}^+ \rangle \\
& \quad + \alpha_2 \langle (2\mathbf{I} + \mathbf{P}^\top \mathbf{P}) \boldsymbol{\ell}^-, \boldsymbol{\ell}^- \rangle \\
& \text{s.t.} \quad \mathbf{1}^\top \boldsymbol{\ell}^+ = -n, \mathbf{1}^\top \boldsymbol{\ell}^- = -n, \text{ and } \boldsymbol{\ell}^+ \perp \boldsymbol{\ell}^-,
\end{aligned} \tag{C.2}$$

where the first two terms correspond to trace terms in (C.1), the last two terms correspond to Frobenius terms of (C.1) and first two constraints are the same as the first two constraints of (C.1). The last constraint with $\ell^+ \leq \mathbf{0}$ and $\ell^- \leq \mathbf{0}$ correspond to the complementarity constraints. By introducing two slack variables $\mathbf{v} = \ell^+$ and $\mathbf{w} = \ell^-$, the problem is written in standard ADMM form:

$$\begin{aligned} \min_{\mathbf{v}, \mathbf{w}, \ell^+, \ell^-} \quad & \iota_S(\mathbf{v}, \mathbf{w}) + h(\ell^+, \ell^-) + \iota_H(\ell^+) + \iota_H(\ell^-) \\ \text{s.t.} \quad & \mathbf{v} - \ell^+ = \mathbf{0}, \quad \mathbf{w} - \ell^- = \mathbf{0}, \end{aligned} \quad (\text{C.3})$$

where $\iota_S(\cdot)$ is the indicator function for the complementarity set $S = \{(\mathbf{v}, \mathbf{w}) : \mathbf{v} \leq 0, \mathbf{w} \leq 0, \mathbf{v} \perp \mathbf{w}\}$, $h(\ell^+, \ell^-)$ is the objective function in (C.2), and $\iota_H(\cdot)$ is the indicator function for the hyperplane $H = \{\ell : \mathbf{1}^\top \ell = -n\}$. The augmented Lagrangian of (C.3) is:

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{v}, \mathbf{w}, \ell^+, \ell^-, \lambda_1, \lambda_2) = & \iota_S(\mathbf{v}, \mathbf{w}) + h(\ell^+, \ell^-) + \iota_H(\ell^+) + \iota_H(\ell^-) \\ & + \lambda_1^\top (\mathbf{v} - \ell^+) + \frac{\rho}{2} \|\mathbf{v} - \ell^+\|_2^2 + \lambda_2^\top (\mathbf{w} - \ell^-) + \frac{\rho}{2} \|\mathbf{w} - \ell^-\|_2^2, \end{aligned} \quad (\text{C.4})$$

where λ_1 and λ_2 are Lagrange multipliers and $\rho > 0$ is the Augmented Lagrangian parameter.

($\mathbf{v}, \mathbf{w}$)-step: The ($\mathbf{v}, \mathbf{w}$)-step of ADMM can be found as the projection onto the complementarity set S :

$$(\mathbf{v}^{k+1}, \mathbf{w}^{k+1}) = \underset{\mathbf{v}, \mathbf{w}}{\operatorname{argmin}} \iota_S(\mathbf{v}, \mathbf{w}) + \frac{\rho}{2} \|\mathbf{v} - \ell^{+k} + \frac{\lambda_1^k}{\rho}\|_2^2 + \frac{\rho}{2} \|\mathbf{w} - \ell^{-k} + \frac{\lambda_2^k}{\rho}\|_2^2 = \Pi_S(\mathbf{y}), \quad (\text{C.5})$$

where $\mathbf{y} = [(\ell^{+k} - \lambda_1^k/\rho)^\top, (\ell^{-k} - \lambda_2^k/\rho)^\top]^\top$ and $\Pi_S(\cdot)$ is the projection operator on the set S .

(ℓ^+, ℓ^-)-step: Using the fact that optimization can be performed separately for ℓ^+ and ℓ^- , ℓ^+ -step can be written as:

$$\begin{aligned} \ell^{+k+1} = & \underset{\ell^+}{\operatorname{argmin}} \quad \mathbf{z}^\top \ell^+ + \alpha_1 \langle (2\mathbf{I} + \mathbf{P}^\top \mathbf{P}) \ell^+, \ell^+ \rangle + \iota_H(\ell^+) + \frac{\rho}{2} \|\mathbf{v}^{k+1} - \ell^+ + \frac{\lambda_1^k}{\rho}\|_2^2 \\ = & \Pi_H[((4\alpha_1 + \rho)\mathbf{I} + 2\alpha_1 \mathbf{P}^\top \mathbf{P})^{-1}(\rho \mathbf{v}^{k+1} + \lambda_1^k - \mathbf{z})], \end{aligned} \quad (\text{C.6})$$

where $\mathbf{z} = 2\mathbf{k} - \mathbf{P}^\top \mathbf{d}$ and $\Pi_H(\cdot)$ is the projection operator on the hyperplane H . Similarly, ℓ^- -step can be written as:

$$\ell^{-k+1} = \Pi_H[((4\alpha_2 + \rho)\mathbf{I} + 2\alpha_2 \mathbf{P}^\top \mathbf{P})^{-1}(\rho \mathbf{w}^{k+1} + \lambda_2^k + \mathbf{z})]. \quad (\text{C.7})$$

Lagrange multipliers update: The updates of Lagrange multipliers are:

$$\lambda_1^{k+1} = \lambda_1^k + \rho(\mathbf{v}^{k+1} - \boldsymbol{\ell}^{+k+1}), \quad (\text{C.8})$$

$$\lambda_2^{k+1} = \lambda_2^k + \rho(\mathbf{w}^{k+1} - \boldsymbol{\ell}^{-k+1}). \quad (\text{C.9})$$

C.1.1 Computational and Storage Complexity

Computational complexity of the optimization procedure described above can be found by determining how many computations are required for each ADMM step. Let $M = n(n-1)/2$ where n is the number of genes. $(\mathbf{v}, \mathbf{w})$ -step can be performed in $O(M)$ time, or $O(n^2)$ time. $(\boldsymbol{\ell}^+, \boldsymbol{\ell}^-)$ -step requires the inversion of the matrix $(4\alpha_1 + \rho)\mathbf{I} + 2\alpha_1\mathbf{P}^\top\mathbf{P}$, which needs to be calculated only once before the optimization iterations. The inverse matrix has a closed form solution which can be found using Woodbury matrix identity. It has a decomposition of the form $\mathbf{A}^\top\mathbf{A}$ where $\mathbf{A}$ is a sparse matrix with $O(n^2)$ non-zero entries. Thus, matrix-vector multiplication of $(\boldsymbol{\ell}^+, \boldsymbol{\ell}^-)$ -step can be done in $O(n^2)$ time. Updates of Lagrangian multipliers can also be performed $O(M)$ time, or $O(n^2)$. Let I be the number of iterations required for the convergence of ADMM. Thus, overall time complexity of scSGL is $O(In^2)$. The storage complexity of scSGL is determined by the size of the inverse matrix required in $(\boldsymbol{\ell}^+, \boldsymbol{\ell}^-)$ -step. Since this matrix has a decomposition of the form $\mathbf{A}^\top\mathbf{A}$, we only need to store $\mathbf{A}$. Thus, the storage complexity of scSGL is $O(n^2)$.

The computational and storage complexity of ADMM is quadratic in the number of genes and is not affected by the number of cells. Note that, scSGL also requires the construction of the kernel matrix before running the optimization. Since there are already very efficient tools to construct kernel matrices [130], we did not include their complexity in the analysis above. Finally, there are recent works in GSP literature for scaling GL methods to learning graphs with millions of nodes [159]. These approaches can be employed to scale scSGL, which we left as a future pursuit.

C.2 AUROC and EPR Results

In the main text, we consider AUPRC based metrics defined above as the main performance metrics for comparison due to inherent sparsity of GRNs. In this section, AUROC and EPR values for synthetic data used in parameter sensitivity analysis are reported in Figures C.1 and C.2, respectively. AUROC and EPR ratios for real datasets are also reported in Figure C.3.



Figure C.1 AUROC values of various methods for synthetic datasets with three different topologies (random, modular and hub) and varying number of genes (top row), dropout ratio (middle row) and number of cells (bottom row).

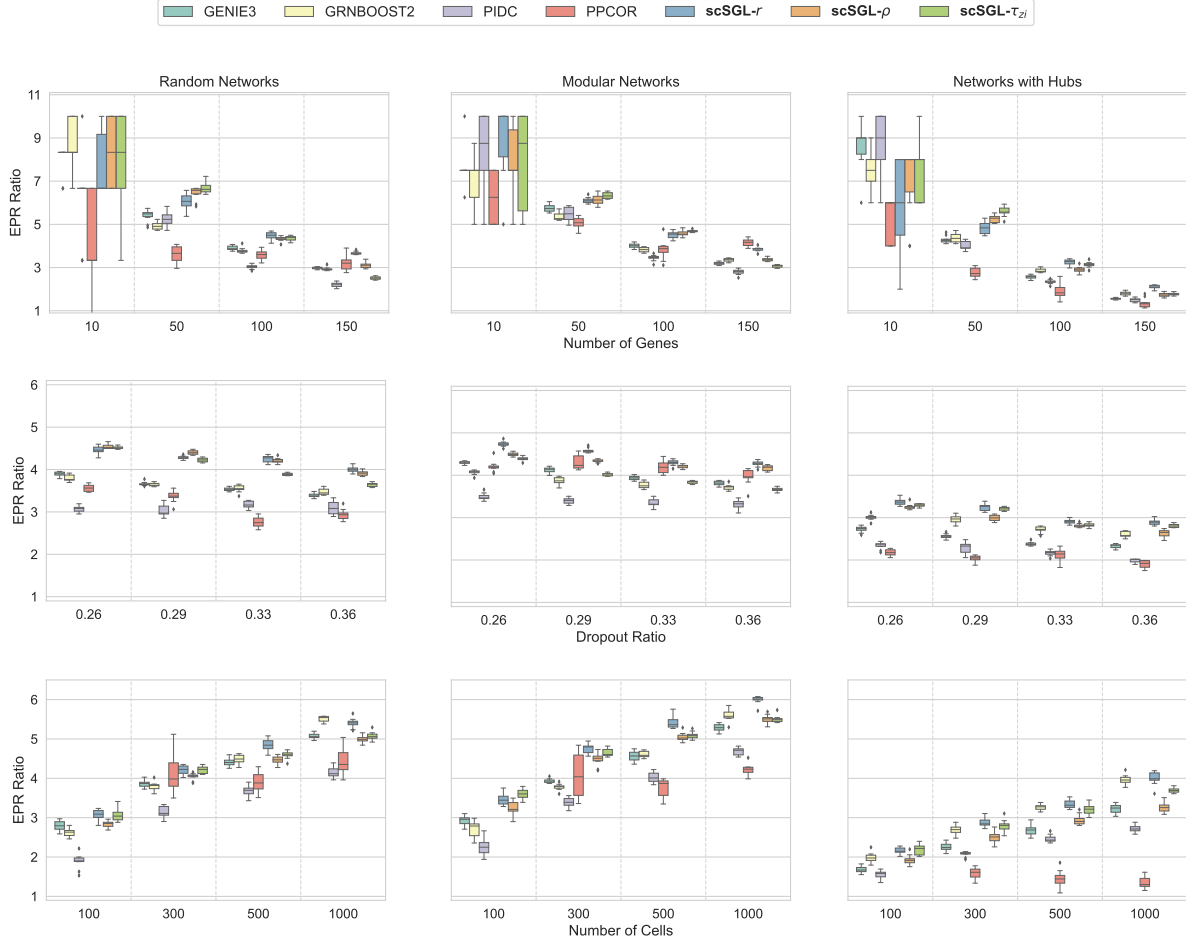


Figure C.2 EPR values of various methods for synthetic datasets with three different topologies (random, modular and hub) and varying number of genes (top row), dropout ratio (middle row) and number of cells (bottom row).

C.3 Synthetic Networks from BEELINE Pipeline

In the main text, we studied the performance of scSGL on curated datasets from BEELINE pipeline [3]. Simulated scRNAseq datasets generated from six synthetic networks using BoolODE are considered. For the sake of completeness, we report the performance of scSGL and benchmarking methods on these datasets here. These six networks correspond to different GRNs, that can produce a variety of trajectories observed in differentiating and developing cells (refer to [3] for the details of these networks). Each dataset includes 10

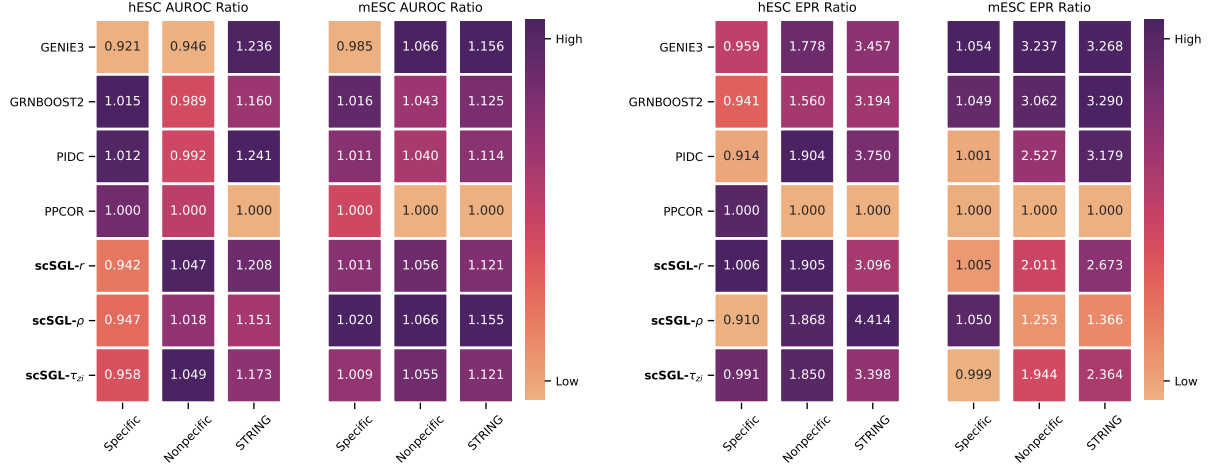


Figure C.3 AUROC and EPR ratios of methods for two real-world scRNAseq datasets. Inferred graphs are compared to three different gene regulatory databases.

realizations of an expression data with 500 cells. Compared to curated datasets analyzed in the main text, these datasets do not include any dropouts. We calculated AUPRC ratios for activating and inhibitory edges separately and average of 10 realizations are reported in Figure C.4.

The figure indicates that scSGL along with PPCOR are the best performing methods for inference of activating edges. For BF, BFC and CY; PPCOR followed by scSGL- r have the highest AUPRC ratios. On the other hand, scSGL- r followed by other kernels and PPCOR have the highest performances for LI, LL and TF. For the inference of inhibitory edges, the best performing method varies across networks. In BF, GRNBOOST2 shows the best performance; in BFC and TF, scSGL- τ and GRNBOOST2 have the highest AUPRC ratios; and for the remaining datasets scSGL- r followed by PPCOR perform better than others. In [3], it is observed that methods perform well on linear networks (LI and LL); while the inference in the remaining networks is harder. AUPRC ratios reported in Figure C.4 are inline with this observation, where AUPRC ratios for linear networks are generally higher than those for BF, BFC, CY and TF. Overall, scSGL- r along with PPCOR are the best performing methods, if the results on activating and inhibitory edges are evaluated together. Finally, when performances of kernels are compared, it can be seen that correlation kernel

shows the highest performance, followed by zero-inflated Kendall.

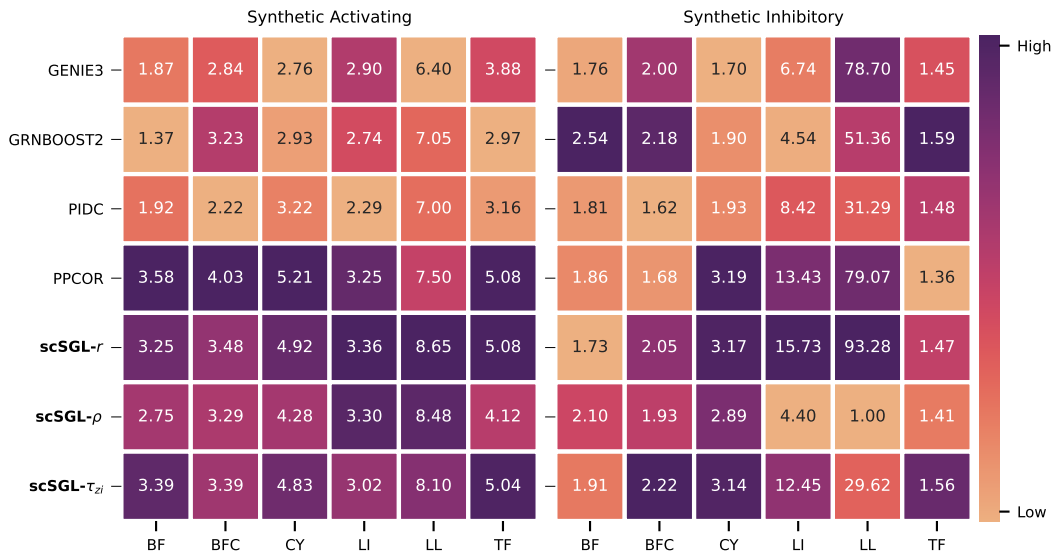


Figure C.4 Performance of scSGL and state-of-the-art methods on curated datasets as measured by AUPRC ratios for activating and inhibitory edges. Each column corresponds to a synthetic network. Abbreviations: LI, linear; CY, cycle; LL, linear long; BF, bifurcating; BFC, bifurcating converging and TF, trifurcating.

C.4 Cell-Type Specific GRN Inference

scSGL is developed based on the assumption that all cells are related to a single GRN. However, single cell datasets are generally a combination of cells arising from varying cell-types, and therefore may necessitate the inference of cell type specific GRNs. Cell-type specific GRN's can be learned in our framework by adding a cell-type clustering step before applying scSGL. One could either group the datasets by using cluster labels provided by the original authors of the experimental study or by clustering the dataset using one of the many clustering algorithms proposed for single cell data [160] (in case pre-defined cell-labels are absent). Assuming independence within cell-groups, we could estimate cell-type specific networks using scSGL for each cluster separately.

In this section, we demonstrate the process for using scSGL to learn cell type specific GRNs and apply this process to the differentiation dataset hESC. We apply scSGL seper-

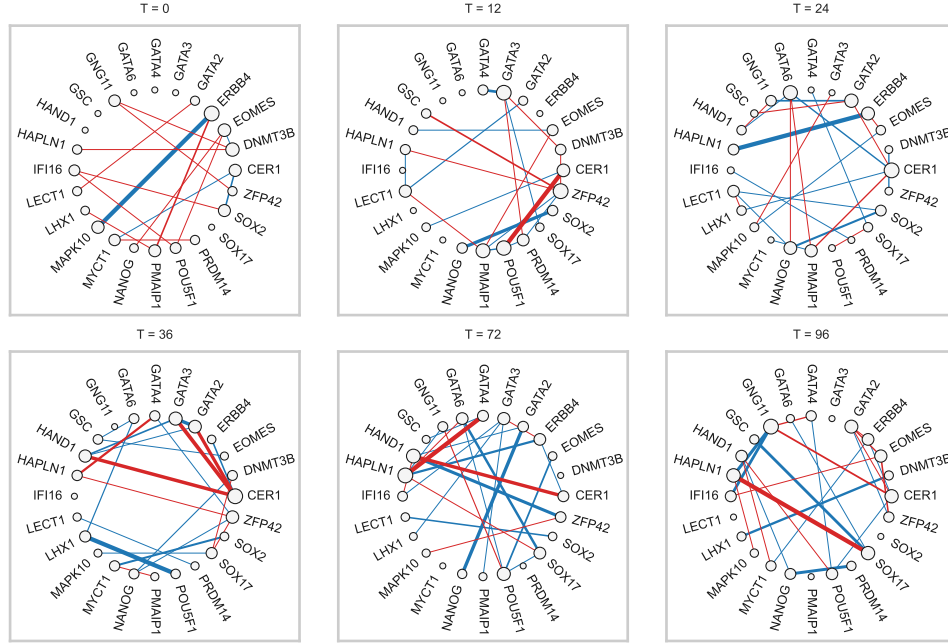


Figure C.5 Edges detected using scSGL- r between 24 Lineage marker genes of hESC at different time points of the differentiation process. Only edges whose absolute edge weights fall into top 10 percent are shown. Edge thicknesses are proportional to their weights, and node sizes are proportional to their degrees.

ately to the hESC dataset clustered by days (0,12,24,36,72 and 96 hrs) and learn scSGL graphs between 24 lineage-specific marker genes [143] at these different time points. Figure C.5 demonstrates the absence of edges from the Gata-family binding proteins at 0h. Gata family binding proteins have been reported as necessary for the development and function of a number of endoderm-derived tissues and cells [146, 161]. Onset of Gata4 and Gata6 expression has been reported to be coincident with the beginning of endoderm gene expression hence the absence of edges from Gata 4 and Gata6 at 0h are indicative of the undifferentiated nature of the single cells [146]. Weak interactions start to emerge at 12 hours of differentiation with inhibition of pluripotency markers Nanog and Sox2. At 24 h of differentiation, we notice a stronger inhibition of Nanog and pluripotency marker Pmaip1 by Gata6, indicating a transition of the cells towards a primitive streak state. Hand1 has been reported to play an essential role in both trophoblast-giant cells differentiation and in cardiac morphogenesis [162]. Inhibition of known DE marker Cer1 by Hand1 and Gata family TF's at 36 and 72

h of differentiation indicates an advanced state of differentiation. The appearance of key DE markers Gata2, Gata4, Gata6, Cer1 and Eomes as hub-nodes in 96-h time point network indicates that the cells have progressed toward the definitive endoderm (DE) state. This analysis clearly demonstrates that scSGL identifies gene network changes from data clustered over time points.

We acknowledge that analyzing the dataset in this manner does not exploit the similarity between the true cell-type specific networks and estimating a single network for the different cell-types ignores the fact that we do not expect the cell-type specific graphs to be identical. Our optimization framework can be extended to jointly learn Laplacians estimated from multiple cell groups but that is out of scope for the current paper and will be considered in future research.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] Rance Nault, Satabdi Saha, Sudin Bhattacharya, Jack Dodson, Samiran Sinha, Tapabrata Maiti, and Tim Zacharewski. Benchmarking of a bayesian single cell rnaseq differential gene expression test for dose–response study designs. *Nucleic acids research*, 50(8):e48–e48, 2022.
- [2] Abdullah Karaaslanli, Satabdi Saha, Selin Aviyente, and Tapabrata Maiti. scsgl: kernelized signed graph learning for single-cell gene regulatory network inference. *Bioinformatics*, 38(11):3011–3019, 2022.
- [3] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.
- [4] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome Biol*, 18(1):174, 2017.
- [5] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 2013.
- [6] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25(10):1491–1498, 2015.
- [7] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.
- [8] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.
- [9] Charles Gawad, Winston Koh, and Stephen R Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188, 2016.
- [10] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.
- [11] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. Zinb-wave: A general and flexible method for signal extraction from single-cell rna-seq data. *BioRxiv*, page 125112, 2017.
- [12] Kwangbom Choi, Yang Chen, Daniel A Skelly, and Gary A Churchill. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome biology*, 21(1):1–16, 2020.

- [13] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):1–10, 2015.
- [14] Andrew McDavid, Greg Finak, Pratip K Chattopadhyay, Maria Dominguez, Laurie Lamoreaux, Steven S Ma, Mario Roederer, and Raphael Gottardo. Data exploration, quality control and testing in single-cell qpcr-based gene expression experiments. *Bioinformatics*, 29(4):461–467, 2013.
- [15] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature genetics*, 53(6):770–777, 2021.
- [16] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13, 2015.
- [17] Andrew McDavid, Raphael Gottardo, Noah Simon, and Mathias Drton. Graphical models for zero-inflated single cell gene expression. *The annals of applied statistics*, 13(2):848, 2019.
- [18] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.
- [19] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.
- [20] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):1–17, 2014.
- [21] Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.
- [22] Justin D Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A David. Naught all zeros in sequence count data are the same. *Computational and structural biotechnology journal*, 18:2789, 2020.
- [23] KS Crump, DG Hoel, CH Langley, and R Peto. Fundamental carcinogenic processes and their implications for low dose risk assessment. *Cancer research*, 36(9_Part_1):2973–2979, 1976.
- [24] Food, Drug Administration, et al. Guidance for industry: exposure-response relationships-study design, data analysis, and regulatory applications. <http://www.fda.gov/cber/gdlns/exposure.pdf>, 2003.

- [25] L Martin. Benchmark dose software (bmds) version 2.1 user’s manual version 2.0. *Washington, DC: United States Environmental Protection Agency, Office of Environmental Information*, 2009.
- [26] National Toxicology Program et al. Ntp research report on national toxicology program approach to genomic dose-response modeling: Research report 5 [internet]. *NTP Research Report on National Toxicology Program*, 2018.
- [27] Qiang Zhang, W Michael Caudle, Jingbo Pi, Sudin Bhattacharya, Melvin E Andersen, Norbert E Kaminski, and Rory B Conolly. Embracing systems toxicology at single-cell resolution. *Current opinion in toxicology*, 16:49–57, 2019.
- [28] J Allen Davis, Jeffrey S Gift, and Q Jay Zhao. Introduction to benchmark dose methods and us epa’s benchmark dose software (bmds) version 2.1. 1. *Toxicology and applied pharmacology*, 254(2):181–191, 2011.
- [29] Beate Vieth, Swati Parekh, Christoph Ziegenhain, Wolfgang Enard, and Ines Hellmann. A systematic evaluation of single cell rna-seq analysis pipelines. *Nature communications*, 10(1):1–11, 2019.
- [30] Zhun Miao, Ke Deng, Xiaowo Wang, and Xuegong Zhang. Desingle for detecting three types of differential expression in single-cell rna-seq data. *Bioinformatics*, 34(18):3223–3224, 2018.
- [31] Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendzierski. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome biology*, 17(1):1–15, 2016.
- [32] Charlotte Sonesson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255–261, 2018.
- [33] Tian Mou, Wenjiang Deng, Fengyun Gu, Yudi Pawitan, and Trung Nghia Vu. Reproducibility of methods to detect differentially expressed genes from single-cell rna sequencing. *Frontiers in genetics*, 10:1331, 2020.
- [34] Maria K Jaakkola, Fatemeh Seyednasrollah, Arfa Mehmood, and Laura L Elo. Comparison of methods to detect differentially expressed genes between single-cell populations. *Briefings in bioinformatics*, 18(5):735–743, 2017.
- [35] Tae Kyun Kim. Understanding one-way anova using conceptual figures. *Korean journal of anesthesiology*, 70(1):22–26, 2017.
- [36] DA Williams. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, pages 103–117, 1971.
- [37] Jan De Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32:1–24, 2010.

- [38] Tim Holland-Letz and Annette Kopp-Schneider. Optimal experimental designs for dose–response studies with continuous endpoints. *Archives of toxicology*, 89(11):2059–2068, 2015.
- [39] Marc Aerts, Matthew W Wheeler, and José Cortiñas Abrahantes. An extended and unified modeling framework for benchmark dose estimation for both continuous and binary data. *Environmetrics*, 31(7):e2630, 2020.
- [40] Matthew W Wheeler, Jose Cortiñas Abrahantes, Marc Aerts, Jeffery S Gift, and Jerry Allen Davis. Continuous model averaging for benchmark dose analysis: Averaging over distributional forms. *Environmetrics*, page e2728, 2022.
- [41] Richard L Schmoyer. Sigmoidally constrained maximum likelihood estimation in quantal bioassay. *Journal of the American Statistical Association*, 79(386):448–453, 1984.
- [42] Colleen Kelly and John Rice. Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, pages 1071–1085, 1990.
- [43] Michel Delecroix, Michel Simioni, and Christine Thomas-Agnan. Functional estimation under shape constraints. *Journaltitle of Nonparametric Statistics*, 6(1):69–89, 1996.
- [44] Brian Neelon and David B Dunson. Bayesian isotonic regression and trend analysis. *Biometrics*, 60(2):398–406, 2004.
- [45] Björn Bornkamp and Katja Ickstadt. Bayesian nonparametric estimation of continuous monotone functions with applications to dose–response analysis. *Biometrics*, 65(1):198–205, 2009.
- [46] Lizhen Lin and David B Dunson. Bayesian monotone regression using gaussian process projection. *Biometrika*, 101(2):303–317, 2014.
- [47] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.
- [48] Lian En Chai, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria. A review on the computational approaches for gene regulatory network construction. *Computers in biology and medicine*, 48:55–65, 2014.
- [49] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13, 2008.
- [50] Seongho Kim. ppcor: an r package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods*, 22(6):665, 2015.
- [51] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe’er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

- [52] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):1–10, 2010.
- [53] Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 2019.
- [54] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. Springer, 2006.
- [55] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biol*, 5(1):e8, 2007.
- [56] Kevin Murphy, Saira Mian, et al. Modelling gene expression data using dynamic bayesian networks. Technical report, Citeseer, 1999.
- [57] Jiguo Cao, Xin Qi, and Hongyu Zhao. Modeling gene regulation networks using ordinary differential equations. In *Next generation microarray bioinformatics*, pages 185–197. Springer, 2012.
- [58] Pierre Geurts et al. dyngenie3: dynamical genie3 for the inference of gene networks from time series expression data. *Scientific reports*, 8(1):1–12, 2018.
- [59] Ziv Bar-Joseph, Georg K Gerber, Tong Ihn Lee, Nicola J Rinaldi, Jane Y Yoo, François Robert, D Benjamin Gordon, Ernest Fraenkel, Tommi S Jaakkola, Richard A Young, et al. Computational discovery of gene modules and regulatory networks. *Nature biotechnology*, 21(11):1337–1342, 2003.
- [60] Keren Bahar Halpern, Rom Shenhav, Orit Matcovitch-Natan, Beata Toth, Doron Lemze, Matan Golan, Efi E Massasa, Shaked Baydatch, Shanie Landen, Andreas E Moor, et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, 542(7641):352–356, 2017.
- [61] Tianhao Mu, Liqin Xu, Yu Zhong, Xinyu Liu, Zhikun Zhao, Chaoben Huang, Xiaofeng Lan, Chengchen Lufei, Yi Zhou, Yixun Su, et al. Embryonic liver developmental trajectory revealed by single-cell rna sequencing in the foxa2egfp mouse. *Communications biology*, 3(1):1–12, 2020.
- [62] Dongyin Guan, Ying Xiong, Trang Minh Trinh, Yang Xiao, Wenxiang Hu, Chunjie Jiang, Pieterjan Dierickx, Cholsoon Jang, Joshua D Rabinowitz, and Mitchell A Lazar. The hepatocyte clock and feeding control chronophysiology of multiple liver cell types. *Science*, 369(6509):1388–1394, 2020.

- [63] Xuelian Xiong, Henry Kuang, Sahar Ansari, Tongyu Liu, Jianke Gong, Shuai Wang, Xu-Yun Zhao, Yewei Ji, Chuan Li, Liang Guo, et al. Landscape of intercellular crosstalk in healthy and nash liver revealed by single-cell secretome gene analysis. *Molecular cell*, 75(3):644–660, 2019.
- [64] Reza Farmahin, Anne Marie Gannon, Rémi Gagné, Andrea Rowan-Carroll, Byron Kuo, Andrew Williams, Ivan Curran, and Carole L Yauk. Hepatic transcriptional dose-response analysis of male and female fischer rats exposed to hexabromocyclododecane. *Food and Chemical Toxicology*, 133:110262, 2019.
- [65] Ivy Moffat, Nikolai L Chepelev, Sarah Labib, Julie Bourdon-Lacombe, Byron Kuo, Julie K Buick, France Lemieux, Andrew Williams, Sabina Halappanavar, Amal I Malik, et al. Comparison of toxicogenomics and traditional approaches to inform mode of action and points of departure in human health risk assessment of benzo [a] pyrene in drinking water. *Critical reviews in toxicology*, 45(1):1–43, 2015.
- [66] A Francina Webster, Nikolai Chepelev, Rémi Gagné, Byron Kuo, Leslie Recio, Andrew Williams, and Carole L Yauk. Impact of genomics platform and statistical filtering on transcriptional benchmark doses (bmd) and multiple approaches for selection of chemical point of departure (pod). *PLoS One*, 10(8):e0136764, 2015.
- [67] Timothy W Gant and Shu-Dong Zhang. In pursuit of effective toxicogenomics. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 575(1-2):4–16, 2005.
- [68] Samarendra Das and Shesh N Rai. Swarnseq: An improved statistical approach for differential expression analysis of single-cell rna-seq data. *Genomics*, 113(3):1308–1324, 2021.
- [69] Minjeong Jeon and Paul De Boeck. Decision qualities of bayes factor and p value-based hypothesis testing. *Psychological Methods*, 22(2):340, 2017.
- [70] Yong Li, Xiao-Bin Liu, and Jun Yu. A bayesian chi-squared test for hypothesis testing. *Journal of Econometrics*, 189(1):54–69, 2015.
- [71] Kelly A Fader, Rance Nault, Mathew P Kirby, Gena Markous, Jason Matthews, and Timothy R Zacharewski. Convergence of hepcidin deficiency, systemic iron overloading, heme accumulation, and rev-erb α/β activation in aryl hydrocarbon receptor-elicited hepatotoxicity. *Toxicology and applied pharmacology*, 321:1–17, 2017.
- [72] Nathalie Percie du Sert, Viki Hurst, Amrita Ahluwalia, Sabina Alam, Marc T Avey, Monya Baker, William J Browne, Alejandra Clark, Innes C Cuthill, Ulrich Dirnagl, et al. The arrive guidelines 2.0: Updated guidelines for reporting animal research. *Journal of Cerebral Blood Flow & Metabolism*, 40(9):1769–1777, 2020.
- [73] Rance Nault, Kelly A Fader, Sudin Bhattacharya, and Tim R Zacharewski. Single-nuclei rna sequencing assessment of the hepatic effects of 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin. *Cellular and Molecular Gastroenterology and Hepatology*, 11(1):147–159, 2021.

- [74] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.
- [75] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):1–15, 2017.
- [76] Michael A Newton, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.
- [77] Yoav Benjamini and Yosef Hochberg. 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, 1995.
- [78] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.
- [79] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.
- [80] Ronald Aylmer Fisher et al. On the "probable error" of a coefficient of correlation deduced from a small sample.(1921). *Contributions to Mathematical Statistics*. 3–32, 1950.
- [81] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.
- [82] Rafael A Irizarry, Daniel Warren, Forrest Spencer, Irene F Kim, Shyam Biswal, Bryan C Frank, Edward Gabrielson, Joe GN Garcia, Joel Geoghegan, Gregory Germino, et al. Multiple-laboratory comparison of microarray platforms. *Nature methods*, 2(5):345–350, 2005.
- [83] Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: power analysis for bulk and single cell rna-seq experiments. *Bioinformatics*, 33(21):3486–3488, 2017.
- [84] Xiuwei Zhang, Chenling Xu, and Nir Yosef. Simulating multiple faceted variability in single cell rna sequencing. *Nature communications*, 10(1):1–16, 2019.
- [85] Alemu Takele Assefa, Jo Vandesompele, and Olivier Thas. Spsimseq: semi-parametric simulation of bulk and single-cell rna-sequencing data. *Bioinformatics*, 36(10):3276–3278, 2020.
- [86] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

- [87] Alessandra Dal Molin, Giacomo Baruzzo, and Barbara Di Camillo. Single-cell rna-sequencing: assessment of differential expression analysis methods. *Frontiers in genetics*, 8:62, 2017.
- [88] Jason R Phillips, Daniel L Svoboda, Arpit Tandon, Shyam Patel, Alex Sedykh, Deepak Mav, Byron Kuo, Carole L Yauk, Longlong Yang, Russell S Thomas, et al. Bmdexpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics*, 35(10):1780–1782, 2019.
- [89] Othman Soufan, Jessica Ewald, Charles Viau, Doug Crump, Markus Hecker, Niladri Basu, and Jianguo Xia. T1000: a reduced gene set prioritized for toxicogenomic studies. *PeerJ*, 7:e7975, 2019.
- [90] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.
- [91] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.
- [92] Kenneth Lange. *MM optimization algorithms*. SIAM, 2016.
- [93] D. R. Hunter and K. Lange. A tutorial on mm algorithms. *The American Statistician*, 58:30—37, 2004.
- [94] S Lang and A Brezger. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13:183–212, 2004.
- [95] Kenneth Lange. The mm algorithm. In *Optimization*, pages 185–219. Springer, 2013.
- [96] Florin Vaida. Parameter convergence for em and mm algorithms. *Statistica Sinica*, pages 831–840, 2005.
- [97] Jan de Leeuw and Kenneth Lange. Sharp quadratic majorization in one dimension. *Computational statistics & data analysis*, 53(7):2471–2484, 2009.
- [98] Kenneth Lange, Joong-Ho Won, Alfonso Landeros, and Hua Zhou. Nonconvex optimization via mm algorithms: Convergence theory. *arXiv preprint arXiv:2106.02805*, 2021.
- [99] Aaron. T. L. Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1):1–14, 2016.
- [100] Matt P Wand. A comparison of regression spline smoothing procedures. *Computational Statistics*, 15(4):443–462, 2000.
- [101] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Cambridge university press, 2003.

- [102] T Robertson, FT Wright, and R Dykstra. *Order Restricted Statistical Inference*. John Wiley&Sons, 1988.
- [103] Suzanne Winsberg and James O Ramsay. Monotonic transformations to additivity using splines. *Biometrika*, 67(3):669–674, 1980.
- [104] Larry Schumaker. *Spline functions: basic theory*. Cambridge University Press, 2007.
- [105] Simon Wood and Maintainer Simon Wood. Package ‘mgcv’. *R package version*, 1(29):729, 2015.
- [106] S Wotherspoon and Burch P. Package ‘zigam’. *R package Github version*, <https://github.com/AustralianAntarcticDataCentre/zigam>, 2016.
- [107] Naomi Moris, Cristina Pina, and Alfonso Martinez Arias. Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics*, 17(11):693–703, 2016.
- [108] Mark WEJ Fiers, Liesbeth Minnoye, Sara Aibar, Carmen Bravo González-Blas, Zeynep Kalender Atak, and Stein Aerts. Mapping gene regulatory networks from single-cell omics data. *Briefings in functional genomics*, 17(4):246–254, 2018.
- [109] Assieh Saadatpour, Guoji Guo, Stuart H Orkin, and Guo-Cheng Yuan. Characterizing heterogeneity in leukemic cells using single-cell gene expression analysis. *Genome biology*, 15(12):1–13, 2014.
- [110] Victoria Moignard, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, Iain C Macaulay, Wajid Jawaid, Evangelia Diamanti, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature biotechnology*, 33(3):269–276, 2015.
- [111] Shuonan Chen and Jessica C Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC bioinformatics*, 19(1):1–21, 2018.
- [112] Lucrezia Patruno, Davide Maspero, Francesco Craighero, Fabrizio Angaroni, Marco Antoniotti, and Alex Graudenzi. A review of computational strategies for denoising and imputation of single-cell transcriptomic data. *Briefings in Bioinformatics*, 22(4):bbaa222, 2021.
- [113] Kyle Akers and TM Murali. Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*, 26:87–97, 2021.
- [114] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):1–17, 2018.
- [115] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

- [116] Xiaowen Dong, Dorina Thanou, Michael Rabbat, and Pascal Frossard. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63, 2019.
- [117] Gonzalo Mateos, Santiago Segarra, Antonio G Marques, and Alejandro Ribeiro. Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Processing Magazine*, 36(3):16–43, 2019.
- [118] Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23):6160–6173, 2016.
- [119] Vassilis Kalofolias. How to learn a graph from smooth signals. In *Artificial Intelligence and Statistics*, pages 920–929, 2016.
- [120] Junhui Hou, Lap-Pui Chau, Ying He, and Huanqiang Zeng. Robust laplacian matrix learning for smooth graph signals. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1878–1882. IEEE, 2016.
- [121] Peter Berger, Gabor Hannak, and Gerald Matz. Efficient graph learning from noisy and incomplete data. *IEEE Transactions on Signal and Information Processing over Networks*, 6:105–119, 2020.
- [122] Sai Kiran Kadambari and Sundeep Prabhakar Chepuri. Learning product graphs from multidomain signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5665–5669. IEEE, 2020.
- [123] Liu Rui, Hossein Nejati, Seyed Hamid Safavi, and Ngai-Man Cheung. Simultaneous low-rank component and graph estimation for high-dimensional graph signals: Application to brain imaging. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4134–4138. IEEE, 2017.
- [124] Gerald Matz and Thomas Dittrich. Learning signed graphs from data. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5570–5574. IEEE, 2020.
- [125] Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W De Luca, and Sahin Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 559–570. SIAM, 2010.
- [126] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.
- [127] Aliaksei Sandryhaila and Jose MF Moura. Discrete signal processing on graphs: Frequency analysis. *IEEE Transactions on Signal Processing*, 62(12):3042–3054, 2014.

- [128] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.
- [129] John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.
- [130] Michael A Skinnider, Jordan W Squair, and Leonard J Foster. Evaluating measures of association for single-cell transcriptomics. *Nature methods*, 16(5):381–386, 2019.
- [131] Thomas P Quinn, Mark F Richardson, David Lovell, and Tamsyn M Crowley. propr: an r-package for identifying proportionally abundant features using compositional data analysis. *Scientific reports*, 7(1):1–9, 2017.
- [132] Ronald S Pimentel, Magdalena Niewiadomska-Bugaj, and Jung-Chao Wang. Association of zero-inflated continuous variables. *Statistics & Probability Letters*, 96:61–67, 2015.
- [133] Inbal Yahav and Galit Shmueli. On generating multivariate poisson data in management science applications. *Applied Stochastic Models in Business and Industry*, 28(1):91–102, 2012.
- [134] Thalia E Chan, Michael PH Stumpf, and Ann C Babbie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, 5(3):251–267, 2017.
- [135] Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.
- [136] Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of computational biology*, 16(2):229–239, 2009.
- [137] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.
- [138] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.
- [139] Zhi-Ping Liu, Canglin Wu, Hongyu Miao, and Hulin Wu. Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, 2015, 2015.
- [140] Luz Garcia-Alonso, Christian H Holland, Mahmoud M Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research*, 29(8):1363–1375, 2019.

- [141] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. Trustrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2018.
- [142] DA Brafman, C Phung, N Kumar, and K Willert. Regulation of endodermal differentiation of human embryonic stem cells through integrin-ecm interactions. *Cell Death & Differentiation*, 20(3):369–381, 2013.
- [143] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jee Choi, Christina Kendzierski, Ron Stewart, and James A Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17(1):1–20, 2016.
- [144] Alistair J Watt, Roong Zhao, Jixuan Li, and Stephen A Duncan. Development of the mammalian liver and ventral pancreas is dependent on gata4. *BMC developmental biology*, 7(1):1–11, 2007.
- [145] Emily M Walker, Cayla A Thompson, and Michele A Battle. Gata4 and gata6 regulate intestinal epithelial cytodifferentiation during development. *Developmental biology*, 392(2):283–294, 2014.
- [146] JB Fisher, K Pulakanti, S Rao, and SA Duncan. Gata6 is essential for endoderm formation from human pluripotent stem cells. *Biology open*, 6(7):1084–1095, 2017.
- [147] Qing Zhou, Hiram Chipperfield, Douglas A Melton, and Wing Hung Wong. A gene regulatory network in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences*, 104(42):16438–16443, 2007.
- [148] Wenjing Shi, Hui Wang, Guangjin Pan, Yijie Geng, Yunqian Guo, and Duanqing Pei. Regulation of the pluripotency marker rex-1 by nanog and sox2. *Journal of biological chemistry*, 281(33):23319–23325, 2006.
- [149] Kathy K Niakan, Hongkai Ji, René Maehr, Steven A Vokes, Kit T Rodolfa, Richard I Sherwood, Mariko Yamaki, John T Dimos, Alice E Chen, Douglas A Melton, et al. Sox17 promotes differentiation in mouse embryonic stem cells by directly regulating extraembryonic gene expression and indirectly antagonizing self-renewal. *Genes & development*, 24(3):312–326, 2010.
- [150] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. Upset: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics*, 20(12):1983–1992, 2014.
- [151] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640, 2014.
- [152] Raphael Petegrosso, Zhuliu Li, and Rui Kuang. Machine learning and statistical methods for clustering single-cell rna-sequencing data. *Briefings in bioinformatics*, 21(4):1209–1223, 2020.

- [153] Zhaoning Wang, Miao Cui, Akansha M Shah, Wei Tan, Ning Liu, Rhonda Bassel-Duby, and Eric N Olson. Cell-type-specific gene regulatory networks underlying murine neonatal heart regeneration at single-cell resolution. *Cell reports*, 33(10):108472, 2020.
- [154] Zhigang Xue, Kevin Huang, Chaochao Cai, Lingbo Cai, Chun-yan Jiang, Yun Feng, Zhenshan Liu, Qiao Zeng, Liming Cheng, Yi E Sun, et al. Genetic programs in human and mouse early embryos revealed by single-cell rna sequencing. *Nature*, 500(7464):593–597, 2013.
- [155] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.
- [156] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.
- [157] Holger Scheel and Stefan Scholtes. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, 2000.
- [158] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78(1):29–63, 2019.
- [159] Vassilis Kalofolias and Nathanaël Perraudin. Large scale graph learning from smooth signals. *arXiv preprint arXiv:1710.05654*, 2017.
- [160] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.
- [161] I-Cheng Ho, Tzong-Shyuan Tai, Sung-Yun Pai, et al. Gata3 and the t-cell lineage: essential functions before and after t-helper-2-cell differentiation. *Nature reviews immunology*, 9(2):125–135, 2009.
- [162] Pual Riley, Lynn Anaon-Cartwright, and James C Cross. The hand1 bhlh transcription factor is essential for placentation and cardiac morphogenesis. *Nature genetics*, 18(3):271–275, 1998.