Summer 2022

Statistical Methods for Analyzing Dependence Structures with Applications in Single-cell Experiments

Zhen Yang

Statistical Methods for Analyzing Dependence Structures with Applications in Single-cell Experiments

by

Zhen Yang

Bachelor of Science
Nanjing University 2014

Master of Science
The George Washington University 2016

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2022

Accepted by:

Yen-Yi Ho, Major Professor

Ray Bai, Committee Member

Hexin Chen, Committee Member

Lianming Wang, Committee Member

Tracey L. Weldon, Vice Provost and Dean of the Graduate School

Acknowledgments

I would first like to thank my advisor Dr. Yen-Yi Ho for patiently guiding and supporting me over the years. I also thank Dr. Ho for introducing me to the fields of dynamic correlations and genomics data analysis. I would also like to thank Dr. Ray Bai, Dr. Hexin Chen, and Dr. Lianming Wang for being on my dissertation committee and offering valuable suggestions.

I would like to thank the Department of Statistics at the University of South Carolina for providing me with the opportunity to pursue a doctoral degree. I would like to thank every professor who taught me and every staff member who helped me. I would also like to express my sincere gratitude to Zichen. He helped me in every stage of my Ph.D. journey. Without his selfless support, my Ph.D. journey would be much harder.

Finally and most importantly, I would like to express my deepest gratitude to my parents for always supporting me and respecting my decisions. I want to thank my wife, Nan, for her love and support. To Keli: Your birth is the best gift for your mom and me.

Abstract

This dissertation focuses on studying methods in dependence structure analysis. In particular, it consists of two topics: (1) modeling dynamic correlation in zero-inflated bivariate count data; and (2) gene co-expression latent factor analysis for cell-type clustering.

In Chapter 2, a zero-inflated negative binomial model for analyzing the dynamic correlation in zero-inflated bivariate count data is proposed. Interactions between biological molecules in a cell are tightly coordinated and often highly dynamic. As a result of these varying signaling activities, changes in gene co-expression patterns could often be observed. The advancements in next-generation sequencing technologies bring new statistical challenges for studying these dynamic changes of gene co-expression. In recent years, methods have been developed to examine genomic information from individual cells. Single-cell RNA sequencing (scRNA-seq) data are count-based, and often exhibit characteristics such as over-dispersion and zero-inflation. To explore the dynamic dependence structure in scRNA-seq data and other zero-inflated count data, new approaches are needed. We consider over-dispersion and zero-inflation in count outcomes and propose a ZEro-inflated Negative binomial dynamic COrrelation model (ZENCO). The observed count data are modeled as a mixture of two components: success amplifications and dropout events in ZENCO. A latent variable is incorporated into ZENCO in order to model the covariate-dependent correlation structure. We conduct simulation studies to evaluate the performance of our proposed method and to compare it with existing approaches. We also illustrate the implementation of our proposed approach using scRNA-seq data in melanoma.

Chapter 3 proposes a cell-type clustering approach that allows for joint analysis of both expression structures and co-expression structures of the data. Due to the complex regulatory mechanisms, biological molecules in a cell often participate in complicated interaction processes. The traditional cell-type clustering approaches use average expression levels of the data and are therefore insufficient for understanding the intricate regulatory mechanisms that underlie different cellular conditions. The co-expression structures can often bring insights into the complex genetic interactions and can help detect correlation changes between pairs of genes across different modulating conditions. Therefore, the co-expression structures can help identify hidden sub-groups in the data and improve the performance of clustering. Our method learns the joint features shared among expression structures and co-expression structures of the data and identifies the unique variation present in each type of structure to further cluster the cell types. The proposed approach is applied to a breast cancer cells data set.

In Chapter 4, a subject-specific random effects model for zero-inflated count-based data is proposed. Tumor heterogeneity is very common and plays important role in therapy design. The development of scRNA-seq technologies brings new opportunities along with challenges for studying tumor heterogeneity. For scRNA-seq data, one of the main analytical approaches is differential co-expression analysis, which can reveal the intricate underlying gene regulatory mechanisms in tumor cells. In recent years, methods have been developed for modeling the dynamic changes of gene co-expression in scRNA-seq data. However, due to the heterogeneous nature of tumors, new approaches are needed. In this chapter, we propose a subject-specific random effects model for zero-inflated count-based data such as scRNA-seq data. A latent variable is incorporated into the model to quantify the correlation dependency structure. We conduct simulation studies to evaluate the performance of our proposed method and to compare it with existing approaches. We also illustrate the implementation of our

proposed approach using scRNA-seq data from a study of immunotherapy resistance in melanoma tumors.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Analysis for dependence structures or dynamic correlation patterns between random quantities is one of the most crucial topics in both statistical theory and genomic applications. In this section, we briefly review the approaches and challenges in the study of dynamic correlation.

Interactions between biological molecules in a cell are tightly coordinated and often highly dynamic (Luscombe et al., 2004; de Lichtenberg et al., 2005). They can change flexibly under different cellular conditions or in response to various external stimulants and signals. As a result of these varying signaling activities, changes in gene co-expression patterns can often be observed in these situations (Li, 2002; Li and Yuan, 2004; de la Fuente, 2010a). Studying these dynamic changes in gene co-expression could reveal these intricate underlying gene regulatory mechanisms.

While it is a challenging task to unravel the complex genetic interactions in a biological system, several statistical approaches have been introduced to describe the co-expression between a pair of genes such as Pearson correlation or rank correlation, F-statistic (Lai et al., 2004), mutual information (Faith et al., 2007), entropy-based approaches (Ho et al., 2007), Gaussian graphical models (Ma et al., 2007), and Bayesian network (Ho et al., 2014). However, these approaches do not account for the fact that genetic circuits can be turned on or off and genes may participate in different regulatory processes under different cellular conditions.

One statistical measure that can capture these dynamic gene correlation changes was proposed by Li (2002). This measure, named dynamic correlation in this paper, quantifies the relationship where the co-expression between two genes is modulated by a third "coordinator" gene. Li (2002) examined these dynamic correlation changes (referred to as liquid association in his paper) in canonical pathways using microarray gene expression data from a model organism, *Saccharomyces cerevisiae*. For a typical genomic study, a pathway-based or a genome-wide screening strategy can be implemented as presented in several studies to effectively identify potential dynamic correlation changes (Dawson and Kendziorski, 2012; Gunderson and Ho, 2014; Wang et al., 2017; Yu, 2018; Kinzy et al., 2019). Li's study and other studies since then, have evidently established its biological validity and popularized it to be a useful tool for analyzing genomic data (Li, 2002; Li et al., 2004; Ho et al., 2007; Zhang et al., 2007; Ho et al., 2011; Wang et al., 2013; Xu et al., 2017; Wang et al., 2017; Khayer et al., 2017; Ai et al., 2019; Kong and Yu, 2019; Wen et al., 2020).

However, when it comes to count data such as RNA sequencing reads, these existing Gaussian-based approaches may not fit the data properly. RNA sequencing (RNA-seq) data are often presented as a count matrix with non-negative counts as the number of reads observed. Count-based models such as the Poisson distribution and the negative binomial distribution are widely used to analyze the RNA-seq data. Karlis and Meligkotsidou (2005) proposed a multivariate Poisson model with covariance structure. Due to both biological and technical variability, RNA-seq count data are often over-dispersed. For over-dispersed data, the variance is larger than the mean, which is a violation of the assumption of the Poisson distribution (mean and variance are equal). To handle over-dispersion, Solis-Trapala and Farewell (2005) used a multivariate Poisson-Gamma mixture model. Robinson et al. (2010) modeled the data using the negative binomial distribution and treated the Poisson distribution as a special case of the negative binomial distribution. Ma et al. (2020) proposed

three flexible Bayesian regression approaches for modeling bivariate correlated count data.

In recent years, the rapid development of next-generation sequencing technologies has made it possible to examine the sequence information from individual cells. Single-cell RNA sequencing (scRNA-seq) analyzes the expression of RNAs from individual cells, while traditional RNA-seq can only analyze the RNAs from mixed cell populations (Bacher and Kendziorski, 2016; Hwang et al., 2018). scRNA-seq gives insight into individual cells' function and behavior at various stages and in various cell types and hence can provide a high-resolution view of dynamic co-expression regulation in a biological system.

However, the analysis of scRNA-seq data is complicated by high levels of technical noise and intrinsic biological variability (Kharchenko et al., 2014). Due to the low amounts of mRNA within individual cells, the counts of single-cell gene expression data contain a large number of zero expression measurements. To avoid stochastic zero counts, Lun et al. (2016) developed a normalization method based on pooling expression values. Pierson and Yau (2015) developed a dimensionality-reduction method considering the dropout characteristics to improve modeling accuracy. Miao et al. (2018) used a zero-inflated negative binomial model to estimate the proportion of real and dropout zeros. Kharchenko et al. (2014) modeled the measurement of each cell as a mixture of two components: one for transcripts that are successfully detected and the other for dropout events during amplification.

1.2 Background Concept

In this section we review the concept of liquid association which will appear in later chapters.

Liquid association is a statistical measure that quantifies the relationship between the correlation of two genes (X_i, X_j) and the expression level of third gene (Z).

Suppose that three random variables X_i, X_j, and Z represent the expression levels of three genes and are already standardized to have mean 0 and variance 1. We can regard the entry-wise product of X_i and X_j as the co-expression measure between them. So the correlation coefficient between X_i and X_j is $E(X_iX_j)$. By conditioning on Z,

$$E(X_iX_j) = E(E(X_iX_j|Z)) = Eg(Z), \tag{1.1}$$

where $g(Z) = E(X_iX_j|Z)$ is the conditional correlation of X_i and X_j given Z. Then the liquid association of X_i and X_j with respect to Z is defined as

$$LA(X_i, X_j|Z) = Eg'(Z), \tag{1.2}$$

where $g'(X)$ is the derivative of $g(X)$ with respective to X. Furthermore, when Z follows a standard normal distribution, $LA(X_i, X_j|Z) = Eg'(Z)$ is equal to a three-product-moment estimator $E(X_iX_jZ)$.

In the above setting of liquid association, the variable Z is given. Oftentimes, Z represents the pathway activities and cannot be measured directly. To estimate the latent variables Z, Yu (2018) uses a new vector $\mathbf{h}$ which is the element-wise product of X_i and X_j. For any gene pair, a product vector $\mathbf{h}$ can be calculated. We then construct a product matrix X_{prod}, each row of which being a product vector $\mathbf{h}$ of a gene pair to represent the co-expression structures of the data. For a product matrix with m rows, Yu (2018) shows that, by applying eigenvalue decomposition to the matrix $X'_{prod}X_{prod}$, the latent factors $Z_k, k = 1, ..., K$ that regulate the coexpression structures of the data can be obtained from

$$\mathbf{z}_k = argmax_{||\mathbf{z}||=1} \sum_m (\mathbf{z} \cdot \mathbf{h})^2, \mathbf{z_s}'\mathbf{z_t} = 0, s \neq t. \tag{1.3}$$

1.3 Structure of the dissertation proposal

The rest of the dissertation proposal is organized as follows. In Chapter 2, we propose a zero-inflated negative binomial model for studying dynamic correlation in

zero-inflated, over-dispersed count data, such as scRNA-seq data. In Chapter 3, we propose a cell-type clustering approach that allows for joint analysis of both expression structures and co-expression structures of the data. In Chapter 4, we propose a subject-specific random effects model which incorporates between-individual random effects to account for the variation of cells collected from different participants.

Chapter 2

Modeling Dynamic Correlation in Zero-inflated Bivariate Count Data

2.1 Introduction

Motivated by the dynamic correlation studies in microarray data, in this chapter, we propose the ZEro-inflated Negative binomial dynamic COrrelation (ZENCO) model. We account for over-dispersion and zero-inflation in count data by considering a mixture model of conditional bivariate negative binomial regressions and zero counts. A latent variable is incorporated into ZENCO to model the covariate-dependent correlation structure. We demonstrate the implementation of ZENCO model using the scRNA-seq data of melanoma cells from Gene Expression Omnibus (GSE116237) and study the difference of dynamic correlations between various phases during treatment with BRAF/MEK inhibitors.

The remainder of the chapter is arranged as follows. In Section 2.2, the detail of the proposed model is introduced. The simulation studies and comparisons are conducted in Section 2.3. In Section 2.4, the analysis of scRNA-seq data generated from melanoma tumor cells is presented. Section 2.5 concludes this chapter with some discussion.

2.2 Method

2.2.1 The ZENCO model

For modeling dynamic co-expression changes, we use $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ to denote the count-based expression levels for three genes. Let X_{ij} represent the gene expression level of the ith gene ($i = 1, 2, 3$) in the jth cells ($j = 1, 2, ...n$), and $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, ..., X_{in})$ represents the gene expression level for the ith gene. In our proposed framework, the marginal distribution of $\mathbf{X}_i$ is modeled as a mixture of dropout component and negative binomial component (non-dropout events). The distribution of $\mathbf{X}_i$ is given by

$$\mathbf{X}_i \sim \begin{cases} \mathcal{I}_0, & \text{with probability } p_i; \\ NB(\mu_i, \phi_i), & \text{with probability } 1 - p_i. \end{cases} \tag{2.1}$$

where $\mathcal{I}_0$ is the distribution with a point mass at zero; p_i is the dropout rate of $\mathbf{X}_i$; μ_i is the mean of the negative binomial component of $\mathbf{X}_i$ and ϕ_i is the dispersion parameter of the negative binomial component. The variance of the negative binomial component of $\mathbf{X}_i$ is $\mu_i(1 + \phi_i\mu_i)$. As ϕ_i goes to 0, $NB(\mu_i, \phi_i) \to Poisson(\mu_i)$.

The dropout rate of a given gene, p_i, is modeled as a function of its mean. The dropout rates are study-specific and can be estimated for a given scRNA-seq dataset. Based on the melanoma data considered in the study, we model the dropout rate using a logistic function: $p = \frac{e^{(b_0+b_1\mu)}}{1+e^{(b_0+b_1\mu)}}$, where μ is the mean of a given gene and b_0, b_1 can be estimated using the expression levels of all available genes in the data (Pierson and Yau, 2015).

Furthermore, we use the indicator $d_{ij} \sim Bernoulli(p_i)$ to describe whether dropout happens or not. If $d_{ij} = 0$, then the ith gene in the jth cell is successfully amplified (non-dropout event). If $d_{ij} = 1$, then dropout happens. According to the combinations of different values of d_{1j} and d_{2j}, there are four different situations for $\mathbf{X}_1$ and

$\mathbf{X}_2$. Their marginal densities can be written as:

$$\begin{cases} X_{1j} \sim NB(\mu_1, \phi_1) \text{ and } X_{2j} \sim NB(\mu_2, \phi_2), & \text{if } d_{1j} = d_{2j} = 0; \\ X_{1j} \sim \mathcal{I}_0 \text{ and } X_{2j} \sim NB(\mu_2, \phi_2), & \text{if } d_{1j} = 1 \text{ and } d_{2j} = 0; \\ X_{1j} \sim NB(\mu_1, \phi_1) \text{ and } X_{2j} \sim \mathcal{I}_0, & \text{if } d_{1j} = 0 \text{ and } d_{2j} = 1; \\ X_{1j} \sim \mathcal{I}_0 \text{ and } X_{2j} \sim \mathcal{I}_0, & \text{if } d_{1j} = d_{2j} = 1. \end{cases} \tag{2.2}$$

When $d_{1j} = d_{2j} = d_{3j} = 0$, the joint distribution of $\mathbf{X}_1$ and $\mathbf{X}_2$ involves a correlation parameter that depends on the expression level of X_{3j}. In other words, the correlation between X_{1j} and X_{2j} could change according to the level of X_{3j} when all three genes (X_{1j}, X_{2j} and X_{3j}) are successfully amplified in the jth cell. If d_{1j}=1 or $d_{2j} = 1$, X_{1j} and X_{2j} are independent, because at least one measurement of X_{1j} and X_{2j} comes from the dropout component.

We model the dependency between $\mathbf{X}_1$ and $\mathbf{X}_2$ and construct our conditional bivariate negative binomial model through a Poisson-Gamma mixture distribution. For $i = 1, 2$ and $j = 1, 2, ..., n$, let

$$X_{ij} \sim Poisson(u_{ij}\mu_i), u_{ij} \sim Gamma(\alpha_i, \alpha_i). \tag{2.3}$$

A negative binomial distribution of $NB(\mu_i, \frac{1}{\alpha_i})$ can be generated by integrating over u_{ij} in (2.3). In this Poisson-Gamma mixture setting, u_{ij} can be considered as the cell-specific random effect. To introduce the conditional correlation between X_{1j} and X_{2j} given X_{3j}, we utilize a latent variable Z and model the conditional correlation implicitly through the cell-specific random effect (u_{ij}).

Let $\mathbf{Z_j} = (Z_{1j}, Z_{2j})'$ be a bivariate normal variable that

$$\mathbf{Z_j} \sim N_2\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_j \\ \rho_j & 1 \end{bmatrix} \Big). \tag{2.4}$$

The correlation, ρ_j, of (Z_{1j}, Z_{2j}) is specified as

$$\log(\frac{1+\rho_j}{1-\rho_j}) = \tau_0 + \tau_1 X_{3j}. \tag{2.5}$$

$\log(\frac{1+\rho_j}{1-\rho_j})$ is the Fisher's Z-transformation for the correlation ρ_j which ensures the correlation ρ_j is within (-1, 1).

Now, we incorporate this latent variable $\mathbf{Z_j}$ into the cell-specific random component (u_{ij}) in the Poisson-Gamma mixture in (2.3) to construct a conditional bivariate negative binomial model of $(X_{1j}, X_{2j})'$ with marginal distribution $X_{1j} \sim NB(\mu_1, \phi_1)$ and $X_{2j} \sim NB(\mu_2, \phi_2)$ and the correlation of (X_{1j}, X_{2j}) depends on X_{3j}. Specifically, for $i = 1, 2$ and $j = 1, 2, ..., n$, let

$$X_{ij} \sim Poisson[F_{\alpha_i}^{-1}\{\Phi(Z_{ij})\}\mu_i], \tag{2.6}$$

where $F_{\alpha_i}(\cdot)$ is the cumulative distribution function of a $Gamma(\alpha_i, \alpha_i)$ distribution with $\alpha_i = 1/\phi_i$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. $F_{\alpha_i}^{-1}$ maps each point in the interval (0,1) to $Gamma(\alpha_i, \alpha_i)$ distribution. Hence, the distribution of $F_{\alpha_i}^{-1}\{\Phi(Z_{ij})\}$ is $Gamma(\alpha_i, \alpha_i)$. The distribution of $X_{ij} \sim Poisson[F_{\alpha_i}^{-1}\{\Phi(Z_{ij})\}\mu_i]$ is then a Poisson-Gamma mixture distribution, which follows the negative binomial density $NB(\mu_i, \phi_i = \frac{1}{\alpha_i})$.

In the model described above, in order to determine the existence of the dynamic co-expression change of $\mathbf{X}_1$, $\mathbf{X}_2$ given $\mathbf{X}_3$, the main parameter of interest is τ_1 in (2.5). If τ_1=0, then the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ does not depend on $\mathbf{X}_3$ and vice versa. In the ZENCO model, we develop a statistical inference procedure via a Bayesian perspective, because it offers a relatively straightforward way to compute $Poisson[F_{\alpha_i}^{-1}\{\Phi(Z_{ij})\}]$ through Markov Chain Monte Carlo (MCMC) sampling. In addition, the posterior distributions of the parameters can be obtained with a set of standard conjugate priors.

Under the hypotheses:

$$H_0 : \tau_1 = 0 \text{ versus } H_1 : \tau_1 \neq 0,$$

the statistical power of the proposed ZENCO approach can be calculated as follows. First, we obtained the posterior sampling distribution of τ_1, then calculated the 95%

equal tail credible interval. Power can be evaluated as the proportion of times when zero is not covered by the 95% credible intervals.

We now describe the likelihood function and the MCMC scheme. Let vector $\boldsymbol{\theta}$ be the notation of all parameters (μ_1, μ_2, μ_3, ϕ_1, ϕ_2, ϕ_3, τ_0, τ_1) in the model. And let $\boldsymbol{\pi}(\boldsymbol{\theta})$ be the prior joint distribution of $\boldsymbol{\theta}$, the likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}|x_1, x_2, x_3) &= \prod_{j=1}^{n} f(x_{1j}, x_{2j}|\mu_1, \mu_2, \phi_1, \phi_2, \tau_0, \tau_1, x_{3j}) f(x_{3j}|\mu_3, \phi_3) \\ &= \prod_{j=1}^{n} \Big\{ \int f(x_{1j}, x_{2j}|\mu_1, \mu_2, \phi_1, \phi_2, \boldsymbol{z}_j) f(\boldsymbol{z}_j|x_{3j}, \tau_0, \tau_1) d\boldsymbol{z}_j \Big\} f(x_{3j}|\mu_3, \phi_3) \\ &= \prod_{j=1}^{n} \Big\{ \int \prod_{i=1}^{2} f(x_{ij}|\mu_i, \phi_i, \boldsymbol{z}_{ij}) f(\boldsymbol{z}_j|x_{3j}, \tau_0, \tau_1) d\boldsymbol{z}_j \Big\} f(x_{3j}|\mu_3, \phi_3), \end{aligned} \tag{2.7}$$

where x_{1j} and x_{2j} are from observed data and $\boldsymbol{z}_j = (z_{1j}, z_{2j})'$. x_{1j} and x_{2j} are independent given $\boldsymbol{z}_j$. Hence, the posterior joint distribution of μ_1, μ_2, μ_3, ϕ_1, ϕ_2, ϕ_3, τ_0, τ_1 given the observations is proportional to

$$\Big[\prod_{j=1}^{n} \Big\{ \int \prod_{i=1}^{2} f(x_{ij}|\mu_i, \phi_i, \boldsymbol{z}_{ij}) f(\boldsymbol{z}_j|x_{3j}, \tau_0, \tau_1) d\boldsymbol{z}_j \Big\} f(x_{3j}|\mu_3, \phi_3) \Big] \boldsymbol{\pi}(\boldsymbol{\theta}),$$

where $f(x_{ij}|\mu_i, \phi_i, \boldsymbol{z}_{ij})$ is the distribution of x_{ij} for $i = 1, 2$:

$$x_{ij} \sim \begin{cases} \mathcal{I}_0, & \text{with probability } p_i; \\ Poisson[F^{-1}_{1/\phi_i}\{\Phi(z_{ij})\}\mu_i], & \text{with probability } 1 - p_i. \end{cases}$$

The dropout rate p_i is study-specific and can be determined using all genes measured in the study as a function of μ_i described previously. And $f(\boldsymbol{z}_j|x_{3j}, \tau_0, \tau_1)$ is the probability density function of a bivariate normal distribution with a covariance matrix structure:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \frac{e^{(\tau_0+\tau_1\times x_{3j})}-1}{e^{(\tau_0+\tau_1\times x_{3j})}+1} \\ \frac{e^{(\tau_0+\tau_1\times x_{3j})}-1}{e^{(\tau_0+\tau_1\times x_{3j})}+1} & 1 \end{bmatrix}.$$

For any given x_{3j}, $\boldsymbol{z}_j$ can be derived as described in (2.4) and (2.5). Finally, $f(x_{3j}|\mu_3, \phi_3)$ is formulated as in (2.1).

For a given gene triplet, the parameter estimation can be carried out using the MCMC algorithm provided in JAGS (Plummer, 2003). We use the normal distribution with mean 0 and variance 4/N as the priors of τ_0 and τ_1, where N is the sample size. This is because the approximate variance of Fisher's Z-transformation $\log(\frac{1+\rho}{1-\rho})$ is $\frac{4}{N-3}$. The priors for μ_1, μ_2, and μ_3 are standard log-normal distributions. The non-informative priors for the dispersion parameters $1/\phi_1$, $1/\phi_2$, and $1/\phi_3$ are the Gamma distribution with mean 100 and relatively large variance 10,000.

The sampling scheme during each MCMC iteration is as follows. For $j = 1, 2, ..., n$, $i = 1, 2, 3$, we sample μ_i from $f(\mu_i|\cdot) \propto f(\mu_i) \prod_{j=1}^{n} f(x_{ij}|\mu_i, \phi_i)$ and sample ϕ_i from $f(1/\phi_i|\cdot) \propto f(1/\phi_i) \prod_{j=1}^{n} f(x_{ij}|\mu_i, \phi_i)$, where $f(x_{ij}|\mu_i, \phi_i)$ is the probability density function of

$$x_{ij} \sim \begin{cases} \mathcal{I}_0, & \text{with probability } p_i; \\ NB(\mu_i, \phi_i), & \text{with probability } 1 - p_i. \end{cases}$$

Then we sample τ_0 from

$$f(\tau_0|\cdot) \propto f(\tau_0) \prod_{j=1}^{n} f(\mathbf{z_j}|\tau_0, \tau_1, x_{3j}),$$

and sample τ_1 from

$$f(\tau_1|\cdot) \propto f(\tau_1) \prod_{j=1}^{n} f(\mathbf{z_j}|\tau_0, \tau_1, x_{3j}),$$

where $f(\mathbf{z_j}|\tau_0, \tau_1, x_{3j}) = N_2\Bigg(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \frac{e^{(\tau_0+\tau_1\times x_{3j})}-1}{e^{(\tau_0+\tau_1\times x_{3j})}+1} \\ \frac{e^{(\tau_0+\tau_1\times x_{3j})}-1}{e^{(\tau_0+\tau_1\times x_{3j})}+1} & 1 \end{bmatrix} \Bigg)$.
In addition, z_{ij} can be sampled from

$$f(z_{ij}|\cdot) \propto f(x_{ij}|z_{ij}, \mu_i, \alpha_i) f(z_{ij}|z_{kj}),\ i, k = 1, 2;\ i \neq k,$$

where $f(z_{ij}|z_{kj}) = N(\rho_j z_{kj}, (1 - {\rho_j}^2))$.

2.2.2 Search strategies

There are several ways to implement the ZENCO approach in a genomic study. We describe a few here: (i) for a given pair of genes ($\mathbf{X}_1$, $\mathbf{X}_2$), screen the whole-genome to

identify the coordinator genes ($\mathbf{X}_3$) that regulate the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$, or (ii) for a given $\mathbf{X}_3$, screen-related pathways or the whole-genome to identify pairs of genes that are modulated by $\mathbf{X}_3$ (m choose 2 gene pairs; m is the total number of genes considered), or (iii) if no prior information about $\mathbf{X}_3$ or ($\mathbf{X}_1$, $\mathbf{X}_2$) is available, screen relevant genetic pathways, or screen the whole-genome to identify potential gene triplets that exhibit dynamic correlation changes (m choose 3 gene triplets). In the experimental data analysis described in Section 4, we demonstrated the second (ii) approach.

When the number of relevant genes under consideration is large (for example $\approx$ 20,000), a pre-screening step is usually beneficial before implementing ZENCO. For example, the algorithm proposed by Gunderson and Ho (2014) or the screening statistic (ζ) introduced in Yu (2018) or filtering out gene with constant expression has been used effectively in the literature.

2.3 Simulation

To evaluate the performance of our proposed ZENCO model and compare it to existing benchmark approaches, we report results from five simulation scenarios below.

2.3.1 Scenario 1: Simulating data from ZENCO

In this first simulation, we demonstrate generating data from the ZENCO model. The simulated data contain count-based expression level of three genes: $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$. In our model, the correlations of $\mathbf{X}_1$ and $\mathbf{X}_2$ are modulated by the level of $\mathbf{X}_3$. This simulation was conducted as follows.

First, we simulated a set of $\{x_{3j}\}_{j=1}^{N}$ from a univariate negative binomial distribution with mean μ_3 and size ϕ_3 and then randomly selected a subset as the dropouts and replaced these $\{x_{3j}\}'s$ with zero. After the simulation of x_{3j}, we calculated correlation coefficient $\rho_j = \frac{e^{(\tau_0+\tau_1\times x_{3j})}-1}{e^{(\tau_0+\tau_1\times x_{3j})}+1}$ for each x_{3j}. Note that for dropouts in $\{x_{3j}\}_{j=1}^{N}$,

we used μ_3 instead of x_{3j} to calculate ρ_j, since the values of those dropouts have nothing to do with the regulatory mechanism of $\mathbf{X}_3$. Then, we generated latent variables $\mathbf{z}_j = (z_{1j}, z_{2j})'$ such that $\mathbf{z}_j \sim N_2\Big(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho_j\\ \rho_j & 1\end{bmatrix}\Big)$ and simulated x_{1j} and x_{2j} using $\mathbf{z}_j$ as described in (2.6). The dependence structure of x_{1j} and x_{2j} is implicitly modeled via $\mathbf{z}_j$. Finally, just like the simulation of x_{3j}, we randomly replaced values of x_{1j} and x_{2j} for dropout events.

Using the simulation approach described above, we generated 10^5 observations from the ZENCO distribution and plotted a panel of conditional distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$ given various levels of $\mathbf{X}_3$ in Figure 2.1. In these figures, we observed that when $\mathbf{X}_3$ is not zero, ρ increases with $\mathbf{X}_3$. When $\mathbf{X}_3$ is zero, the correlations of $\mathbf{X}_1$ and $\mathbf{X}_2$ are small and show reduced dependency with respect to $\mathbf{X}_3$. This is due to the zero value observation of $\mathbf{X}_3$ being a mixture of true zero and dropout. In other words, some zero values of $\mathbf{X}_3$ come from the negative binomial distribution, others come from dropout events.

2.3.2 Scenario 2: Comparisons to existing approaches

To evaluate the performance of our proposed ZENCO model, we performed power analysis and compare ZENCO to three other existing approaches. For testing the existence of dynamic co-expression changes, our hypotheses are set up as:

$$H_0 : \tau_1 = 0 \text{ versus } H_1 : \tau_1 \neq 0.$$

First, we compared ZENCO to a bivariate negative binomial regression without considering the zero-inflated components. Similarly to ZENCO, the statistical power of this method can be calculated as the percentage of times that the posterior 95% credible intervals of τ_1 do not cover zero. The ZENCO model and the model without considering the zero-inflated components were both carried out using the MCMC al-

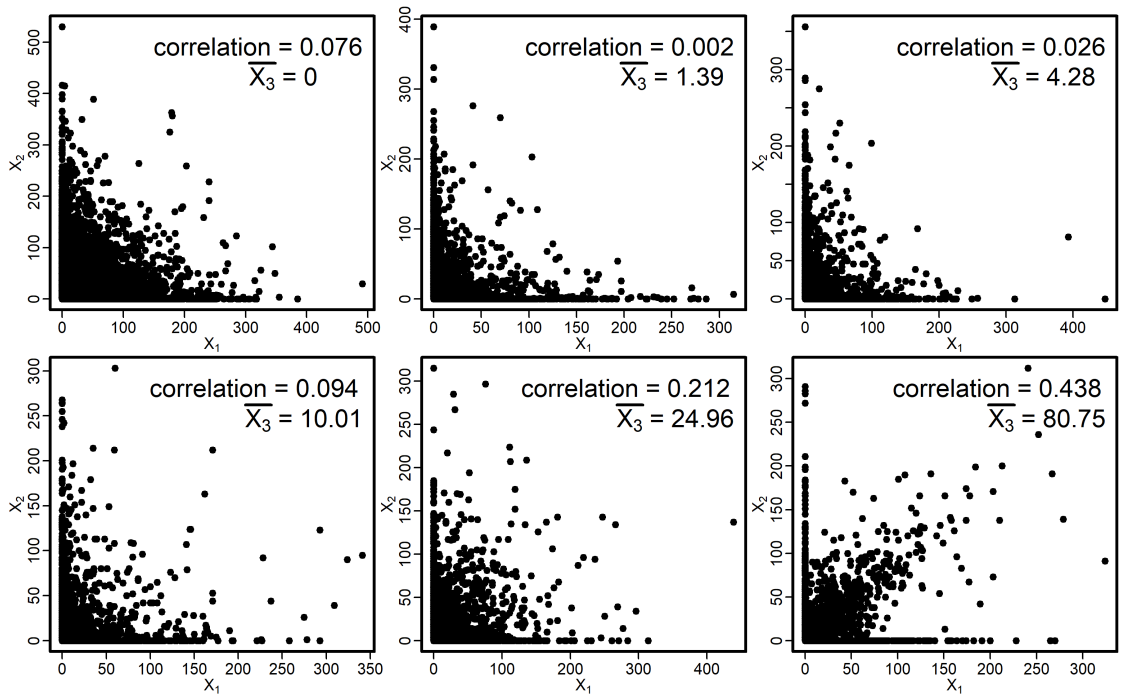

Figure 2.1. Profile plots of $(\mathbf{X}_1, \mathbf{X}_2|\mathbf{X}_3)$ with varying $\mathbf{X}_3$ ($\mu_1 = \mu_2 = \mu_3 = 15$, $\phi_1 = \phi_2 = \phi_3 = 4$, $\tau_0 = 0$, and $\tau_1 = 0.05$)

gorithm with 20,000 iterations, and 10,000 burn-ins.

Second, we compared ZENCO to the existing benchmark approach introduced by Li (2002). This existing approach was later applied to scRNA-seq data by Yu (2018). This test statistic according to the three-product-moment measure is written as: $T_{LA} = \frac{\widehat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)}{SE\{\widehat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)\}}$, where $\mathbf{X}_1^*$, $\mathbf{X}_2^*$, $\mathbf{X}_3^*$ are the standardized $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ with mean 0, variance 1, and $\widehat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)$ is the three-product-moment estimator for the dynamic correlation. $SE\{\widehat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)\}$, the standard error of $\widehat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)$, can be estimated via bootstrap. T_{LA} can be used to test whether the correlation of $\mathbf{X}_1$, $\mathbf{X}_2$ depends on $\mathbf{X}_3$, that is $H_0 : \tau_1 = 0$ (Li, 2002; Ho et al., 2011). The distribution of T_{LA} under the null hypothesis and associated p-value can be obtained using a permutation approach.

The third comparison is to fit the negative binomial count data with the conditional normal model (CNM-Full) (Ho et al., 2011). Assuming data are from the

conditional bivariate normal distribution instead of the conditional bivariate negative binomial distribution, the test statistic of this method can be estimated using a generalized estimating equation-based procedure (Yan and Fine, 2004) and a p-value associated with the test statistic can be obtained. The powers of these two methods (T_{LA} and CNM-Full) can be calculated by counting the percentage of times when p-values associated with τ_1 are less than 0.05.

We simulated 1,000 observations from ZENCO model by fixing $\mu_1 = \mu_2 = \mu_3 = 15$, $\phi_1 = \phi_2 = \phi_3 = 4$ and $\tau_0 = 0$, and then varied τ_1 values and performed power analyses. The simulated values of $\mu_1, \mu_2, \mu_3, \phi_1, \phi_2, \phi_3$ are based on the estimates obtained from the real data analysis. Figure 2.2 shows the power curves of the four methods. We observed that our proposed ZENCO method outperforms the other three methods. In addition, fitting the negative binomial count-based data using Gaussian-based models reduces statistical power drastically. This is because ZENCO accounts for both zero-inflation and overdispersion of the data, and hence achieves better power to detect dynamic dependence structure.

2.3.3 Scenario 3: Estimation efficiency

In this simulation scenario, we evaluated the estimation efficiency of the ZENCO model and reported mean squared errors (MSE), mean bias errors (MBE), and 95% empirical coverage probabilities under various settings. Three sets of simulation studies were done with sample sizes 200, 500, and 1,000. For each simulation study, we generated 1,000 datasets. We used the parameter estimated values obtained from the real data analysis in Section 4 and set the true values of the parameters as follows: $\mu_1 = \mu_2 = \mu_3 = 15$, $\phi_1 = \phi_2 = \phi_3 = 4$, $\tau_0 = 0.01$, and $\tau_1 = 0.05$. The true values of the parameters associated with dropout rate were similar to the values obtained based on the real data: $b_0 = 0.14$ and $b_1 = -0.02$ (dropout rates for $\mathbf{X}_1$ and $\mathbf{X}_2$ are both 0.44).

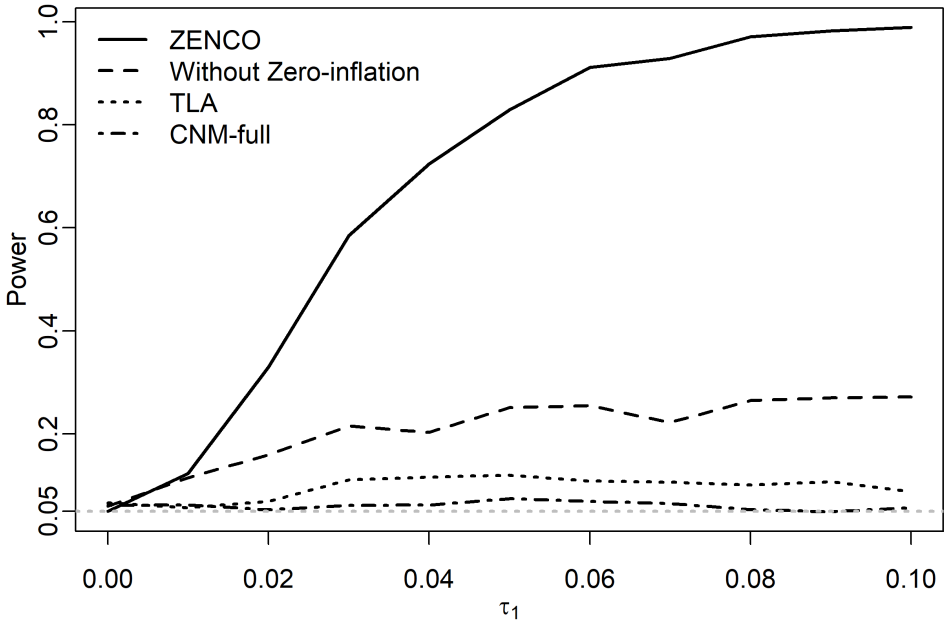

Figure 2.2. Power curves comparing various methods. Both TLA and CNM-Full approaches are Gaussian-based models.

The empirical 95% coverage probabilities from the posterior distributions and the length of credible intervals are shown in Table 2.1. In Table 2.1, we also presented the parameter estimates using a negative binomial model without zero-inflation. The empirical 95% coverage probability is calculated as the percentage of times when the 95% credible intervals covering the true parameter value based on 1,000 MCMC simulations. The simulation results shown in Table 2.1 suggest that ZENCO model provides a much better 95% coverage probability than a negative binomial regression method model without zero-inflation.

MSEs and MBEs are shown in Table 2.2. The MBE of a given parameter β is calculated as $\frac{1}{N}\sum_{i=1}^{N}(\widehat{\beta}_i - \beta)$; N is the number of simulation iterations (N=1,000). Based on the simulation results in Table 2.2, ZENCO model has smaller MSEs and MBEs comparing with the non-zero-inflated negative binomial regression method.

Table 2.1. Coverage probability of 95% credible intervals (CIs) and interval lengths based on 1,000 MCMC simulations ($\tau_0 = 0.01$, $\tau_1 = 0.05$)

		Without Zero-inflation		With Zero-inflation	
	Parameter	Coverage probability	CI length	Coverage probability	CI length
$N = 200$	τ_0	1.000	0.237	1.000	0.246
	τ_1	0.154	0.041	0.957	0.095
$N = 500$	τ_0	1.000	0.223	1.000	0.244
	τ_1	0.006	0.022	0.961	0.059
$N = 1,000$	τ_0	0.957	0.205	1.000	0.242
	τ_1	0.000	0.015	0.954	0.040

Table 2.2. Mean square errors (MSE) and mean bias errors (MBE) based on 1,000 MCMC simulations ($\tau_0 = 0.01$, $\tau_1 = 0.05$)

		Without Zero-inflation		With Zero-inflation	
	Parameter	MSE	MBE	MSE	MBE
$N = 200$	τ_0	0.001	0.005	0.000	-0.008
	τ_1	0.002	-0.039	0.001	-0.006
$N = 500$	τ_0	0.002	0.024	0.000	-0.009
	τ_1	0.002	-0.040	0.000	-0.001
$N = 1,000$	τ_0	0.004	0.048	0.000	-0.009
	τ_1	0.002	-0.041	0.000	0.000

2.3.4 Scenario 4: Robustness

In order to assess the robustness of the ZENCO method under model misspecification, we conducted three sets of simulations where the data are generated via a negative binomial model without zero-inflation. The three sets of simulation studies were performed with sample sizes 200, 500, and 1,000, and each with 1,000 simulation iterations. The true values of parameters were set as $\mu_1 = \mu_2 = \mu_3 = 15$, $\phi_1 = \phi_2 = \phi_3 = 4$, $\tau_0 = 0.01$, and $\tau_1 = 0.05$. We analyzed the simulated datasets using a negative binomial regression method without zero-inflation and the ZENCO method.

The empirical 95% coverage probabilities from posterior distributions and the length of credible intervals using the above two models are shown in Table 2.3; the MSEs and MBEs are shown in Table 2.4. The results shown in Table 2.3 and Table 2.4 suggest that our proposed estimation procedure in ZENCO is fairly robust even when the data are generated from a non-zero-inflated negative binomial setting.

Table 2.3. Coverage probability of 95% credible intervals (CIs) and interval lengths for τ_0 and τ_1 estimates based on 1,000 MCMC simulations with non-zero inflated data ($\tau_0 = 0.01$, $\tau_1 = 0.05$)

		Without Zero-inflation		With Zero-inflation	
	Parameter	Coverage probability	CI length	Coverage probability	CI length
$N = 200$	τ_0	1.000	0.235	1.000	0.238
	τ_1	0.951	0.034	0.957	0.037
$N = 500$	τ_0	1.000	0.220	1.000	0.223
	τ_1	0.957	0.021	0.949	0.022
$N = 1,000$	τ_0	0.999	0.200	0.999	0.203
	τ_1	0.953	0.015	0.958	0.016

Table 2.4. Mean square errors (MSE) and mean bias errors (MBE) based on 1,000 MCMC simulations with non-zero inflated data ($\tau_0 = 0.01$, $\tau_1 = 0.05$)

		Without Zero-inflation		With Zero-inflation	
	Parameter	MSE	MBE	MSE	MBE
$N = 200$	τ_0	0.000	-0.009	0.000	-0.010
	τ_1	0.000	0.000	0.000	0.002
$N = 500$	τ_0	0.001	-0.009	0.001	-0.009
	τ_1	0.000	0.000	0.000	0.001
$N = 1,000$	τ_0	0.001	-0.008	0.001	-0.008
	τ_1	0.000	0.000	0.000	0.001

2.3.5 Scenario 5: A multiple-gene setting

In this simulation scenario, we turn our attention to a multiple-gene setting. Our goal here is to demonstrate our proposed approach could capture dependencies among multiple genes through multiple pairwise searches. We set $b_0 = 0.65$ and $b_1 = -0.015$, which is similar to the values obtained based on the real data and then simulated 5 genes (10 gene pair combinations) with $\mu_1 = 15$, $\mu_2 = 19$, $\mu_3 = 10$, $\mu_4 = 15$, $\mu_5 = 12$, $\phi_1 = 4$, $\phi_2 = 5$, $\phi_3 = 6$, $\phi_4 = 4$, $\phi_5 = 3$. The true values of the 10 $\tau_1's$ range from 0.005 to 0.05, while the true value of τ_0 was set as 0. The empirical 95% coverage probabilities and MBEs of 10 $\tau_1's$ are shown in Table 2.5. The results indicate that our method demonstrated desirable performance under a multiple-gene setting.

Table 2.5. Coverage probability of 95% credible interval and MSEs for τ_1 estimates based on 1,000 MCMC simulations for 5 genes (10 gene pairs)

Pair	τ_1	$\widehat{\tau_1}$	MSE	Coverage probability
(G_1, G_2)	0.0050	0.0054	0.0003	0.952
(G_1, G_3)	0.0100	0.0091	0.0004	0.962
(G_1, G_4)	0.0150	0.0128	0.0003	0.971
(G_1, G_5)	0.0200	0.0161	0.0004	0.960
(G_2, G_3)	0.0250	0.0211	0.0004	0.964
(G_2, G_4)	0.0300	0.0256	0.0004	0.952
(G_2, G_5)	0.0350	0.0292	0.0004	0.930
(G_3, G_4)	0.0400	0.0344	0.0004	0.949
(G_3, G_5)	0.0450	0.0385	0.0005	0.941
(G_4, G_5)	0.0500	0.0435	0.0004	0.949

2.4 Experimental Data Analysis

We used the proposed ZENCO model to analyze the melanoma data set described in Rambow et al. (2018). The scRNA-seq data were obtained from Gene Expression Omnibus (GEO accession number: GSE116237). The data set consists of 57,445 genes and 674 melanoma cells. To study minimal residual disease (MRD) as well as relapse during melanoma treatment, Rambow et al. (2018) performed single-cell RNA sequencing using malignant cells from BRAF-mutant patient-derived xenograft melanoma cohorts treated with BRAF/MEK inhibitor (dabrafenib/trametinib).

During the course of continuous treatment with BRAF/MEK inhibitor, the transition of tumor cells can be categorized into three phases: phase 1 is in the early stage when all treated lesions rapidly shrunk upon initial treatment (BRAF-inhibitor sensitive); phase 2 is the second stage when drug-tolerant tumor cells remain viable upon continuous treatment (MRD); in phase 3, relapse is observed and tumor cells exhibit adaptive resistance to continuous BRAF inhibition treatment (BRAF-inhibitor resistance). Among the 674 melanoma cells in the data set, there are 155 phase 1 cells, 199 phase 2 cells, and 148 phase 3 cells. More details can be found in Rambow et al. (2018).

To gain insight into transcriptional switches of genetic circuits in tumor cells during the course of BRAF-inhibitor treatment, we set out to identify gene pairs that interact with BRAF differently between BRAF-inhibitor sensitive cells (phase 1) and BRAF-inhibitor resistance cells (phase 3). Hence in this analysis, we chose BRAF as $\mathbf{X}_3$ and conducted the pairwise analysis for genes in the melanoma pathway described in the KEGG database (Kanehisa and Goto, 2000). According to the melanoma pathway in KEGG database, 72 genes were identified as melanoma-associated genes. The data were first preprocessed by the procedures described in McCarthy et al. (2017). After removing low expressed genes (maximum count across all cells less than 5) and genes with more than 70% zeros in either phase 1 cells or phase 3 cells, 28 genes were selected for further analysis.

The study-specific parameters, b_0, b_1, associated with dropout rates can be estimated using the logistic function $p = \frac{e^{(b_0+b_1\mu)}}{1+e^{(b_0+b_1\mu)}}$. In the logistic function, we used the sample mean to estimate μ. After calculating the dropout rate as the proportion of cells with zero counts, a non-linear least-squares approach was then applied to calculate b_0 and b_1.

We implemented ZENCO analyses for 351 gene pair combinations in phase 1 cells and phase 3 cells and obtained the estimates of τ_1. To identify the gene pairs that interact with BRAF differently, we chose gene pairs that are in both phase 1 and phase 3 cells and calculated the differences of τ_1 estimates between the two phases. The top 30 gene pairs with the largest differences of τ_1 between phase 3 and phase 1 are shown in Table 2.6. The first two columns in Table 2.6 are the names of two genes. $\tau_1(P1)$ is the estimated τ_1 in phase 1 cells, and $\tau_1(P3)$ is the estimated τ_1 in phase 3 cells. $\Delta\tau_1$ is defined as $\tau_1(P3) - \tau_1(P1)$. It quantifies the change of dynamic co-expression in relation to BRAF between phase 3 and phase 1 cells.

From Table 2.6, we observed that genes PDGFC and FGFR1 have the largest $|\Delta\tau_1|$ between phase 1 and phase 3 cells. In phase 1 cells, the estimate of τ_1 for PDGFC

and FGFR1 is 0.045 and the 95% credible interval does not contain 0. In phase 3 cells, the estimate of τ_1 is close to 0. This suggests that the regulatory mechanism between BRAF and the gene pair (PDGFC, FGFR1) changes between phase 1 and phase 3 cells. Czyz (2019) pointed out that melanoma cells somehow acquire the ability to grow independent of the two growth factors: FGFR1, PDGFC which helps melanoma cells to gain resistance toward BRAF treatment. Our finding from Table 2.6 is consistent with this finding. Interestingly, many top gene pairs listed in Table 2.6 are from the mitogen-activated protein kinase (MAPK) and phosphoinositide 3-kinase (PI3K) signaling pathways. Our analysis findings support the hypotheses described in Villanueva et al. (2011).

In the above analysis, the convergence of MCMC was assessed using the Gelman-Rubin convergence statistic (Gelman et al., 1992). The convergence statistics were close to 1 for all τ_1 estimates in all 351 gene pairs. The trace plots of the top 5 gene pairs are shown in Figure 2.3. In our real data application, it took 67 minutes to implement ZENCO with 3 chains (100,000 iterations each) for all 351 gene combinations using 13 computing cluster nodes (each with 28 2.4 GHz Intel Xeon E5-2680 v4 processors).

2.5 Discussion

In this chapter, we presented a zero-inflated negative binomial dynamic correlation model for studying covariate-dependent correlations in zero-inflated, over-dispersed count data, such as scRNA-seq data. In our model, the correlation of two genes is regulated by the expression level of the third gene; a phenomenon we named dynamic correlation in this chapter. This novel dynamic correlation focuses on studying the changes of conditional correlation. It is a different measure from the partial correlation coefficient. The partial correlation quantifies the amount of residual correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ after regression on $\mathbf{X}_3$ to adjust for the influence of $\mathbf{X}_3$ (Li, 2002).

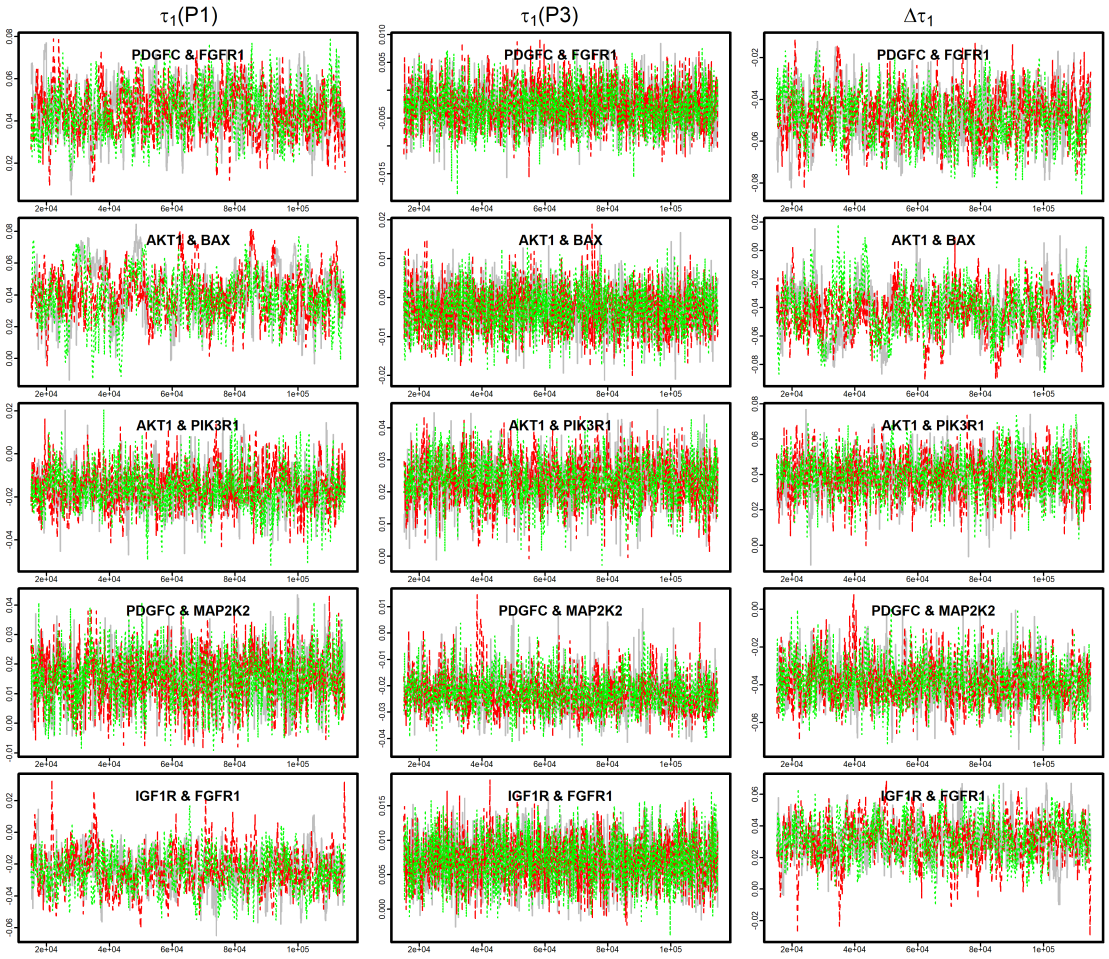

Figure 2.3. Trace plots of the top 5 gene pairs reported in the experimental data analysis

The proposed model in this chapter takes both over-dispersion and zero-inflation of the data into consideration. With the proper choice of the values of parameters τ_0 and τ_1, the relationship between conditional correlation and the expression level of the third gene can be positive or negative. As demonstrated by our simulation studies, the ZENCO model significantly outperforms other existing approaches.

Two other prior distributions for the dispersion parameters ϕ_1, ϕ_2, and ϕ_3 have been implemented: an informative Gamma distribution on $\frac{1}{\phi}$ and a half-t distribution on $\sqrt{\phi}$. Our sensitivity analysis suggests that the ϕ_1, ϕ_2 and, ϕ_3 estimates are robust regardless of prior distribution assumptions. The Gamma distribution with mean 100 and relatively large variance 10,000 used in this chapter is more general and has

slightly better performance in MCMC parameter estimates.Moreover, in our model, ρ is the correlation of the latent variable Z. The Fisher transformation of ρ is assumed to be linear with X_3. In a more general setting, the relationship between $\log(\frac{1+\rho}{1-\rho})$ and $\mathbf{X}_3$ does not have to be linear. And our model can be easily adapted to other functional forms.

In the melanoma data analysis, $\mathbf{X}_3$ was used to denote the expression level of BRAF. And ZENCO model was implemented for each pairwise combination of $\mathbf{X}_1$ and $\mathbf{X}_2$ in the KEGG melanoma pathway. Using this search strategy, we found the pairs of genes whose BRAF-associated dynamic correlations change significantly between different phases during treatment. In Table 2.6, we reported the top genes with the largest $|\Delta\tau_1|$. Several existing type I error control approaches can be used in conjunction with the Bayesian model framework in ZENCO such as Käll et al. (2008); Dawson and Kendziorski (2012). As described in Section 2, there are several ways to implement ZENCO in a genomic study. If a pre-filtering step is used before implementing ZENCO, considerations described in Dawson and Kendziorski (2012); van Iterson et al. (2010) could be helpful to maintain type I error control.

Furthermore, in our application, $\mathbf{X}_3$ was used to denote the gene expression level of the BRAF gene because of its pivotal role in melanoma treatment and relapse in the study. In practice, the $\mathbf{X}_3$ can be easily modified to represent the activity level of a biological process or different cell types, or various cellular conditions such as tumor status, survival probability, degree of inflammation, metastasis potential,...etc. Also, $\mathbf{X}_3$ can be easily extended to represent a linear combination of several covariates or biological processes to accommodate the complexity of biological systems in other applications.

Because several existing procedures are available for pre-processing scRNA-seq data to remove low-magnitude background noise, in the ZENCO model, the dropout component is modeled as a degenerate distribution with a point mass at zero. How-

ever, the method can be easily adapted to allow a low-magnitude Poisson distribution to model the background noise in the dropout component.

In this chapter, our focus is on the changes in co-expression patterns between a gene pair. It is plausible that there might exist higher-order interactions between genes (more than two genes), a generalization of our approach to higher dimensions is feasible. However, special treatments need to be considered to guarantee the positive definiteness of the variance-covariance matrix in higher-dimension.

Table 2.6. Top table of dynamic correlations differences. $\Delta\tau_1$ is the difference between τ_1 estimates in Phase 3 (P3) and Phase 1 (P1).

#	Gene1	Gene2	$\tau_1(P1)$	$\tau_1(P3)$	$\Delta\tau_1$
1	PDGFC	FGFR1	0.045 (0.021, 0.068)	-0.003 (-0.010, 0.005)	-0.047 (-0.072,-0.023)
2	AKT1	BAX	0.040 (0.008, 0.071)	-0.003 (-0.014, 0.008)	-0.043 (-0.075,-0.010)
3	AKT1	PIK3R1	-0.016 (-0.035, 0.004)	0.024 (0.009, 0.038)	0.040 (0.015, 0.062)
4	PDGFC	MAP2K2	0.016 (-0.002, 0.032)	-0.023 (-0.036,-0.006)	-0.039 (-0.059,-0.013)
5	IGF1R	FGFR1	-0.024 (-0.048, 0.000)	0.007 (0.000, 0.014)	0.032 (0.006, 0.056)
6	MDM2	CCND1	0.021 (0.007, 0.031)	-0.011 (-0.018,-0.004)	-0.031 (-0.044,-0.017)
7	AKT1	ARAF	-0.025 (-0.047, 0.002)	0.007 (-0.007, 0.018)	0.031 (0.002, 0.056)
8	AKT1	MAP2K1	0.025 (0.004, 0.057)	-0.006 (-0.017, 0.009)	-0.030 (-0.063,-0.006)
9	AKT1	MAPK1	-0.003 (-0.012, 0.006)	0.026 (0.007, 0.055)	0.029 (0.007, 0.058)
10	KRAS	PDGFC	0.012 (-0.005, 0.024)	-0.017 (-0.042, 0.005)	-0.029 (-0.057,-0.002)
11	IGF1R	MAP2K2	0.025 (0.002, 0.056)	-0.004 (-0.011, 0.006)	-0.028 (-0.060,-0.004)
12	PTEN	PDGFC	-0.022 (-0.036,-0.004)	0.007 (-0.003, 0.014)	0.028 (0.008, 0.044)
13	PTEN	PIK3R1	0.031 (0.007, 0.050)	0.005 (-0.006, 0.014)	-0.027 (-0.048,-0.002)
14	BAX	POLK	0.025 (0.006, 0.048)	0.000 (-0.012, 0.010)	-0.026 (-0.051,-0.003)
15	KRAS	NRAS	0.017 (-0.003, 0.034)	-0.008 (-0.015, 0.002)	-0.024 (-0.043,-0.003)
16	ARAF	RB1	0.020 (0.008, 0.032)	-0.004 (-0.009, 0.002)	-0.024 (-0.037,-0.011)
17	AKT1	RAF1	-0.016 (-0.033,-0.003)	0.007 (-0.004, 0.017)	0.023 (0.006, 0.042)
18	NRAS	MAPK1	0.017 (0.002, 0.029)	-0.005 (-0.013, 0.006)	-0.021 (-0.037,-0.004)
19	PIK3R1	MDM2	0.020 (0.004, 0.035)	-0.001 (-0.010, 0.008)	-0.021 (-0.038,-0.002)
20	IGF1R	TP53	-0.016 (-0.034, 0.002)	0.005 (-0.003, 0.011)	0.020 (0.002, 0.039)
21	BAK1	POLK	-0.018 (-0.030,-0.006)	0.002 (-0.006, 0.010)	0.020 (0.006, 0.034)
22	AKT3	MAP2K2	0.016 (0.005, 0.025)	-0.003 (-0.011, 0.007)	-0.018 (-0.030,-0.006)
23	PTEN	KRAS	-0.005 (-0.016, 0.011)	0.012 (0.003, 0.020)	0.017 (0.000, 0.030)
24	BAD	RAF1	-0.016 (-0.031,-0.006)	0.000 (-0.009, 0.008)	0.016 (0.002, 0.032)
25	IGF1R	CDK6	0.014 (-0.001, 0.026)	-0.002 (-0.008, 0.003)	-0.016 (-0.029,-0.001)
26	RB1	CCND1	0.011 (0.000, 0.020)	-0.004 (-0.010, 0.004)	-0.014 (-0.025,-0.002)
27	AKT2	FGFR1	-0.003 (-0.015, 0.006)	0.011 (0.004, 0.017)	0.014 (0.002, 0.027)
28	BAD	TP53	-0.001 (-0.010, 0.007)	0.013 (0.002, 0.021)	0.014 (0.001, 0.026)
29	NRAS	BAK1	0.001 (-0.008, 0.008)	0.014 (0.006, 0.022)	0.014 (0.002, 0.025)
30	AKT2	BAK1	-0.004 (-0.013, 0.005)	0.010 (0.000, 0.019)	0.014 (0.001, 0.026)

Chapter 3

Clustering using gene co-expression latent factors

3.1 Introduction

Biological molecules in a cell often participate in complicated interaction processes. Due to the complex regulatory mechanisms, genes are often tightly regulated (de la Fuente, 2010b). To understand the roles of genes, both the average expression levels of genes and interactions between genes are needed to be taken into consideration. Genetic interactions analysis can be used to associate genes with biological processes or to discern gene regulatory mechanisms (Van Dam et al., 2018). With recent advances in next-generation sequencing, researchers are now able to study genetic interactions systematically. In addition, interactions between genes are often highly dynamic under different cell types or cellular conditions such as tumor status (Luscombe et al., 2004; de Lichtenberg et al., 2005). Changes in gene interactions can often result in changes in co-expression patterns between genes. Therefore, differential co-expression analysis can be used to identify sets of genes whose expressions change in a coordinated fashion across different cell types.

The studies of how genes interact in a cell are important when clustering cell types, since genes in different cell types could interact differently. Clustering has been the routine of single-cell data analysis and cell-type identifying. The single-cell RNA sequencing technologies have made it possible to discover the groupings of a set of cells on the basis of transcriptome similarity. Diverse types of clustering approaches

have been developed. For example, the two widely-used clustering algorithms are k-means clustering and hierarchical clustering. However, both k-means and hierarchical clustering have their own limitations when it comes to high-dimensional data sets (Kiselev et al., 2019). Therefore, graph-based clustering approaches are becoming more and more popular in scRNA-seq data and have been incorporated into some user-friendly packages such as Seurat (Stuart et al., 2019) and scanpy (Wolf et al., 2018).

Although great progress has been made in terms of clustering approaches, the effects of the genetic interactions have not been taken into consideration. The traditional clustering approaches use average expression levels of the data and are therefore insufficient for understanding the intricate regulatory mechanisms that underlie different cellular conditions. The co-expression structures can often bring insights into the complex genetic interactions (Yu, 2018) and can help detect correlation changes between pairs of genes across different modulating conditions. Therefore, the co-expression structures can help identify hidden sub-groups in the data and improve the performance of clustering.

Latent pathway activities can affect both the average expression levels of genes and the co-expression levels between genes. Therefore, both expression levels and co-expression levels contain a part of the latent pathway activities information. Some pathway activities may only affect the average expression levels of genes, some may only affect the co-expression levels between genes, and some may affect both. Dimensionality reduction approaches such as JIVE (Lock et al., 2013) and MSFA (De Vito et al., 2019) consider the mixed factors and can be used to decompose the data (Cantini et al., 2021). The ability of the dimensionality reduction approaches to separately estimate the joint structure and source-specific structures can contribute significantly to the interpretation of the latent pathway activities. We consider both the expression structures and co-expression structures of the data to find the latent pathway

activities in different sources. We then use the latent factors to conduct cell-type clustering.

In this chapter, we propose a cell-type clustering approach that allows for joint analysis of both expression structures and co-expression structures of the data. Our method learns the joint features shared among data sources and identifies the unique variation present in each source to further cluster the cell types. The remainder of the chapter is arranged as follows. In Section 3.2, the detail of the proposed model is introduced. The simulation studies and comparisons are conducted in Section 3.3. In Section 3.4, the analysis of scRNA-seq data generated from breast cancer cells is presented.

3.2 Method

We consider two matrices in the model: original matrix $\mathbf{X}$ and product matrix $\mathbf{X_{prod}}$. $\mathbf{X_{prod}}$ is the differential co-expression patterns calculated from $\mathbf{X}$. Each matrix has n columns representing n cells in gene expression data. $\mathbf{X}$ is a $p_1 \times$ n matrix, and $\mathbf{X}_{prod}$ is a $p_2 \times$ n matrix. In our method, we treat them as two data sources. The original matrix $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_{p_1})^T$ is the matrix of gene expression data. $\mathbf{X}_i$ is the vector of the expressions of the ith gene. Without loss of generality, we normalized the data with mean 0 and variance 1 for every gene. Hence, the correlation of the ith gene and the jth gene

$$Corr(\mathbf{X}_i, \mathbf{X}_j) = E(\mathbf{X}_i \mathbf{X}_j). \tag{3.1}$$

The product of gene pairs $\mathbf{X}_i$ and $\mathbf{X}_j$ estimates correlation and captures the interaction between these two genes. Since the number of genes in $\mathbf{X}_{p_1 \times n}$ is p_1, there will be $p' = \binom{p_1}{2}$ gene pairs in total. We can construct a new matrix $\mathbf{W}_{p' \times n} = (\mathbf{W}_1, ..., \mathbf{W}_{p'})^T$ with $\binom{p_1}{2}$ rows and n columns. Each row of $\mathbf{W}_{p' \times n}$ is equal to the entry-wise product of $\mathbf{X}_i$ and $\mathbf{X}_j$, $\forall\ i \neq j$. Usually, the number of genes p_1 in the gene expression data is over 10^4. In this situation, there are around 10^8 different gene pairs. Therefore, we

need to prescreen the gene pairs and only keep the gene pairs that are informative on the cellular states. In other words, we need to find the gene pairs that are likely to have dynamic co-expression patterns in different types of cells in the data. The difficulty is that the cell types of the data are unknown. Yu (2018) solves this problem using a new metric, called Liquid Association Coefficient (LAC) score without knowing the cellular states or sub-classes in advance. The LAC score is defined as

$$LAC_{ij} = r(|X_i|, |X_j|) - |r(X_i, X_j)|, \tag{3.2}$$

where r() is the Pearson correlation coefficient. It has been shown that gene pairs with higher LAC scores are more likely to have dynamic correlation patterns. We then use LAC scores for all gene pairs to find the top p_2 potential dynamic correlated gene pairs and store them in the new matrix $\mathbf{X_{prod}}$.

The latent pathway activities can affect both the original matrix $\mathbf{X}$ and the product matrix $\mathbf{X_{prod}}$. Therefore, factor analysis is needed to find the factors associated with the latent processes. In addition, considering the high dimensionality of $\mathbf{X}$ and $\mathbf{X_{prod}}$, feature selection or dimension reduction are needed to speed up calculations. Traditional approaches only perform dimension reduction on original matrix $\mathbf{X}$, while our model uses both expression matrix $\mathbf{X}$ and co-expression matrix $\mathbf{X_{prod}}$. Since the size and the scale of $\mathbf{X}$ and $\mathbf{X_{prod}}$ often differ, we center and scale both matrices to have a mean 0 and a variance of 1 within each row. The model can be defined as follows:

$$\begin{aligned} \mathbf{X} &= \mathbf{J_1} + \mathbf{I_1} + \mathbf{U_1} \\ \mathbf{X_{prod}} &= \mathbf{J_2} + \mathbf{I_2} + \mathbf{U_2} \end{aligned} \tag{3.3}$$

Let $\mathbf{I_1}$ be the matrix representing the individual structure of $\mathbf{X}$, and let $\mathbf{I_2}$ be the matrix representing the individual structure of $\mathbf{X_{prod}}$. $\mathbf{J_1}$ and $\mathbf{J_2}$ represent the joint factor matrices of $\mathbf{X}$ and $\mathbf{X_{prod}}$. $\mathbf{U_1}$ and $\mathbf{U_2}$ are the matrices of residuals. In a factorized model, for $i = 1, 2$, the joint structure $\mathbf{J_i}$ can be written as $\mathbf{W_i F}$. The individual structure matrix $\mathbf{I_i}$ can be written as $\mathbf{B_i F_i}$. Therefore, the model can be

written as:

$$\begin{aligned} \mathbf{X} &= \mathbf{W_1 F} + \mathbf{B_1 F_1} + \mathbf{U_1} \\ \mathbf{X_{prod}} &= \mathbf{W_2 F} + \mathbf{B_2 F_2} + \mathbf{U_2} \end{aligned} \tag{3.4}$$

The joint structure is represented by the common factor matrix $\mathbf{F}$. The loading matrices $\mathbf{W_i}$ indicate how these joint factors are expressed in the rows of the corresponding matrix. Individual structure $\mathbf{I_i}$ is represented by the factor matrix $\mathbf{F_i}$ with variable loading matrix $\mathbf{B_i}$. The original matrix $\mathbf{X}$ comprises joint structure $\mathbf{W_1 F}$, individual structure $\mathbf{B_1 F_1}$, and residual $\mathbf{U_1}$. Both the joint and individual structure of $\mathbf{X}$ can be obtained using traditional dimension reduction. The joint structure $\mathbf{W_2 F}$ cannot provide any additional information than $\mathbf{W_1 F}$, because they share the same factor matrix $\mathbf{F}$. Therefore, the individual structure $\mathbf{B_2 F_2}$ is the additional information provided by product matrix $\mathbf{X_{prod}}$.

The joint structure and individual structure between the original matrix and product matrix in the above model can be found by Joint and Individual Variation Explained (JIVE) (Lock et al., 2013). JIVE is a decomposition of data as the sum of two factorizations: a low-rank matrix capturing joint features across two data sources, and low-rank omics-specific factor matrices capturing individual features. It's worth noting that the matrix capturing individual features of co-expression structures contains additional information which the first-order structures can not provide. The ranks of estimated joint and individual structures can be determined by the number of components in PCA with the proportion of variance explained > 0.05. The ranks can also be determined by JIVE using permutation testing. Since the individual structure $\mathbf{I_2} = \mathbf{B_2 F_2}$ contains additional information provided by the product matrix, which has not been considered in the traditional clustering approaches, we can use this additional information for further clustering. The further clustering using $\mathbf{I_2}$ to find sub-classes of cell types is the same procedure as the traditional clustering methods. Therefore, any clustering method that works well on traditional gene ex-

pression data can be used on $\mathbf{I_2}$. According to (Kiselev et al., 2019), the combination of shared-nearest-neighbor graphs and Louvain community detection, which has been used in packages such as Seurat and scanpy, is one of the popular choices of clustering methods. We adapt the same methods for further clustering the cell types using $\mathbf{I_2}$.

3.3 Simulation

As a basic illustration, we use three latent vectors $\mathbf{z_1}$, $\mathbf{z_2}$, and $\mathbf{z_3}$ in the simulation to represent three sets of latent pathway activities. $\mathbf{z_1}$ controls the individual coexpression structure of the data, $\mathbf{z_2}$ controls the individual average expression structure of the data, and $\mathbf{z_3}$ controls the joint structure of the data. The simulated data contains two blocks, block 1 and block 2, with simple patterns corresponding to individual and joint structures respectively. In each block, we simulated 400 genes with 1000 samples. In block 1, genes are simulated with $\mathbf{z_1}$ and $\mathbf{z_2}$. In block 2, genes are simulated with $\mathbf{z_3}$. The patterns of $\mathbf{z_1}$, $\mathbf{z_2}$, and $\mathbf{z_3}$ are in Figure 3.1.

The data simulation schematic is demonstrated in Figure 3.2. The first heatmap represents the original matrix of the data. The columns of the heatmap are samples and the row of the heatmap are genes. The first half of original matrix is block 1, and second half is block 2. Block 1 is from gene 1 to gene 400, and block 2 is from gene 401 to gene 800. The second heatmap in Figure 3.2 is the ideal product matrix we are trying to simulate. The columns of the heatmap are samples and the row of the heatmap are gene pairs. The product matrix also have two blocks with different patterns. Moreover, we can observe that the block 2 in original matrix and the block 2 in the product matrix share the same pattern, while block 1 in these two matrices have different individual structures.

Specifically, to simulate block 1 of the data matrix, we first generated the latent vector $\mathbf{z_1}$ corresponding to 1000 samples. In this simulation study, the first 500 samples of $\mathbf{z_1}$ are -1 and the remaining 500 samples are 1. We then generated a latent

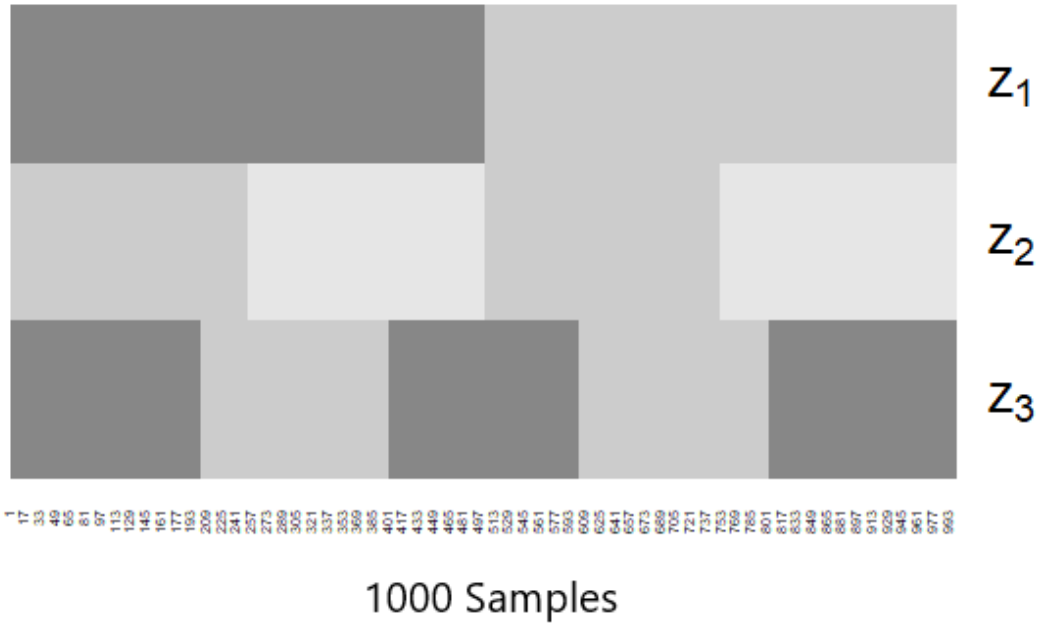

Figure 3.1. The patterns of latent vectors $\mathbf{z_1}$, $\mathbf{z_2}$, and $\mathbf{z_3}$. Different colors represent different values.

$\mathbf{z_2}$ with 1000 elements. Samples 1 to 250 and samples 501 to 750 in $\mathbf{z_2}$ are 1 and the remaining samples are 2. Then, correlation ρ was calculated as $\frac{exp(\tau_1 * z_1 + \tau_0) - 1}{exp(\tau_1 * z_1 + \tau_0) + 1}$. The 200 gene pairs (400 genes) were generated one by one through bivariate normal distribution using ($\mathbf{z_2}$, $\mathbf{z_2}$) as the mean and ρ as the correlations. The random noise generated from standard normal distribution was added to each gene. Using the simulation setting described above, the correlations and means of genes in block 1 are respectively controlled by $\mathbf{z_1}$ and $\mathbf{z_2}$. This represents the individual structure between the original matrix and product matrix of block 1. To generate block 2, we use latent vector $\mathbf{z_3}$ which has different pattern as $\mathbf{z_1}$ and $\mathbf{z_2}$. Then, ρ was calculated as $\frac{exp(\tau_1 * z_3 + \tau_0) - 1}{exp(\tau_1 * z_3 + \tau_0) + 1}$. The 200 gene pairs were generated one by one through bivariate normal distribution using ($\mathbf{z_3}$, $\mathbf{z_3}$) as the mean and ρ as the correlations. Therefore, genes in block 2 have joint structures for the original matrix and product matrix.

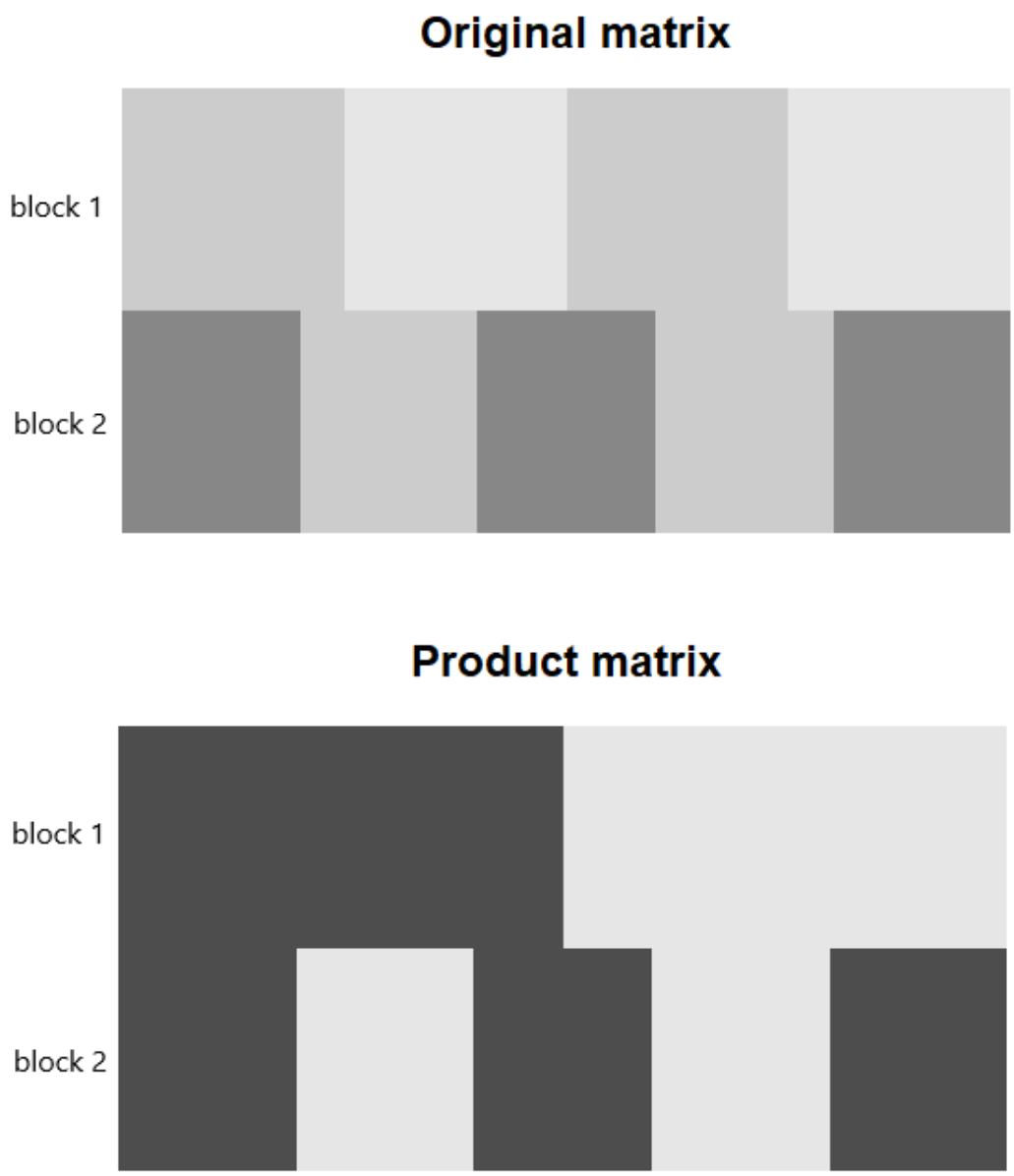

Figure 3.2. Simulation schematic for original matrix and product matrix in simulation. Different colors represent different values.

We then implemented the proposed method to cluster the samples. We first normalized the simulated data matrix with mean 0 and variance 1 for every gene. To select the gene pairs that were most likely to have the dynamic correlation, we used a pre-screening method proposed in (Yu, 2018). We computed the LAC scores for

all possible gene pairs and stored the gene pairs with LAC scores above 0.08 in the product matrix $\mathbf{X_{prod}}$. Our goal is to identify both the joint structure and individual structures of two matrices. Figure 3.3 shows the heatmaps of the original matrix and product matrix, the JIVE estimates for joint structure, individual structures, and random noises. The first row of Figure 3.3 represents the original matrix. We can see that the heatmap of the original matrix is the same as the original matrix in Figure 3.2. The second row of Figure 3.3 displays the product matrix of the simulated data. Our proposed method successfully find the individual structures and joint pattern between these two matrices. Actually, the individual pattern of the product matrix has the same pattern as $\mathbf{z_1}$. The individual pattern of the original matrix has the same pattern as $\mathbf{z_2}$. And the joint pattern between the original matrix and product matrix has the same pattern as $\mathbf{z_3}$.

Figure 3.4 shows the UMAP plot of the original simulated data. We can see that the four different clusters can be found using the original matrix. In fact, the original matrix only contains the patterns of $\mathbf{z_1}$ and $\mathbf{z_3}$. This is why the UMAP plot of it has four clusters ($\mathbf{z_1}$ and $\mathbf{z_3}$ both have two classes). The four clusters can be further clustered if we use the information from the product matrix. We use cluster 1 in Figure 3.4 as an example. Cluster 1 can not be further clustered using the original matrix. However, we can further cluster it using the individual pattern of the product matrix. Figure 3.5 shows the sub-cluster of cluster 1. The colors in Figure 3.5 indicate there are two different sub-clusters. It shows that cluster 1 can be further clustered using the individual pattern of the product matrix, which is also the pattern of $\mathbf{z_2}$.

3.4 Experimental Data Analysis

To demonstrate how our approach can be used to further cluster cell types, we analyzed the breast cancer data described in (Yeo et al., 2020; Fultang et al., 2021). The scRNA-seq data were obtained from Gene Expression Omnibus (GEO accession

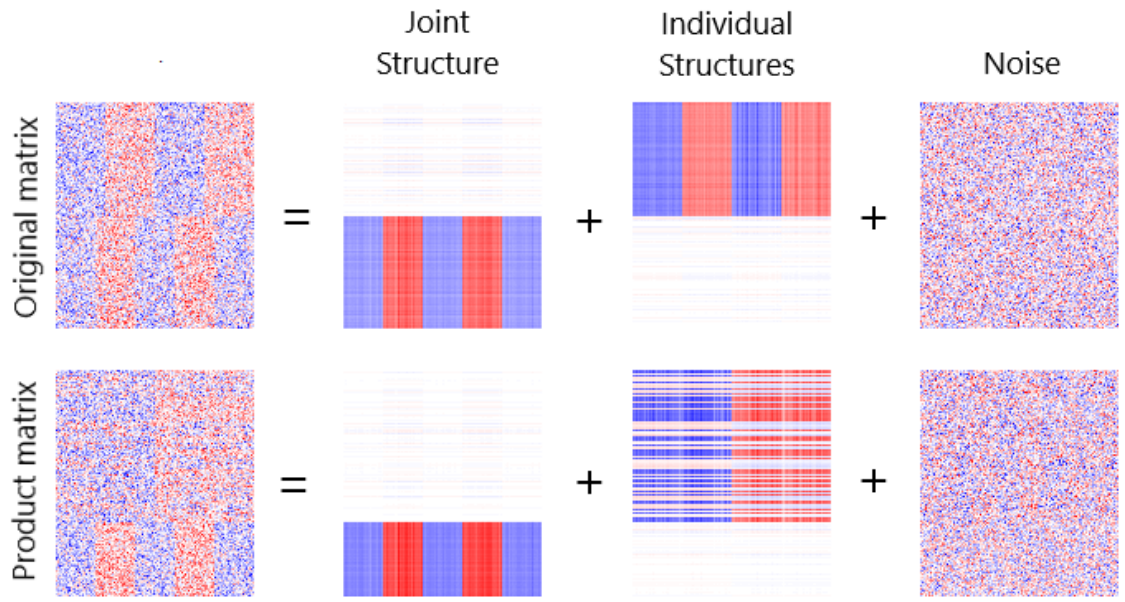

Figure 3.3. Heatmaps from the JIVE output.

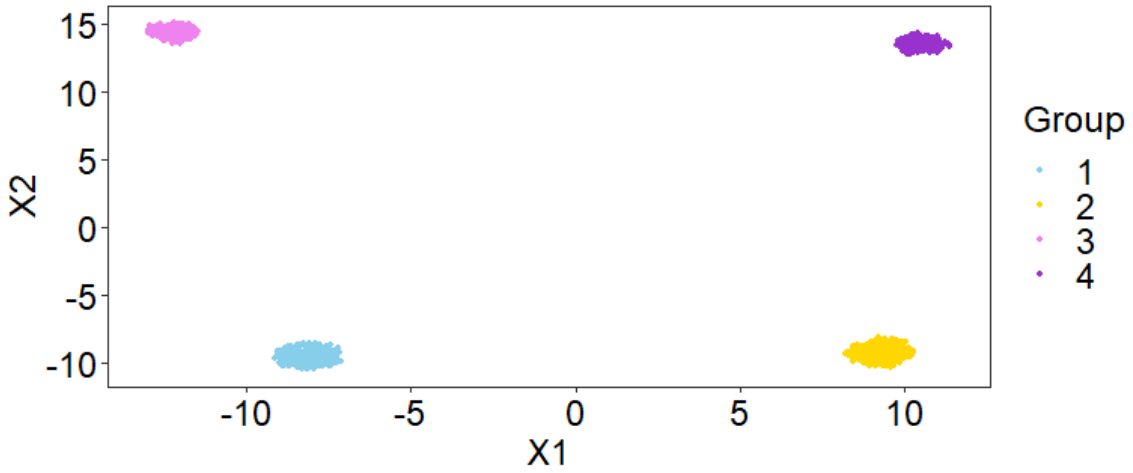

Figure 3.4. UMAP plot of original matrix in simulation.

number: GSE123366). The data set consists of 13,745 cells from four samples (1903 4T1 cells, 3102 BRCA1-null cells, 4173 PyMT cells, and 4567 Neu cells). In our study, we focused on finding the sub-clusters of BRCA1-null cells. The BRCA1-null data are from two different sample libraries. We first integrated the BRCA1-null cells from these two sample libraries using the Seurat package in R. Before the integration, we filtered cells that have unique feature counts over 6,000 or less than 200, normalized the data, and selected the top 2000 high cell-to-cell variation features in each dataset.

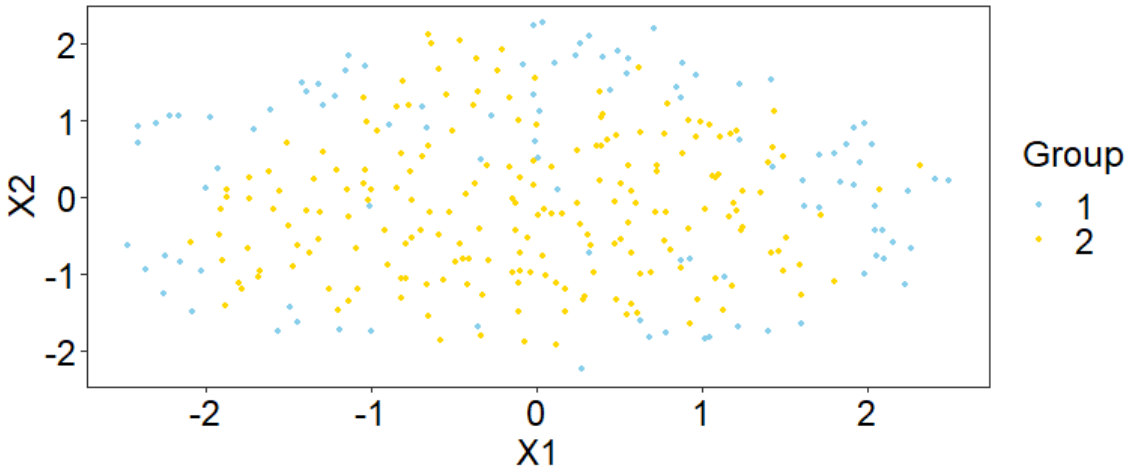

Figure 3.5. UMAP plot of cluster 1 only. The different colors represent different sub-clusters of cluster 1 found using individual pattern of the product matrix.

We then used the integrated dataset of BRCA1-null cells to conduct cell types clustering. The integrated dataset contains information of the expression level of genes or features. Traditional approaches such as the Seurat package conduct the cell-type clustering using only the integrated dataset itself. We first used the Seurat package to cluster the cell types of BRCA1-null cells. The UMAP plot of the clusters is in Figure 3.6. Traditional approaches can only cluster cell types to this extent because only the information of the original matrix is used.

To further cluster the cell types in Figure 3.6, more information are needed. We calculated the LAC score for each gene pair in integrated data to select the highly dynamic correlated gene pairs and generated the product matrix. The joint structure and individual structures between the integrated dataset and its product matrix were obtained from JIVE. In this analysis, we focused on cluster 0 in Figure 3.6, because the number of cells in cluster 0 is larger than in other clusters. The individual structure of the product matrix, which is the additional information provided by the product matrix, is denoted as $\mathbf{I_2}$. The UMAP plot of $\mathbf{I_2}$ matrix is shown in Figure 3.7. Figure 3.7 shows that $\mathbf{I_2}$ can be used for cell-type clustering just like the original matrix does. We used the $\mathbf{I_2}$ matrix to further cluster cell types in cluster 0. Figure

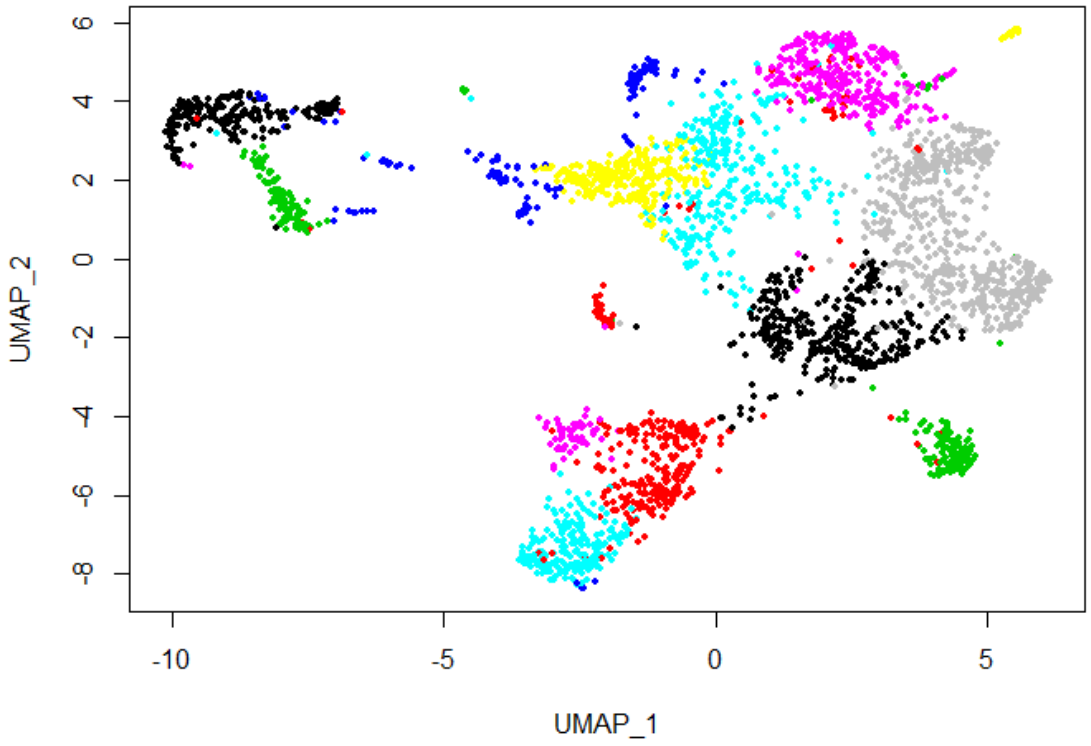

Figure 3.6. UMAP plot showing the individual clusters using original matrix.

3.8 is the UMAP plot for cluster 0 only. From Figure 3.8, we know that $\mathbf{I_2}$ contains additional information that can be used to further cluster the cells in cluster 0. Using $\mathbf{I_2}$, cluster 0 can be further clustered as 7 subclusters. Among those 7 subclusters in cluster 0, most cells are in subclusters 1, 3, 4, and 6. The heatmap of all cluster 0 cells in $\mathbf{I_2}$ matrix is shown in Figure 3.9. We can see that the four subclusters have distinct patterns in the heatmap. The rows of the matrix are different gene pairs. The heatmap shows that the co-expression patterns of gene pairs in different subclusters are different.

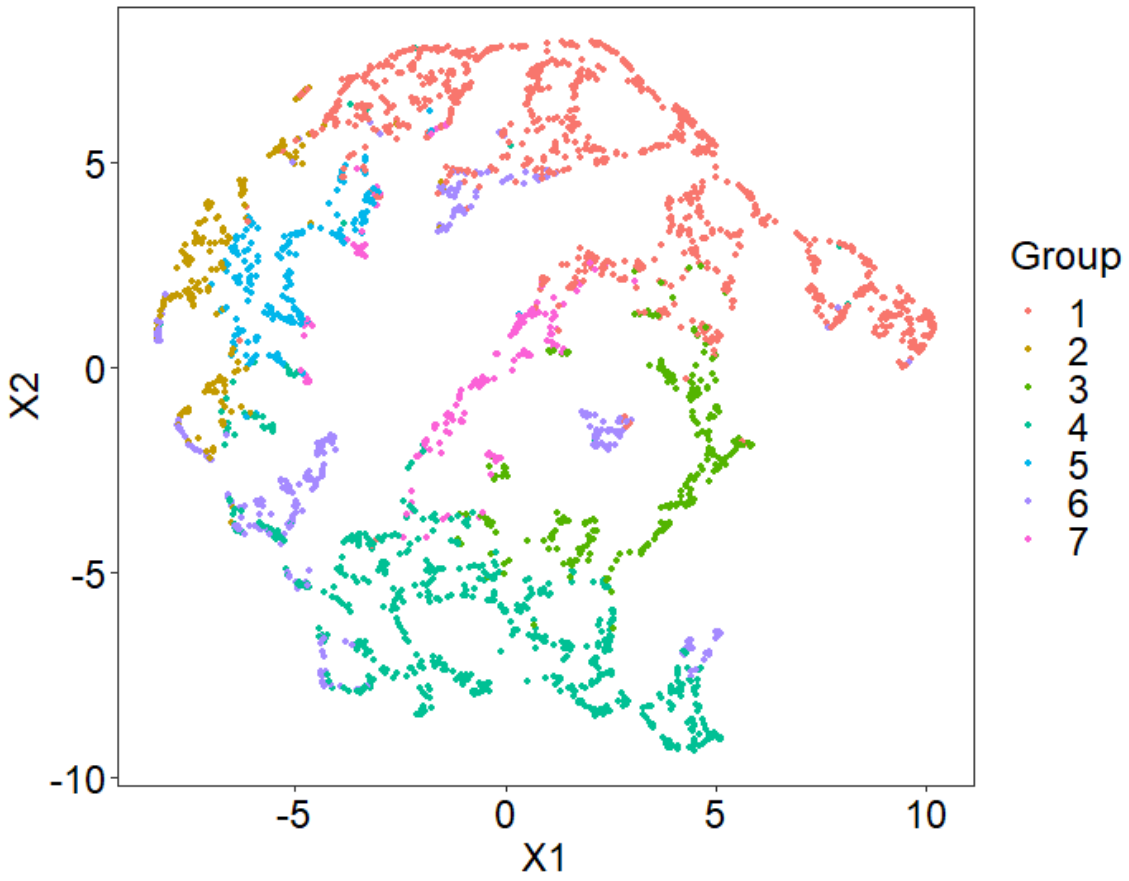

Figure 3.7. UMAP plot showing the individual clusters using $\mathbf{I_2}$ matrix.

3.5 Discussion

In this chapter, we presented a cell-type clustering approach that allows for joint analysis of both expression structure and co-expression structures of the data. Compared with traditional cell-type clustering approaches which only consider average expression structures of data, our approach can bring insight into the genetic interactions and help identify hidden sub-groups in clustering. The joint component and two individual components from the joint analysis can reveal different types of latent pathway activities.

Our method uses factor analysis and doesn't require distribution assumptions. Moreover, our method can adapt to any factor analysis approach. For instance, when analyzing zero-inflated data, we can use zero-inflated factor analysis (Pierson and

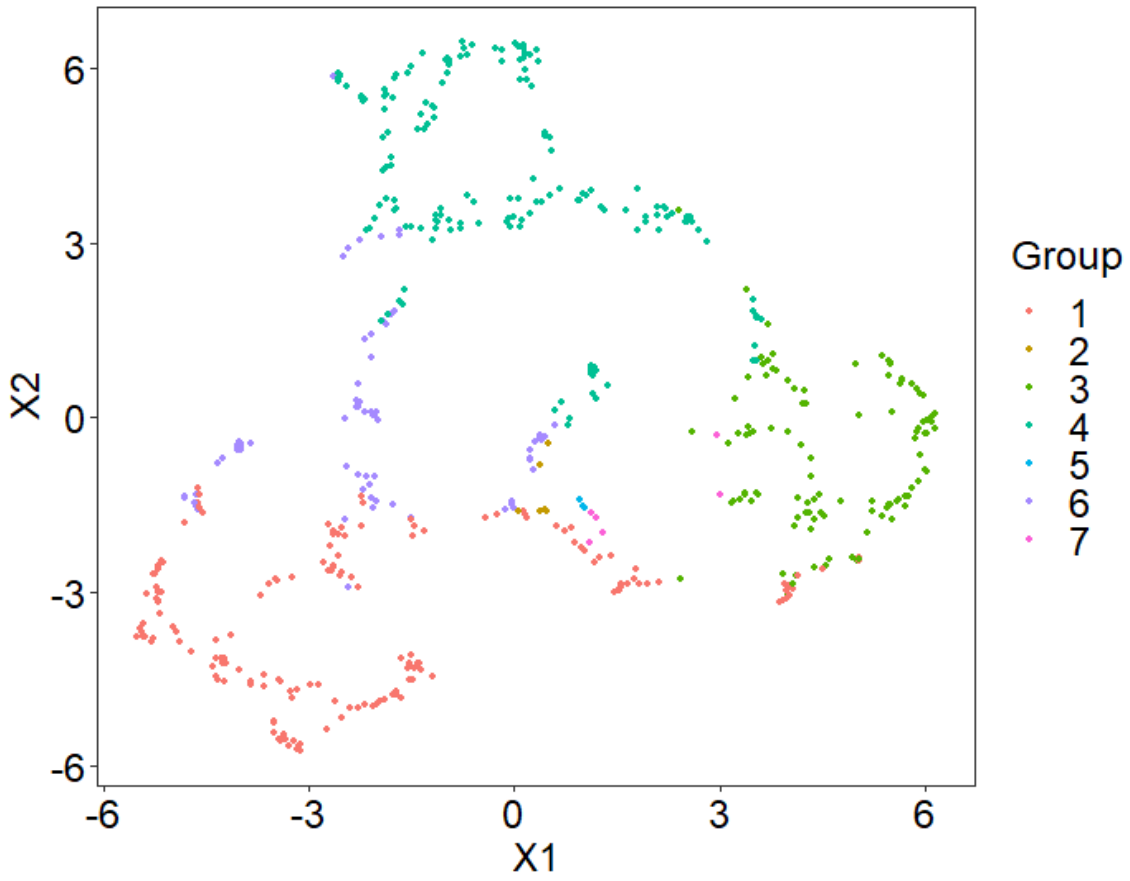

Figure 3.8. UMAP plot showing the subclusters of cluster 0 using $\mathbf{I_2}$ matrix.

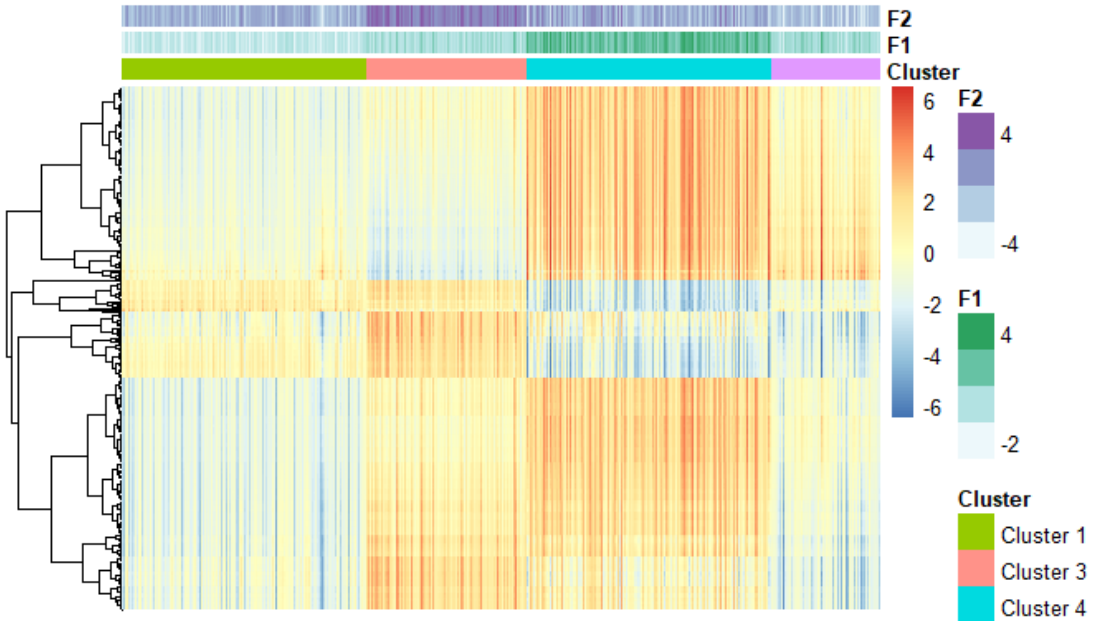

Figure 3.9. Heatmap of cluster 0 cells in $\mathbf{I_2}$ matrix.

Yau, 2015). Moreover, in this paper, we use shared-nearest-neighbor graphs and Louvain community detection for cell type clustering. Other clustering approaches can also be used if needed. Moreover, not only hard clustering can be used in our model, but soft clustering approaches such as fuzzy clustering, which allows data points to belong to multiple clusters, can also be used.

Our focus in this paper is on cell-type clustering. The joint analysis of both expression and co-expression has a lot of potential applications. This is because the output of the joint analysis is three matrices: one joint structure and two individual structures. We can apply the existing approaches for gene expression analysis to these three matrices. However, transformations need to be considered to make sure the three matrices meet the required assumptions such as normality for each expression analysis approach.

Chapter 4

Modeling Dynamic Correlation in Zero-inflated Bivariate Count Data with Random Effects

4.1 Introduction

Recent single-cell RNA sequencing (scRNA-seq) studies have given insight into individual cells' behavior and function at various stages. For the scRNA-seq data, there are two main analytical approaches: differential expression analysis and differential co-expression analysis. Differential expression analysis considers how the average expression levels of genes vary across different biological conditions and ignores the regulatory mechanisms among different genes. Differential co-expression analysis, on the other hand, takes genetic interactions into consideration and can reveal the intricate underlying gene regulatory mechanisms.

Interactions under different cellular conditions are tightly coordinated and often highly dynamic (Luscombe et al., 2004; de Lichtenberg et al., 2005). In response to various external stimulants and signals, genetic interactions can change flexibly. Therefore, changes in gene co-expression patterns can often be observed (Li, 2002; Li and Yuan, 2004; de la Fuente, 2010a). One statistical measure that can capture these genetic interaction changes was proposed by Li (2002). This measure, referred to as liquid association in his paper, quantifies the relationship where the co-expression between two genes is modulated by a third "coordinator" gene. Li's ground-breaking work has increased the interest in analyzing genetic interactions (Li, 2002; Li et al., 2004; Ho et al., 2007, 2011; Wang et al., 2017; Yu, 2018).

The existing Gaussian-based approaches work well for microarray data but may not fit the count data such as RNA sequencing reads properly. Ma et al. (2020) proposed three Bayesian regression approaches on bivariate count data to model the complex dependency structures. Chapter 2 of this dissertation extended the bivariate Poisson-gamma mixture model in Ma et al. (2020) to a ZEro-inflated Negative binomial dynamic COrrelation (ZENCO) model which accounts for both the over-dispersion and zero-inflation in count data. Therefore, ZENCO can be used to capture the genetic interaction changes in scRNA-seq data. The motivating data in ZENCO is scRNA-seq data from a patient-derived xenograft (PDX) melanoma model (Rambow et al., 2018), so ZENCO assumes homogeneity between cells and doesn't take the subject-specific random effects into consideration. However, cancer scRNA-seq datasets often contain cells from multiple individuals. Tirosh et al. (2016) observed that expression profiles of cells from the same individual tend to cluster together even after data preprocessing and normalization. When analyzing single tumor cells from multiple patients, inter-tumor heterogeneity needs to be accounted for in the analysis. In this chapter, we propose a subject-specific random effects model which is an extension of ZENCO. The proposed model incorporates between-individual random effects to account for the variation between expression measurements of cells collected from different participants.

The remainder of the chapter is arranged as follows. In Section 4.2, the detail of the proposed model is introduced. The simulation studies with different scenarios are conducted in Section 4.3. In Section 4.4, the analysis of scRNA-seq data generated from multiple melanoma tumors is presented. Section 4.5 concludes this chapter with some discussion.

4.2 Methods

Suppose there are s different subjects. For kth subject, there are n_k different cells. Let X_{ikj} represent the count-based expression level of the ith gene ($i = 1, 2$) in the jth cell ($j = 1, 2, ...n_k$) of the kth subject ($k = 1, 2, ..., s$). X_{3kj} represents a third variable that controls the correlation between X_{1kj} and X_{2kj}. The marginal distribution of X_{ikj} ($i = 1, 2$) is modeled as a zero-inflated negative binomial component. The distribution of $\mathbf{X}_{ikj}$ is given by

$$\mathbf{X}_{ikj} \sim \begin{cases} \mathcal{I}_0, & \text{with probability } p_i; \\ NB(\mu_{ik}, \phi_i), & \text{with probability } 1 - p_i. \end{cases} \tag{4.1}$$

μ_{ik} are random effects with hyperparameters μ_i and σ_{μ_i},

$$\mu_{ik} \sim N(\mu_i, \sigma_{\mu_i}), k = 1, 2, ..., s \tag{4.2}$$

Following the same logic as the ZENCO model, we construct our conditional bivariate negative binomial model through a Poisson-Gamma mixture setting. For $i = 1, 2$, $j = 1, 2, ..., n_k$ and $k = 1, 2, ..., s$, let

$$X_{ikj} \sim Poisson(u_{ikj}\mu_{ik}), u_{ikj} \sim Gamma(\alpha_i, \alpha_i). \tag{4.3}$$

A negative binomial distribution of $NB(\mu_{ik}, \frac{1}{\alpha_i})$ can be generated by integrating over u_{ikj}. The conditional correlation between X_{1kj} and X_{2kj} given X_{3kj} is from a latent variable $Z_{kj} = (Z_{1kj}, Z_{2kj})$.

The correlation, ρ_{kj}, of (Z_{1kj}, Z_{2kj}) is specified as

$$\log(\frac{1 + \rho_{kj}}{1 - \rho_{kj}}) = \tau_{0k} + \tau_{1k} X_{3kj}. \tag{4.4}$$

$\log(\frac{1+\rho_{kj}}{1-\rho_{kj}})$ is the Fisher's Z-transformation for the correlation ρ_{kj}, τ_{0k} and τ_{1k} follows subject-specific random effects model with mean equal to τ_0 and τ_1 respectively.

$$\begin{aligned} \tau_{0k} &\sim N(\tau_0, \sigma_{\tau_0}), k = 1, 2...s \\ \tau_{1k} &\sim N(\tau_1, \sigma_{\tau_1}), k = 1, 2...s \end{aligned} \tag{4.5}$$

Let

$$X_{ikj} \sim Poisson[F_{\alpha_i}^{-1}\{\Phi(Z_{ikj})\}\mu_{ik}], \tag{4.6}$$

where $F_{\alpha_i}(\cdot)$ is the cumulative distribution function of a $Gamma(\alpha_i, \alpha_i)$ distribution with $\alpha_i = 1/\phi_i$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. $F_{\alpha_i}^{-1}$ maps each point in the interval (0,1) to $Gamma(\alpha_i, \alpha_i)$ distribution. Hence, the distribution of $F_{\alpha_i}^{-1}\{\Phi(Z_{ikj})\}$ is $Gamma(\alpha_i, \alpha_i)$. The distribution of $X_{ikj} \sim Poisson[F_{\alpha_i}^{-1}\{\Phi(Z_{ikj})\}\mu_{ik}]$ is then a Poisson-Gamma mixture distribution, which follows the negative binomial density $NB(\mu_{ik}, \phi_i = \frac{1}{\alpha_i})$.

The sampling scheme during each MCMC iteration is as follows. For $j = 1, 2, ..., n_k$, $i = 1, 2$, and $k = 1, 2, ..., s$, we sample μ_{ik} from

$$f(\mu_{ik}|\cdot) \propto f(\mu_{ik}) \prod_{j=1}^{n_k} f(x_{ikj}|\mu_{ik}, \phi_i),$$

where $f(x_{ikj}|\mu_{ik}, \phi_i)$ is the probability density function of

$$x_{ikj} \sim \begin{cases} \mathcal{I}_0, & \text{with probability } p_i; \\ NB(\mu_{ik}, \phi_i), & \text{with probability } 1 - p_i. \end{cases}$$

The hyperparameters μ_i and σ_{μ_i} can be sampled from

$$f(\mu_i|\cdot) \propto f(\mu_i) \prod_{k=1}^{s} f(\mu_{ik}|\mu_i, \sigma_{\mu_i}),$$
$$f(\sigma_{\mu_i}|\cdot) \propto f(\sigma_{\mu_i}) \prod_{k=1}^{s} f(\mu_{ik}|\mu_i, \sigma_{\mu_i}).$$

We then sample ϕ_i from

$$f(1/\phi_i|\cdot) \propto f(1/\phi_i) \prod_{k=1}^{s} \prod_{j=1}^{n_k} f(x_{ikj}|\mu_{ik}, \phi_i),$$

Next, we sample τ_{0k} from

$$f(\tau_{0k}|\cdot) \propto f(\tau_{0k}) \prod_{j=1}^{n_k} f(\mathbf{z_{kj}}|\tau_{0k}, \tau_{1k}, x_{3kj}),$$

and sample τ_{1k} from

$$f(\tau_{1k}|\cdot) \propto f(\tau_{1k}) \prod_{j=1}^{n_k} f(\mathbf{z_{kj}}|\tau_{0k}, \tau_{1k}, x_{3kj}),$$

where $f(\mathbf{z_{kj}}|\tau_{0k}, \tau_{1k}, x_{3kj}) = N_2\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \frac{e^{(\tau_{0k}+\tau_{1k}\times x_{3jk})}-1}{e^{(\tau_{0k}+\tau_{1k}\times x_{3jk})}+1} \\ \frac{e^{(\tau_{0k}+\tau_{1k}\times x_{3jk})}-1}{e^{(\tau_{0k}+\tau_{1k}\times x_{3jk})}+1} & 1 \end{bmatrix} \Big)$.

The hyperparameters τ_0, σ_{τ_0}, τ_1, and σ_{τ_1} can be sampled from

$$f(\tau_0|\cdot) \propto f(\tau_0) \prod_{k=1}^{s} f(\tau_{0k}|\tau_0, \sigma_{\tau_0}),$$
$$f(\tau_1|\cdot) \propto f(\tau_1) \prod_{k=1}^{s} f(\tau_{1k}|\tau_1, \sigma_{\tau_1}),$$
$$f(\sigma_{\tau_0}|\cdot) \propto f(\sigma_{\tau_0}) \prod_{k=1}^{s} f(\sigma_{\tau_0}|\tau_0, \sigma_{\tau_0}),$$
$$f(\sigma_{\tau_1}|\cdot) \propto f(\sigma_{\tau_1}) \prod_{k=1}^{s} f(\sigma_{\tau_1}|\tau_1, \sigma_{\tau_1}).$$

In addition, z_{ikj} can be sampled from

$$f(z_{ikj}|\cdot) \propto f(x_{ikj}|z_{ikj}, \mu_{ik}, \alpha_i) f(z_{ikj}|z_{mkj}),\ i, m = 1, 2;\ i \neq m,$$

where $f(z_{ikj}|z_{mkj}) = N(\rho_{kj} z_{mkj}, (1 - {\rho_{kj}}^2))$.

4.3 Simulation

4.3.1 Scenario 1: Simulating data from proposed model

In this first simulation, we demonstrate generating data from the proposed model. The simulated data contain count-based expression levels of genes $\mathbf{X}_1$, $\mathbf{X}_2$ regulated by $\mathbf{X}_3$ from two types of subjects. For instance, subject 1 can come from an immune-resistance tumor and subject 2 can come from an immune-sensitive tumor (Jerby-Arnon et al., 2018). $\mathbf{X}_3$, in this case, is the immune resistance score.

Using the simulation approach described in ZENCO, we generated 10^5 observations from the proposed model for each subject with subject-specific τ_1 and plotted a panel of conditional distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$ given various levels of $\mathbf{X}_3$ in Figure 4.1. In this simulation study, τ_1 in subject 1 is zero, while τ_1 in subject 2 is not zero. The first row is for subject 1 and the second row is for subject 2. In these figures, we observed that different subjects will have different correlation patterns between

$\mathbf{X}_1$ and $\mathbf{X}_2$ when τ_1 has subject-specific random effects. It's necessary to account for random effects for co-expression levels of gene pairs if we are interested in the dynamic correlation changes.

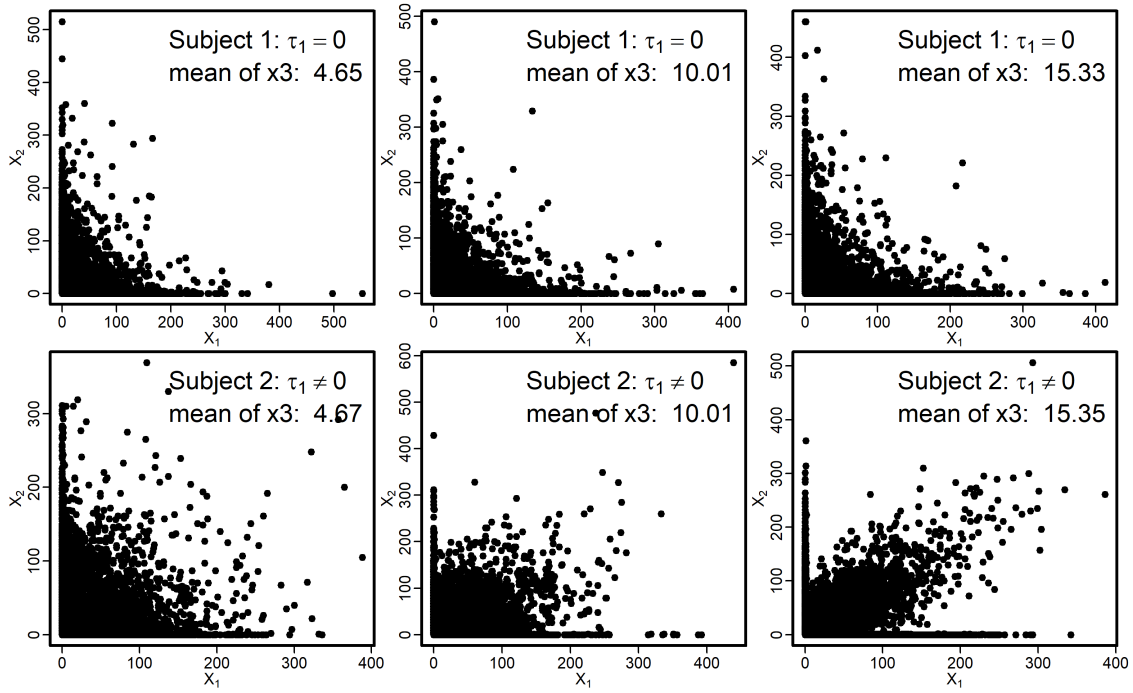

Figure 4.1. Profile plots of $(\mathbf{X}_1, \mathbf{X}_2|\mathbf{X}_3)$ with varying τ_1 ($\tau_1 = 0$ vs $\tau_1 \neq 0$)

4.3.2 Scenario 2: Estimation efficiency

To evaluate the estimation efficiency of our proposed model, we simulated 20 subjects, each with 100 cells. The simulated data contain the level of three variables: $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, and the indicator variable for different subjects $\mathbf{s}$. Therefore, the data have 2,000 rows and 4 columns. The correlations of $\mathbf{X}_1$ and $\mathbf{X}_2$ are controlled by the level of $\mathbf{X}_3$. This simulation was conducted as follows.

For each subject, a subject-specific τ_{0k} was sampled from $N(0, 0.01)$. Similarly, a subject-specific τ_{1k} was sampled from $N(0, 0.0001)$. Different values of τ_{0k} and τ_{1k} are correspond to different subjects. We then used the similar simulation strategy we

used in Chapter 2 to simulate 100 cells for each subject. First, we simulated a set of $\{x_{3kj}\}_{j=1}^{100}$ from a normal distribution with mean 0 and variance 1. We then calculated correlation coefficient $\rho_{kj} = \frac{e^{(\tau_{0k}+\tau_{1k}\times x_{3kj})}-1}{e^{(\tau_{0k}+\tau_{1k}\times x_{3kj})}+1}$ for each x_{3kj}. The latent variables $\mathbf{z}_{kj} = (z_{1kj}, z_{2kj})'$ were generated such that $\mathbf{z}_{kj} \sim N_2\Big(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho_{kj}\\ \rho_{kj} & 1\end{bmatrix}\Big)$ and simulated x_{1kj} and x_{2kj} using $\mathbf{z}_{kj}$. For each subject, data with 100 cells were generated. Since there are 20 subjects in total, the simulated data have 2,000 cells. The indicator variable for different subjects was added as the fourth column of the simulated data. We then evaluated the estimation efficiency of the proposed random effects model and compared the random effects model with a model without considering random effects. We reported 95% empirical coverage probabilities, mean squared errors (MSE), and mean bias errors (MBE) in Table 4.1 and Table 4.2 for both models. We can observe that when the data have random effects, our proposed model outperforms the model without considering random effects. The coverage probability of 95% credible intervals in the model without subject-specific random effects is much lower than in the proposed model. Also, the MSE and MBE are much higher in the model without subject-specific random effects.

Table 4.1. Coverage probability of 95% credible intervals (CIs), interval lengths based on 1,000 MCMC simulations ($\tau_0 = 0$, $\tau_1 = 0.3$, 20 subjects, 100 cells per subject).

	Proposed model		Model without random effects	
Parameter	Coverage probability	CI length	Coverage probability	CI length
μ_1	0.953	2.707	0.938	2.478
μ_2	0.948	2.705	0.931	2.478
ϕ_1	0.956	0.449	0.955	0.450
ϕ_2	0.940	0.449	0.939	0.448
τ_0	0.977	0.473	0.993	0.286
τ_1	0.972	0.473	0.379	0.288

Table 4.2. Mean square errors (MSE), and mean bias errors (MBE) based on 1000 MCMC simulations ($\tau_0 = 0$, $\tau_1 = 0.3$, 20 subjects, 100 cells per subject).

	Proposed model		Model without random effects	
Parameter	MSE	MBE	MSE	MBE
μ_1	0.454	-0.082	0.456	-0.102
μ_2	0.482	-0.027	0.479	-0.048
ϕ_1	0.013	0.003	0.013	-0.003
ϕ_2	0.013	0.000	0.014	-0.005
τ_0	0.012	-0.006	0.003	-0.001
τ_1	0.012	-0.009	0.029	-0.161

4.3.3 Scenario 3: Comparisons to existing approaches

To evaluate the performance of our proposed random effects model, we performed power analysis and compare the model to two other existing approaches. For testing the existence of dynamic co-expression changes, our hypotheses are set up as:

$$H_0 : \tau_1 = 0 \text{ versus } H_1 : \tau_1 \neq 0.$$

First, we compared the proposed model to a model without considering random effects. The statistical power of this method can be calculated as the percentage of times that the posterior 95% credible intervals of τ_1 do not cover zero. The random effects model and the model without random effects were both carried out using the MCMC algorithm with 20,000 iterations, and 10,000 burn-ins.

The third method is to fit the negative binomial count data with the conditional normal model (CNM-Full). The parameters in CNM-Full are estimated using generalized estimation equations (GEE)-based approach. The powers of CNM-Full can be calculated by counting the percentage of times when p-values associated with τ_1 are less than 0.05.

We simulated 20 subjects each with 100 observations from the proposed model by letting $\phi_1 = \phi_2 = 2$, μ_1, $\mu_2 \sim N(30, 1)$ and $\tau_0 \sim N(0, 0.01)$, and then varied τ_1 values and performed power analyses. The simulated values of $\mu_1, \mu_2, \phi_1, \phi_2$ are based on the estimates obtained from the real data analysis. Figure 4.2 shows the power curves of

the three methods. We observed that both the random effects model and the CNM-Full outperform the model without random effect. When τ_1 is greater than 0.25, our proposed method outperforms the other two methods. When τ_1 is smaller than 0.25, CNM-Full performs slightly better than our random effects model. This is because Bayesian tests tend to be more conservative than the frequentist tests (Gelman and Tuerlinckx, 2000). Also, GEE is robust to model misspecification (Hubbard et al., 2010). However, GEE is not able to present subject-random effects. In this power analysis, only the mean of τ_1 is calculated. Oftentimes, in real data analysis, we also need the subject-specific random effects of τ_1. In addition, Chapter 2 has shown that when X_3 has dropouts, the performance of CNM-Full will decline drastically.

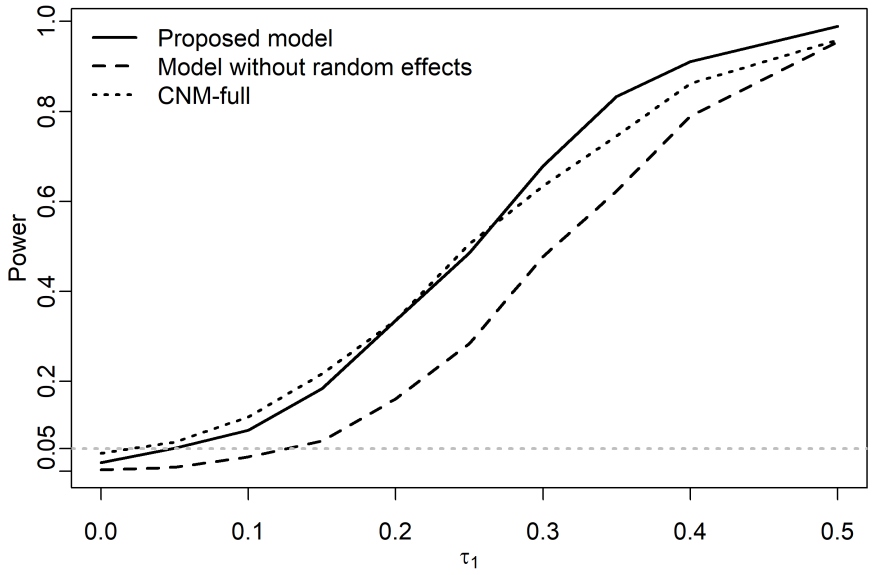

Figure 4.2. Power curves comparing model without random effect and CNM-Full model.

4.4 Experimental Data Analysis

We used the proposed model to analyze the melanoma tumor data described in Jerby-Arnon et al. (2018). The data were obtained from Gene Expression Omnibus (GEO accession number: GSE115978). The data set consists of 7186 cells from 33 melanoma

tumors comprised of 2,987 cells from 17 newly collected patient tumors and 4,199 cells from 16 patient tumors that Tirosh et al. (2016) previously reported. The cells were annotated as different cell subsets such as malignant cells, CD8+ and CD4+ T cells, B cells, natural killer (NK) cells, macrophages, cancer-associated fibroblasts (CAFs), and endothelial cells based on expression profiles and inferred copy-number variation (CNV) profiles. In our data analysis, we focused on malignant cells and only used the tumors with at least 70 malignant cells.

To study the immunotherapy resistance in melanoma tumors, Jerby-Arnon et al. (2018) identified a malignant-cell exclusion program consisting of genes induced or repressed by malignant cells in "cold" (high-resistance to immunotherapy) versus "hot" (low-resistance to immunotherapy) tumors. An overall expression (OE) score of gene sets was defined in their paper to quantify the level of resistance to immunotherapy for each cell or tumor. In our study, we scored each cell with its OE of gene set with genes induced by malignant cells in the resistance program. The OE score was used as X_3 in this analysis. We then conducted the pairwise analysis for genes in the melanoma pathway described in the KEGG database (Kanehisa and Goto, 2000).

To identify the gene pairs that are significantly associated with immunotherapy resistance, we implemented the proposed model and considered tumor heterogeneity by estimating τ_1 and random effects of τ_1. For most gene pairs, the overall τ_1 estimates are not significant. This is because τ_1 has subject-specific random effects. In some tumors, τ_1 estimates are significant, while in other tumors, τ_1 are not significant. Even if τ_1 estimates in different tumors are all significant, the estimates can have different signs. Therefore, it's important and necessary to show the random effects of τ_1. The gene pairs with significant overall τ_1 estimates and standard deviation of τ_1 random effects are shown in Table 4.3. Figure 4.3 shows the estimates of the random effects for τ_1 in those gene pairs. Different colors in Figure 4.3 represent different tumors. We can see there are tumor-specific random effects of τ_1 and our method

successfully presented the random effects of τ_1. Positive τ_1 means the gene pair tends to be correlated in cold tumors, while negative τ_1 means the gene pair tends to be correlated in hot tumors. Most gene pairs in Table 4.3 and Figure 4.3 have negative τ_1 estimates, which means those gene pairs are more correlated in low-resistance to immunotherapy tumors.

Table 4.3. Table of significant τ_1 estimates and standard deviation (SD) of τ_1 random effects in experimental data analysis

#	Gene1	Gene2	τ_1	τ_1 random effects SD
1	PIK3CD	KRAS	-0.892(-1.856,-0.029)	0.248
2	PIK3CD	BAD	-0.893(-1.804,-0.240)	0.158
3	PIK3CD	CDK4	-0.904(-1.561,-0.285)	0.403
4	PIK3CD	CDK6	-0.708(-1.362,-0.046)	0.105
5	AKT1	FGF5	-0.686(-1.213,-0.128)	0.120
6	KRAS	CDK4	-0.893(-1.767,-0.268)	0.306
7	CDK4	POLK	-0.904(-1.727,-0.148)	0.263
8	CDK4	FGF5	-1.092(-1.714,-0.452)	0.231
9	BAK1	FGFR1	-1.792(-3.247,-0.462)	0.132
10	POLK	CDK6	-0.635(-1.380,-0.049)	0.082
11	FGF5	FGFR1	1.203(0.072, 2.439)	0.295

4.5 Discussion

In this chapter, we presented a random effects model for studying changes of conditional correlation in scRNA-seq data with subject-specific random effects. In our model, the correlation of two genes is regulated by the third variable. We assume the third variable is continuous and normally distributed because the motivating data of this study uses continuous scores as X_3.

The third variable can also be the expression level of the third gene, which means X_3 follows a zero-inflated negative binomial distribution. Although this is a different setting that we didn't cover in this chapter, we developed two approaches: Gaussian copula with random effects and ZENCO with random effects. The results of simulation studies for these two approaches are provided in Appendix A.

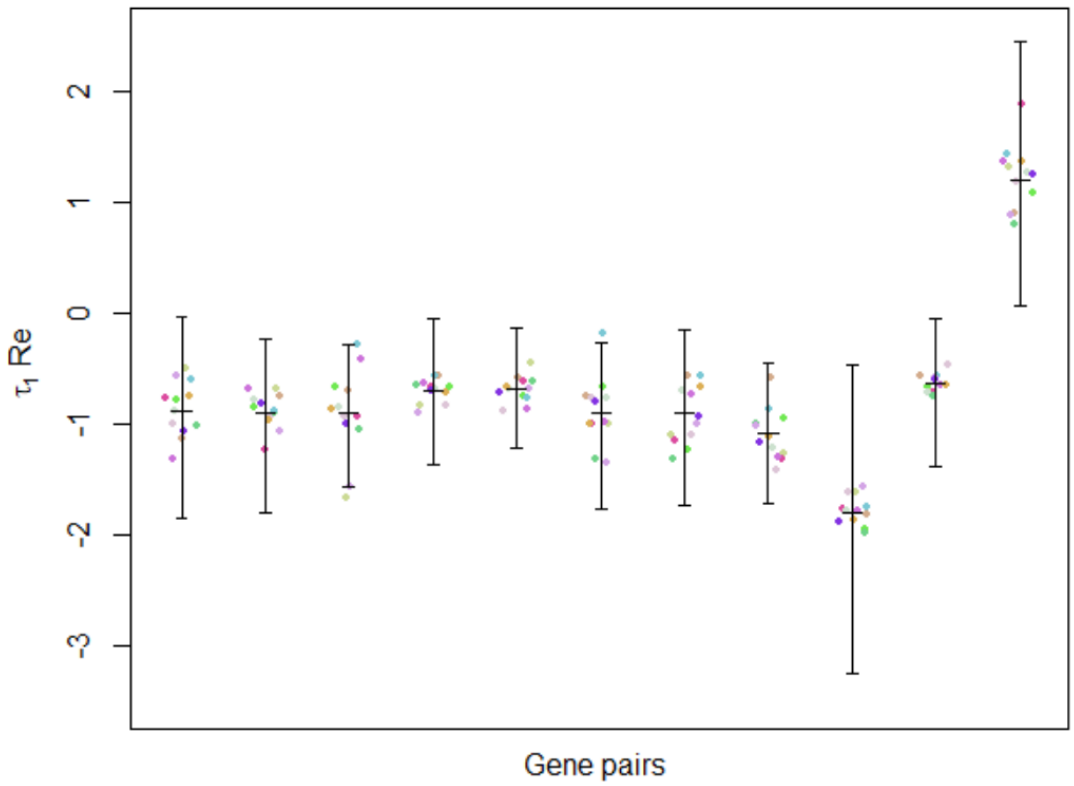

Figure 4.3. Random effects estimates for τ_1 for gene pairs in Table 4.3. Different color represents different tumors.

In this study, we assume the dispersion parameters ϕ_1 and ϕ_2 don't have subject-specific random effects, because the estimates from real data show ϕ_1 and ϕ_2 don't have strong random effects. This makes sense, because small changes in dispersion parameters may lead to big changes in variance.

Our proposed model outperforms the model without considering subject-specific random effects when the data have random effects. In practical applications, we usually don't know if the data have strong subject-specific random effects or not. We can compute log-pseudo-marginal likelihood (LPML) for models with or without random effects and do model selection based on two models' LPMLs.

Bibliography

Ai, D., Li, X., Pan, H., Chen, J., Cram, J. A., and Xia, L. C. (2019). Explore mediated co-varying dynamics in microbial community using integrated local similarity and liquid association analysis. *BMC Genomics*, 20(2):185.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, 17(1):63.

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature communications*, 12(1):1–12.

Czyz, M. (2019). Fibroblast growth factor receptor signaling in skin cancers. *Cells*, 8(6):540.

Dawson, J. A. and Kendziorski, C. (2012). An empirical Bayesian approach for identifying differential coexpression in high-throughput experiments. *Biometrics*, 68(2):455–465.

de la Fuente, A. (2010a). From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends in Genetics : TIG*, 26:326–333.

de la Fuente, A. (2010b). From 'differential expression'to 'differential networking'– identification of dysfunctional regulatory networks in diseases. *Trends in genetics*, 26(7):326–333.

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2019). Multi-study factor analysis. *Biometrics*, 75(1):337–346.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8.

Fultang, N., Chakraborty, M., and Peethambaran, B. (2021). Regulation of cancer stem cells in triple negative breast cancer. *Cancer Drug Resistance*, 4(2):321–342.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Gelman, A., Roberts, G. O., Gilks, W. R., et al. (1996). Efficient metropolis jumping rules. *Bayesian statistics*, 5(599-608):42.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gelman, A. and Tuerlinckx, F. (2000). Type s error rates for classical and bayesian single and multiple comparison procedures. *Computational statistics*, 15(3):373–390.

Gunderson, T. and Ho, Y.-Y. (2014). An efficient algorithm to explore liquid association on a genome-wide scale. *BMC Bioinformatics*, 15(1):371.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli*, pages 223–242.

Ho, Y.-Y., Cope, L., Dettling, M., and Parmigiani, G. (2007). Statistical methods for identifying differentially expressed gene combinations. *Methods in Molecular Biology*, 408:171–191.

Ho, Y.-Y., Cope, L. M., and Parmigiani, G. (2014). Modular network construction using eQTL data: an analysis of computational costs and benefits. *Frontiers in Genetics*, 5:40.

Ho, Y.-Y., Parmigiani, G., Louis, T. A., and Cope, L. M. (2011). Modeling liquid association. *Biometrics*, 67(1):133–141.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Satariano, S. A., Jewell, N., Bruckner, T., and Satariano, W. A. (2010). To gee or not to gee: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, pages 467–474.

Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, 50(8):1–14.

Jerby-Arnon, L., Shah, P., Cuoco, M. S., Rodman, C., Su, M.-J., Melms, J. C., Leeson, R., Kanodia, A., Mei, S., Lin, J.-R., et al. (2018). A cancer cell program promotes t cell exclusion and resistance to checkpoint blockade. *Cell*, 175(4):984–997.

Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008). Posterior error probabilities and false discovery rates: two sides of the same coin. *Journal of Proteome Research*, 7(01):40–44.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30. accessed on 02/24/19.

Karlis, D. and Meligkotsidou, L. (2005). Multivariate poisson regression with covariance structure. *Statistics and Computing*, 15(4):255–265.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740.

Khayer, N., Marashi, S.-A., Mirzaie, M., and Goshadrou, F. (2017). Three-way interaction model to trace the mechanisms involved in Alzheimer's disease transgenic mice. *PLoS One*, 12(9).

Kinzy, T. G., Starr, T. K., Tseng, G. C., and Ho, Y.-Y. (2019). Meta-analytic framework for modeling genetic coexpression dynamics. *Statistical Applications in Genetics and Molecular Biology*, 18(1):1–12.

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282.

Kong, Y. and Yu, T. (2019). A hypergraph-based method for large-scale dynamic correlation study at the transcriptomic scale. *BMC Genomics*, 20(1):397.

Lai, Y., Wu, B., Chen, L., and Zhao, H. (2004). A statistical method for identifying differential gene–gene co-expression patterns. *Bioinformatics*, 20(17):3146–3155.

Li, K.-C. (2002). Genome-wide coexpression dynamics: theory and application. *Proceedings of the National Academy of Sciences*, 99(26):16875–16880.

Li, K.-C., Liu, C.-T., Sun, W., Yuan, S., and Yu, T. (2004). A system for enhancing genome-wide coexpression dynamics study. *Proceedings of the National Academy of Sciences*, 101(44):15561–15566.

Li, K.-C. and Yuan, S. (2004). A functional genomic study on NCI's anticancer drug screen. *The Pharmacogenomics Journal*, 4(2):127–135.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523.

Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):75.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312.

Ma, S., Gong, Q., and Bohnert, H. J. (2007). An *Arabidopsis* gene network based on the graphical Gaussian model. *Genome Research*, 17:1614–1625.

Ma, Z., Hanson, T. E., and Ho, Y.-Y. (2020). Flexible bivariate correlated count data regression. *Statistics in Medicine*, 39(25):3476–3490.

McCarthy, D. J., Campbell, K. R., Lun, A. T., and Wills, Q. F. (2017). Scater: preprocessing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186.

Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, 34(18):3223–3224.

Pierson, E. and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):1–10.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*.

Rambow, F., Rogiers, A., Marin-Bejar, O., Aibar, S., Femel, J., Dewaele, M., Karras, P., Brown, D., Chang, Y. H., Debiec-Rychter, M., et al. (2018). Toward minimal residual disease-directed therapy in melanoma. *Cell*, 174(4):843–855.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Solis-Trapala, I. L. and Farewell, V. T. (2005). Regression analysis of overdispersed correlated count data with subject specific covariates. *Statistics in Medicine*, 24(16):2557–2575.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196.

Van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J. P. (2018). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592.

van Iterson, M., Boer, J. M., and Menezes, R. X. (2010). Filtering, FDR and power. *BMC Bioinformatics*, 11(1):450.

Villanueva, J., Vultur, A., and Herlyn, M. (2011). Resistance to BRAF inhibitors: unraveling mechanisms and future treatment options. *Cancer Research*, 71(23):7137–7140.

Wang, L., Liu, S., Ding, Y., Yuan, S., Ho, Y.-Y., and Tseng, G. C. (2017). Meta-analytic framework for liquid association. *Bioinformatics*, 33(14):2140–2147.

Wang, L., Zheng, W., Zhao, H., and Deng, M. (2013). Statistical analysis reveals co-expression patterns of many pairs of genes in yeast are jointly regulated by interacting loci. *PLoS Genetics*, 9(3):e1003414.

Wen, X., Gao, L., and Hu, Y. (2020). LAcemodule: Identification of competing endogenous RNA modules by integrating dynamic correlation. *Frontiers in Genetics*, 11:235.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5.

Xu, X., Wang, M., Li, L., Che, R., Li, P., Pei, L., and Li, H. (2017). Genome-wide trait-trait dynamics correlation study dissects the gene regulation pattern in maize kernels. *BMC Plant Biology*, 17(1):163.

Yan, J. and Fine, J. (2004). Estimating equations for association structures. *Statistics in Medicine*, 23(6):859–874.

Yeo, S. K., Zhu, X., Okamoto, T., Hao, M., Wang, C., Lu, P., Lu, L. J., and Guan, J.-L. (2020). Single-cell rna-sequencing reveals distinct patterns of cell state heterogeneity in mouse models of breast cancer. *Elife*, 9:e58810.

Yu, T. (2018). A new dynamic correlation algorithm reveals novel functional aspects in single cell and bulk RNA-seq data. *PLoS Computational Biology*, 14(8):e1006391.

Zhang, J., Ji, Y., and Zhang, L. (2007). Extracting three-way gene interactions from microarray data. *Bioinformatics*, 23(21):2903–2909.

Appendix A

Random effects approaches when $\mathbf{X}_3$ is zero-inflated count data

In the setting of Chapter 2, the third variable $\mathbf{X}_3$ is zero-inflated count data. However, in Chapter 4, to fit the setting of data set we used, we assume the third variable $\mathbf{X}_3$ follows a normal distribution. To model the dynamic correlation in zero-inflated bivariate count data when $\mathbf{X}_3$ is also zero-inflated and count-based, more works are needed to be done. In this Appendix, we briefly introduce two approaches that can be used when $\mathbf{X}_3$ is zero-inflated count data.

A.1 ZENCO with Random Effects

The first approach is an direct extension of ZENCO in Chapter 2. Suppose there are s different subjects. For kth subject, there are n_k different cells. Let X_{ikj} represent the gene expression level of the ith gene ($i = 1, 2, 3$) in the jth cell ($j = 1, 2, ...n_k$) of the kth subject($k = 1, 2, ..., s$). The correlation, ρ_{kj}, of (Z_{1kj}, Z_{2kj}) is specified as

$$\log(\frac{1+\rho_{kj}}{1-\rho_{kj}}) = \tau_{0k} + \tau_1 X_{3kj}. \tag{A.1}$$

$\log(\frac{1+\rho_{kj}}{1-\rho_{kj}})$ is the Fisher's Z-transformation for the correlation ρ_{kj}.

The distribution of τ_{0k} is as follow:

$$\tau_{0k} \sim N(\tau_0, \sigma_{\tau_0}), k = 1, 2...s \tag{A.2}$$

We simulated 50 subjects, each with 100 cells. The simulated data contain the expression level of three genes: $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, and the indicator variable for different

subjects. The simulated data have 5,000 rows and 4 columns. This simulation was conducted as follows.

For each subject, a subject-specific τ_{0k} was sampled from $N(0, 1)$. Different values of τ_{0k} are correspond to different subjects. We then used the same simulation strategy we used in ZENCO paper to simulate 100 cells for each subject. Since there are 50 subjects in total, the simulated data have 5,000 cells. The indicator variable for different subjects was added as the forth column of the simulated data. The mean squared errors (MSE), mean bias errors (MBE), and 95% empirical coverage probabilities are presented in Table A.1.

Table A.1. Coverage probability of 95% credible intervals (CIs), interval lengths, Mean square errors (MSE), and mean bias errors (MBE) based on 1,000 MCMC simulations ($\tau_0 = 0$, $\tau_1 = 0.05$, 50 participants, 100 cells per participant) using ZENCO with random effects.

Parameter	Coverage probability	CI length	MSE	MBE
μ_1	0.946	2.356	0.366	-0.057
μ_2	0.944	2.356	0.382	-0.075
ϕ_1	0.943	0.672	0.030	-0.007
ϕ_2	0.951	0.667	0.028	-0.003
τ_0	0.960	1.054	0.066	-0.011
τ_1	0.951	0.044	0.000	0.000

Moreover, according to Gelman (2006), we can also implement a redundant multiplicative re-parameterization to reduce autocorrelation and improve Markov Chain Monte Carlo (MCMC) convergence:

$$\tau_{0k} \sim N(\tau_0 + \xi\eta_k, \sigma_{\tau_0}), k = 1, 2...s \tag{A.3}$$

$$\xi \sim N(0, \tau_\xi)$$

$$\eta_k \sim N(0, \tau_\eta)$$

$$\tau_\eta \sim Gamma(0.5, 0.5)$$

In this setting, the variance of the random effect $\xi\eta_k$ will follow a half-Cauchy distribution (Gelman, 2006).

A.2 Copula Model with Random Effects

The second model is based on the Copula model in Ma et al. (2020). The MCMC sampling scheme is described as follows:

1. Initiate $\boldsymbol{\eta}_i = (\mu_i, \phi_i)$, τ_0, σ_{τ_0}, τ_1, $\boldsymbol{\tau_{0k}}$, and $\{z_{kj}\}_{k=1}^{s}$, $j = 1, \ldots, n_k$. Set $t = 1$.

2. For $i = 1, 2$ and $j = 1, \ldots, n_k$,

   a) Calculate

$$f(z_{ikj} \mid z_{(-i)kj}, else) = \Phi\left[\frac{\Phi^{-1}[F_i(x_{ikj}; \boldsymbol{\eta}_i)] - \rho_{kj} z_{(-i)kj}}{\sqrt{1-\rho_{kj}^2}}\right] - \Phi\left[\frac{\Phi^{-1}[F_i(x_{ikj}-1; \boldsymbol{\eta}_i)] - \rho_{kj} z_{(-i)kj}}{\sqrt{1-\rho_{kj}^2}}\right];$$

   b) Sample

$$z_{ikj} \sim N\left(\rho_{kj} z_{(-i)kj},\ 1-\rho_{kj}^2\right)$$

   truncated at $(\Phi^{-1}[F_i(x_{ikj}-1; \boldsymbol{\eta}_i)], \Phi^{-1}[F_i(x_{ikj}; \boldsymbol{\eta}_i)])$;

   c) Sample $\boldsymbol{\eta}_i^{(t)}$ from

$$f(\boldsymbol{\eta}_i|\cdot) \propto \left[\prod_{k=1}^{s}\prod_{j=1}^{n_k} f(z_{ikj} \mid z_{(-i)kj}, else)\right] f(\boldsymbol{\eta}_i),$$

   where $f(\boldsymbol{\eta}_i)$ represents the prior distributions of $\boldsymbol{\eta}_i$.

3. Sample $(\tau_1, \boldsymbol{\tau_{0k}})^{(t)}$ from

$$f(\tau_1, \boldsymbol{\tau_{0k}}|\cdot) \propto \left[\prod_{k=1}^{n_k}\prod_{j=1}^{n_c} f(\mathbf{z}_{kj}|\tau_1, \boldsymbol{\tau_{0k}})\right] f(\tau_1)\, f(\boldsymbol{\tau_{0k}}),$$

4. Sample $(\tau_0, \sigma_{\tau_0})^{(t)}$ from

$$f(\tau_0, \sigma_{\tau_0}|\boldsymbol{\tau_{0k}}) \propto f(\boldsymbol{\tau_{0k}}|\tau_0, \sigma_{\tau_0})\, f(\tau_0)\, f(\sigma_{\tau_0})$$

5. Set $t = t + 1$ and return to Step 2.

A random-walk adaptive Metropolis algorithm (Haario et al., 2001) was used in the above sampling scheme. In general, for a d-dimensional vector $\boldsymbol{\theta}^*$, to sample $\boldsymbol{\theta}^* \sim \text{MVN}(\boldsymbol{\theta}^{(t-1)}, C_t)$ where

$$C_t = \begin{cases} C_0 & t \leq t_0, \\ s_d \cdot \left[\text{cov}(\boldsymbol{\theta}^{(t-h)}, \ldots, \boldsymbol{\theta}^{(t-1)}) + \epsilon \cdot \mathbf{I}_d\right] & t > t_0, \end{cases} \tag{A.4}$$

we accept $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ with probability

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^*|\cdot)}{f(\boldsymbol{\theta}^{(t-1)}|\cdot)} \cdot \frac{\phi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*, C_t)}{\phi(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}, C_t)}\right\}. \tag{A.5}$$

We tuned s_d and ϵ to make the acceptance rate around 23%. Gelman et al. (1996) recommend to set $s_d = 2.4^2/d$ and $\epsilon = 0.005$ to make sure that the acceptance rate is optimal. In step 3, we want to sample τ_1 and $\boldsymbol{\tau_{0k}}$ simultaneously. If we use the recommended s_d and ϵ for both of them to make sure the overall acceptance rate is optimal, we can see their trace plots don't show good mixing. Therefore, we tuned s_d and ϵ for τ_1 and $\boldsymbol{\tau_{0k}}$ separately.

Using same simulation setting in ZENCO with random effects, we applied the copula model with random effects and presented 95% empirical coverage probabilities, MSE and MBE in Table A.2.

Table A.2. Coverage probability of 95% credible intervals (CIs), interval lengths, Mean square errors (MSE), and mean bias errors (MBE) based on 1,000 MCMC simulations ($\tau_0 = 0$, $\tau_1 = 0.05$, 20 participants, 500 cells per participants) using Copula model with random effects.

Parameter	Coverage probability	CI length	MSE	MBE
μ_1	0.962	0.422	0.011	-0.002
μ_2	0.959	0.422	0.011	-0.005
τ_0	0.956	0.963	0.052	0.007
τ_1	0.872	0.011	0.000	0.000
τ_y	0.959	1.436	0.132	0.003