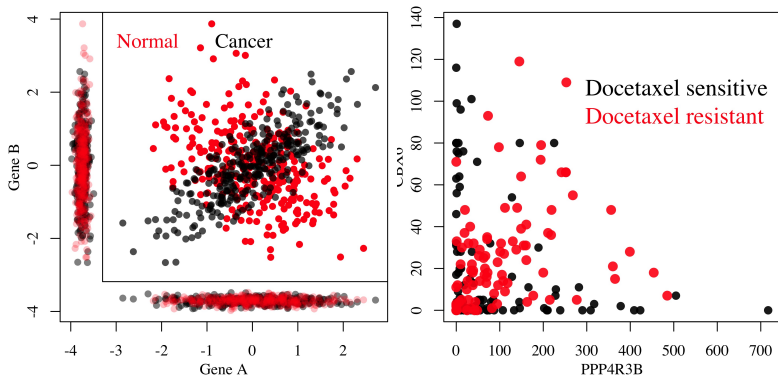


Models for Correlated Data

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Example: Differential Coexpression



Genetic molecules and gene products participate in complex inter-connected pathways in biological systems.

Modeling Liquid Association

Yen-Yi Ho,^{1,*} Giovanni Parmigiani,² Thomas A. Louis,³ and Leslie M. Cope^{4,**}

¹McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

²Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.

³Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

⁴The Sidney Kimmel Cancer Center, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.


**email:* yho@jhsph.edu

***email:* lcope1@jhmi.edu

SUMMARY. In 2002, Ker–Chau Li introduced the liquid association measure to characterize three-way interactions between genes, and developed a computationally efficient estimator that can be used to screen gene expression microarray data for such interactions. That study, and others published since then, have established the biological validity of the method, and clearly demonstrated it to be a useful tool for the analysis of genomic data sets. To build on this work, we have sought a parametric family of multivariate distributions with the flexibility to model the full range of trivariate dependencies encompassed by liquid association. Such a model could situate liquid association within a formal inferential theory. In this article, we describe such a family of distributions, a trivariate, conditional normal model having Gaussian univariate marginal distributions, and in fact including the trivariate Gaussian family as a special case. Perhaps the most interesting feature of the distribution is that the parameterization naturally parses the three-way dependence structure into a number of distinct, interpretable components. One of these components is very closely aligned to liquid association, and is developed as a measure we call modified liquid association. We develop two methods for estimating this quantity, and propose statistical tests for the existence of this type of dependence. We evaluate these inferential methods in a set of simulations and illustrate their use in the analysis of publicly available experimental data.

KEY WORDS: Gene expression; Generalized estimating equations; Higher-order interaction; Liquid association; Non-Gaussian multivariate distribution.

Flexible bivariate correlated count data regression

Zichen Ma¹  | Timothy E. Hanson² | Yen-Yi Ho¹

¹Department of Statistics, University of South Carolina, Columbia, South Carolina

²Medtronic Inc., Minneapolis, Minnesota

Correspondence

Zichen Ma, Department of Statistics, 127C LeConte College, University of South Carolina, Columbia, SC 29208.
Email: zichen@email.sc.edu

Multivariate count data are common in many disciplines. The variables in such data often exhibit complex positive or negative dependency structures. We propose three Bayesian approaches to modeling bivariate count data by simultaneously considering covariate-dependent means and correlation. A direct approach utilizes a bivariate negative binomial probability mass function developed in Famoye (2010, *Journal of Applied Statistics*). The second approach fits bivariate count data indirectly using a bivariate Poisson-gamma mixture model. The third approach is a bivariate Gaussian copula model. Based on the results from simulation analyses, the indirect and copula approaches perform better overall than the direct approach in terms of model fitting and identifying covariate-dependent association. The proposed approaches are applied to two RNA-sequencing data sets for studying breast cancer and melanoma (BRCA-US and SKCM-US), respectively, obtained through the International Cancer Genome Consortium.

KEYWORDS

bivariate count data regression, covariate-dependent correlation, Gaussian copula, liquid association, Poisson-gamma mixture model

Modeling dynamic correlation in zero-inflated bivariate count data with applications to single-cell RNA sequencing data

Zhen Yang  | Yen-Yi Ho 

Department of Statistics, University of South Carolina, Columbia, South Carolina, USA

Correspondence

Zhen Yang, Department of Statistics, University of South Carolina, Columbia, SC, USA.

Email: zheny@email.sc.edu

Funding information

National Cancer Institute, Grant/Award Number: 1R21CA264353-01

Abstract

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 coexpression patterns could often be observed. The advancements in next-generation sequencing technologies bring new statistical challenges for studying these dynamic changes of gene coexpression. 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 overdispersion and zero inflation. To explore the dynamic dependence structure in scRNA-seq data and other zero-inflated count data, new approaches are needed. In this paper, we consider overdispersion 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 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 from a study of minimal residual disease in melanoma.

KEYWORDS

correlated count data, covariate-dependent correlation, dynamic coexpression, liquid association, single-cell RNA sequencing, zero inflation

Motivation for Dynamic Association

- Biological pathways could be turned on or off under different cellular conditions.

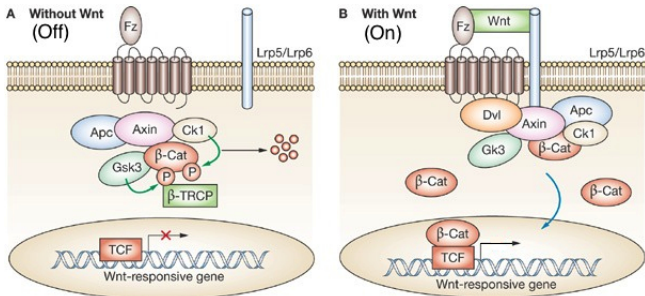
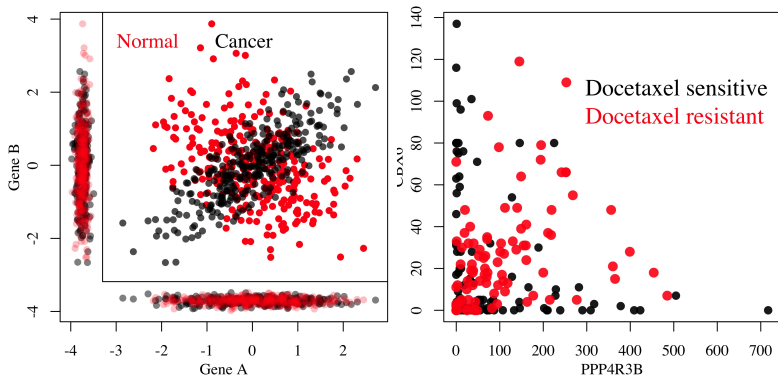


Figure: On/off switch of Wnt/ β -catenin pathway.

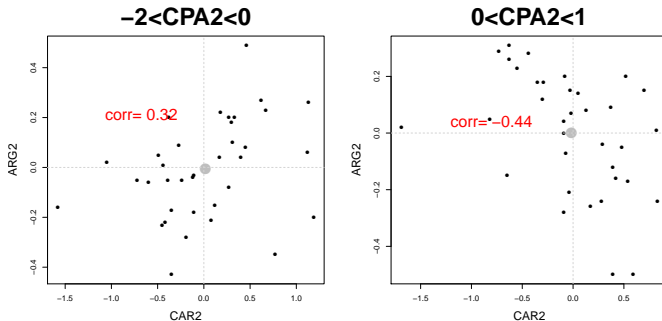
Differential Coexpression



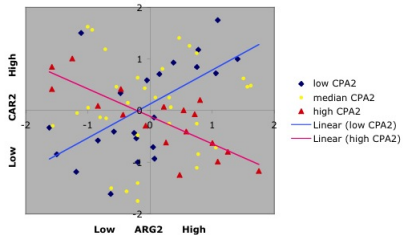
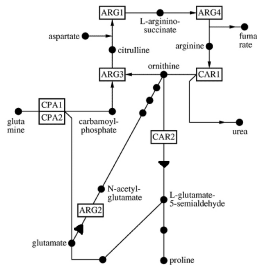
Genetic molecules and gene products participate in complex inter-connected pathways in biological systems.

Experimental Data in Yeast Urea Cycle

ARG2, CAR2 | *CPA2*

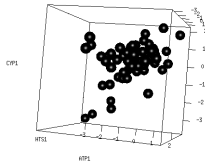


Spellman et al. Molecular Biology of the Cell 9 (1998)

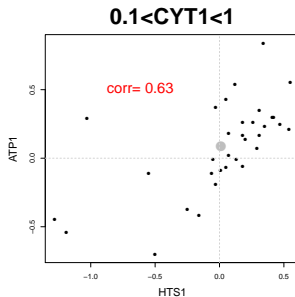
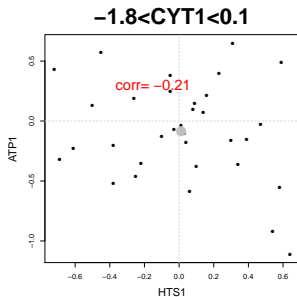


- high CPA2 : signal for arginine demand.
- up-regulation of ARG2 and down-regulation of CAR2 prevents ornithine from leaving the urea cycle.
- When the demand is relieved, CPA2 is lowered, CAR2 is up-regulated.

Yeast Electron Transport Pathway



HTS1, ATP1 | CYT1



Measuring Liquid Association

Assuming $X_3 \sim N(0, 1)$

The co-expression of $(X_1, X_2 | X_3)$:

$$h(X_3) = \text{cor}(X_1, X_2 | X_3),$$

Measuring Liquid Association

Assuming $X_3 \sim N(0, 1)$

The co-expression of $(X_1, X_2 | X_3)$:

$$h(X_3) = \text{cor}(X_1, X_2 | X_3),$$

$$\text{MLA}(X_1, X_2 | X_3) \equiv E\left[\frac{\partial h(X_3)}{\partial X_3}\right],$$

Measuring Liquid Association

Assuming $X_3 \sim N(0, 1)$

The co-expression of $(X_1, X_2 | X_3)$:

$$h(X_3) = \text{cor}(X_1, X_2 | X_3),$$

$$\text{MLA}(X_1, X_2 | X_3) \equiv E\left[\frac{\partial h(X_3)}{\partial X_3}\right],$$

By Stein's Lemma

$$= E[\text{cor}(X_1, X_2 | X_3) X_3].$$

Measuring Liquid Association

Assuming $X_3 \sim N(0, 1)$

The co-expression of $(X_1, X_2 | X_3)$:

$$h(X_3) = \text{cor}(X_1, X_2 | X_3),$$

$$\text{MLA}(X_1, X_2 | X_3) \equiv E\left[\frac{\partial h(X_3)}{\partial X_3}\right],$$

By Stein's Lemma

$$= E[\text{cor}(X_1, X_2 | X_3) X_3].$$

$$|\text{MLA}| \leq E(|X_3|) = \sqrt{\frac{2}{\pi}} \approx 0.798.$$

$$\text{Direct Estimate } \widehat{\text{MLA}} = \frac{\sum_i^M \widehat{\text{cor}}_i \overline{X_{3i}}}{M}.$$

Measuring Liquid Association: Model-Based Approach

The Conditional Normal Model (CNM)

Consider X_1, X_2, X_3 with mean 0 and variance 1.

$$X_3 \sim N(0, 1)$$

$$X_1, X_2 | X_3 \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma\right), \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

$$\mu_1 = \beta_1 X_3,$$

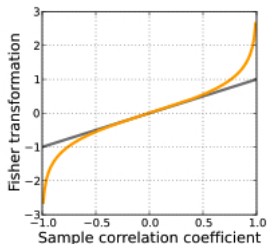
$$\mu_2 = \beta_2 X_3,$$

$$\log \sigma_1^2 = \alpha_3 + \beta_3 X_3,$$

$$\log \sigma_2^2 = \alpha_4 + \beta_4 X_3,$$

$$\log \left[\frac{1 + \rho}{1 - \rho} \right] = \alpha_5 + \beta_5 X_3.$$

$$CNM(X_1, X_2, X_3) = f(X_1, X_2 | X_3) f(X_3).$$



Liquid Association Pattern in CNM

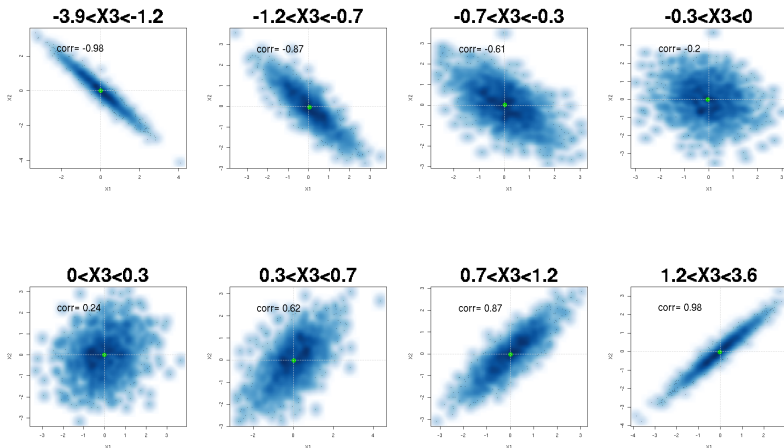


Figure: Conditional distribution of $(X_1, X_2|X_3)$ when $\beta_5 = 3$ (MLA = 0.63, $\frac{0.63}{0.798} = 0.79$), $\alpha_3 = \alpha_4 = \alpha_5 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

Theorem 1 : *Let X_1 , X_2 and X_3 be standardized random variables with mean 0 and variance 1 such that*

(i) $X_3 \sim N(0, 1)$, and

(ii) both $E(X_1|X_3) = E(X_2|X_3) = 0$, for all X_3 ,

(iii) both $\sigma_1(X_1|X_3) = \sigma_2(X_2|X_3) = 1$, for all X_3

If $E[h'(X_3)]$ and $E(X_1 X_2 X_3)$ exist,

then $MLA(X_1, X_2 | X_3) = E(X_1 X_2 X_3)$.

Generalized Estimating Equations (GEE): A Modern Love Story

GEE extends GLM and analyzes correlated data with

- Normal or non-normal response variable (Y)
- Y s that are linearly or non-linearly link to covariates
- X s can be combinations of discrete and continuous variables.

Application of GEE

- Nested data
 - Dyadic relationships
 - Family studies
 - School and organizational studies
- Repeated measures
 - Longitudinal Data Analysis
- Within subjects designs
 - Pre/post designs

The score function

$$u(\beta) = \frac{1}{n} \sum_{i=1}^n D_i' \Sigma_i^{-1} (y_i - \mu_i)$$

GEE1

$$u(\beta) = \frac{1}{n} \sum_{i=1}^n D_i' \hat{\Sigma}_i^{-1} (y_i - \mu_i)$$

Semi-Parametric Estimation Equation

$$g_1(\mu_i) = X_{1i}\beta$$

$$g_2(\sigma^2) = X_{2i}\gamma$$

$$g_3(\rho_i) = X_{3i}\alpha$$

$$u(\beta, \gamma, \alpha) = \sum_{i=1} \begin{pmatrix} D_{1i} & 0 & 0 \\ 0 & D_{2i} & 0 \\ 0 & 0 & D_{3i} \end{pmatrix} \begin{pmatrix} V_{1i} & 0 & 0 \\ 0 & V_{2i} & 0 \\ 0 & 0 & V_{3i} \end{pmatrix} \begin{pmatrix} Y_i - \mu_i \\ s_i - \sigma^2 \\ z_i - \rho_i \end{pmatrix}$$

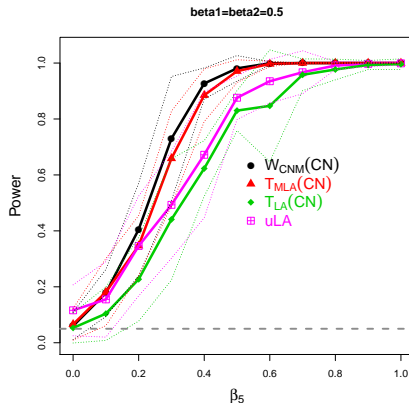
$$D_{1i} = \frac{\partial \mu_i}{\partial \beta^T}$$

$$s_{it} = (Y_{it} - \mu_{it})^2 / v_{it}$$

$$z_{its} = (Y_{it} - \mu_{it})(Y_{is} - \mu_{is}) / \sqrt{\text{Var}(Y_{it}) \text{Var}(Y_{is})}$$

Yan and Fine 2004a.

Comparison of Methods

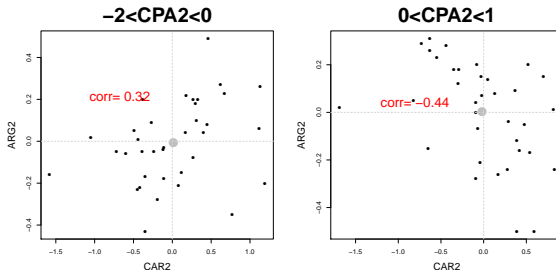


$$\begin{aligned}\mu_1 &= \beta_1 X_3, \\ \mu_2 &= \beta_2 X_3, \\ \log \sigma_1^2 &= \alpha_3 + \beta_3 X_3, \\ \log \sigma_2^2 &= \alpha_4 + \beta_4 X_3, \\ \log \left[\frac{1 + \rho}{1 - \rho} \right] &= \alpha_5 + \beta_5 X_3.\end{aligned}$$

- Power \rightarrow the ability to detect signals.
- When the data fits the model, model-based approach is more powerful.

Experimental Data Example

ARG2, CAR2 | CPA2

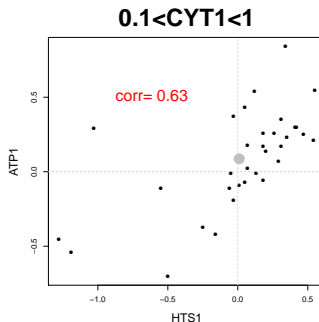
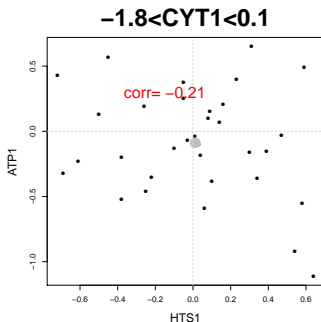


$MLA = -0.33 \left(\frac{0.33}{0.798} = 0.41 \right)$, 95% C.I. = (-0.10, -0.56), $p \text{ value} < 0.001$.

Ho. LiquidAssociation: R/Bioconductor package (2009)

Experimental Data Example

HTS1, ATP1 | CYT1



$\hat{MLA} = 0.32$, $\hat{\beta}_5 = 1.309$, 95% C.I. = (0.16, 0.49), p value=0.01.

Genome-Wide Liquid Association Analysis

Genome-wide LA analysis for $\binom{7000}{3}$ triplets:
Screening: Bin X_3 into low, median, high.

$$\begin{aligned} \text{MLA} &= \frac{\text{change in coexpression}}{\text{change in } X_3} = \frac{\text{change in } \rho}{\text{change in } X_3} = \frac{\rho_{\text{high}} - \rho_{\text{medium}}}{1 - 0} \\ &+ \frac{\rho_{\text{medium}} - \rho_{\text{low}}}{0 - (-1)} = \rho_{\text{high}} - \rho_{\text{low}} \end{aligned}$$

Gunderson, Ho et al. 2014 BMC Bioinformatics
fastLiquidAssociation: R/Bioconductor package (2014)

Conclusion

- Analyses of gene expression data suggested that liquid association is present and measurable in many biological systems.
- CNM and MLA are useful procedures to examine liquid association among three genes.
- The described modeling procedures can be applied using the LiquidAssociation R package. <http://www.bioconductor.org>, and fastLiquidAssociation for genome-wide liquid association analysis
- Empirical bayesian estimation approach is also available (EBcoexpress: R/Bioconductor package).
- Ongoing work in extending liquid association in higher-dimension focuses on ensuring positive definite the property of correlation matrix.
- Citations appeared applying liquid association in epigenetic analysis, protein pathway analysis in cancer data.

Motivation

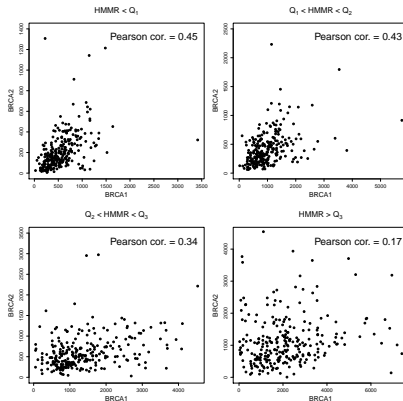


Figure: Scatterplot of $BRCA1$ vs. $BRCA2$ given $HMMR$ less than its Q_1 (top left), between Q_1 and median (top right), between median and Q_3 (bottom left), and above Q_3 (bottom right).

Introduction

- This dynamic correlation was coined as *liquid association* (Li, 2002).
- **Goal:** Regress bivariate response (Y_1, Y_2) onto covariate x , taking into consideration the covariate-dependent correlation between Y_1 and Y_2 , i.e. the liquid association between Y_1 and Y_2 .
- Further, we treat Y_1 and Y_2 as integer-valued counts, potentially over-dispersed, to reflect the nature of RNA-seq data.

Approaches

For $i = 1, 2, \dots, n$, let $\mathbf{Z}_i = [Z_{i1}, Z_{i2}]'$ be such that

$$\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n \stackrel{\text{ind}}{\sim} N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix} \right), \quad (1)$$

where ρ_i is such that

$$\log \left(\frac{1 + \rho_i}{1 - \rho_i} \right) = \tau_0 + \tau_1 X_i.$$

Note: Each Z_{ij} is marginally $N(0, 1)$, and $\Phi(Z_{ij})$ marginally $U(0, 1)$, $\Phi(\cdot)$ being the CDF of $N(0, 1)$.

Approach 1. Poisson-gamma mixture model

- Let θ be the collection of all parameters, i.e. all the regression coefficients.
- The likelihood is given by

$$L(\theta|\mathbf{y}, \mathbf{z}) = \prod_{i=1}^n \left\{ \left[\prod_{j=1}^2 p(y_{ij} | \mu_{ij}, \phi_{ij}, z_{ij}) \right] q(\mathbf{z}_i | \tau_0, \tau_1) \right\},$$

where $p(\cdot)$ is the PMF in (3), and $q(\cdot)$ is the PDF in (2).

- Need to consider prior distributions for θ . (In practice, independent $N(0, 1)$ priors were used and seemed to suffice.)
- This model can be easily fitted in R using JAGS (Martyn Plummer, 2017 [v. 4.3.0]).

Approach 2. Gaussian copula model

Let $F_{ij}(\cdot)$ be the CDF of a negative binomial distribution with mean μ_{ij} and over-dispersion ϕ_{ij} . Observe that

$$P(Y_1 = y_{i1}, Y_2 = y_{i2}) = P\{\Phi^{-1}[F_{ij}(y_{i1} - 1)] < Z_{i1} \leq \Phi^{-1}[F_{ij}(y_{i1})], \\ \Phi^{-1}[F_{ij}(y_{i2} - 1)] < Z_{i2} \leq \Phi^{-1}[F_{ij}(y_{i2})]\}, \quad (2)$$

where $[Z_{i1}, Z_{i2}]'$ is defined as in (2).

- Likelihood based on (4) requires calculating n double integrals within each MCMC loop, computationally infeasible.
- Pitt et al. (2006) considered a sampling scheme that alternates between $[Z_{i1}|Z_{i2}]$ and $[Z_{i2}|Z_{i1}]$.

Approach 2. Gaussian copula model

```
1 for  $t = 1$  to  $T$  do
2   Sample  $(\beta_1, \gamma_1)$  from
      
$$\pi(\beta_1, \gamma_1 | \cdot) \propto \pi(\beta_1, \gamma_1) \prod_{i=1}^n \left[ \Phi \left\{ \frac{\Phi^{-1}[F_{ij}(y_{i1})] - \rho_i z_{i2}}{1 - \rho_i^2} \right\} - \right. \\ \left. \Phi \left\{ \frac{\Phi^{-1}[F_{ij}(y_{i1} - 1)] - \rho_i z_{i2}}{1 - \rho_i^2} \right\} \right];$$

3   for  $i = 1$  to  $n$  do
4     Sample  $z_{i1}$  from  $N(\rho_i z_{i2}, 1 - \rho_i^2)$  truncated to
       $(\Phi^{-1}[F_{ij}(y_{i1} - 1)], \Phi^{-1}[F_{ij}(y_{i1})]);$ 
5   end
```

Figure: Sampling scheme of the Gaussian copula model.

Approach 2. Gaussian copula model

```
6 | Sample  $(\beta_2, \gamma_2)$  from  

$$\pi(\beta_2, \gamma_2 | \cdot) \propto \pi(\beta_2, \gamma_2) \prod_{i=1}^n \left[ \Phi \left\{ \frac{\Phi^{-1}[F_{ij}(y_{i2})] - \rho_i z_{i1}}{1 - \rho_i^2} \right\} - \right. \\ \left. \Phi \left\{ \frac{\Phi^{-1}[F_{ij}(y_{i2} - 1)] - \rho_i z_{i1}}{1 - \rho_i^2} \right\} \right];$$
  
7 | for  $i = 1$  to  $n$  do  
8 |   | Sample  $z_{i2}$  from  $N(\rho_i z_{i1}, 1 - \rho_i^2)$  truncated to  
   |    $(\Phi^{-1}[F_{ij}(y_{i2} - 1)], \Phi^{-1}[F_{ij}(y_{i2})])$ ;  
9 |   end  
10 | Sample  $\tau$  from  

$$\pi(\tau | \cdot) \propto \pi(\tau) \prod_{i=1}^n q(z_i | \tau).$$
  
11 end
```

Figure: Sampling scheme of the Gaussian copula model (cont'd).

Approach 2. Gaussian copula model

- JAGS is not available for this model.
- We adopted an adaptive Metropolis-Hastings scheme (Haario et al., 2001) in updating the regression coefficients.
- The JAGS code for Approach 1 and the R code for Approach 2 is available on Github at
<https://github.com/ZichenMa-USC/Correlated-bivariate-count-data-regression>.

Real data application

- In terms of the motivating example, the response \mathbf{Y} is the expression levels of *BRCA1* and *BRCA2*. *HMMR* is the covariate x .
- We fitted both models on the data.
- The total number of iterations in the Markov chain is 10000, with the first 2000 being burn-in and taking every 10th sample point as thinning.

Real data application

Table: Model comparison on regressing *BRCA1/2* against *HMMR*

	PSIS-LOO	PBF
Poisson-gamma	-10.3	1.000
Copula	-10.2	1.105

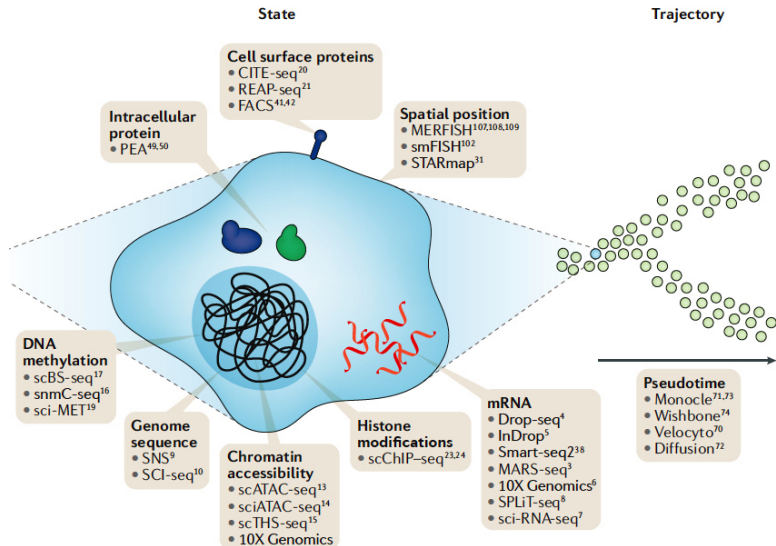
Table: Point and interval estimates of τ_1 on regressing *BRCA1/2* against *HMMR*

	$\hat{\tau}_1$	95% CI for τ_1
Poisson-gamma	-0.00038	(-0.00054, -0.00020)
Copula	-0.00064	(-0.00066, -0.00062)

Future work

- Generalization to multivariate regression in higher dimensions. The difficulty is how to make the covariance matrix positive definite.
- In the univariate negative binomial distribution, the variance is a function of the mean. Borrowing this idea, in the multivariate situation, we may model the covariance indirectly as a function of the mean.

Single-Cell Multi-Omics Data



Dynamic Coexpression

Dynamic coexpression changes: the correlations of two genes, X_1 and X_2 can be mediated by a third variable, X_3 .

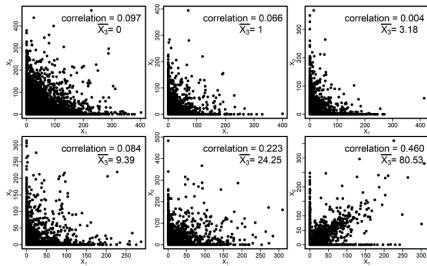


Figure: Simulated example of dynamic coexpression changes

- Single-cell RNA sequencing (scRNA-seq) data are count-based
- Zero-inflation (Yang and Ho, 2021)

Motivating Example

- Biological pathways are highly dynamic. Cancer cells can acquire drug resistance by establishing alternative bypass signaling pathways after exposure to therapeutic agents.

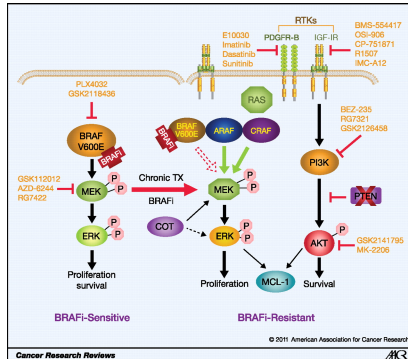
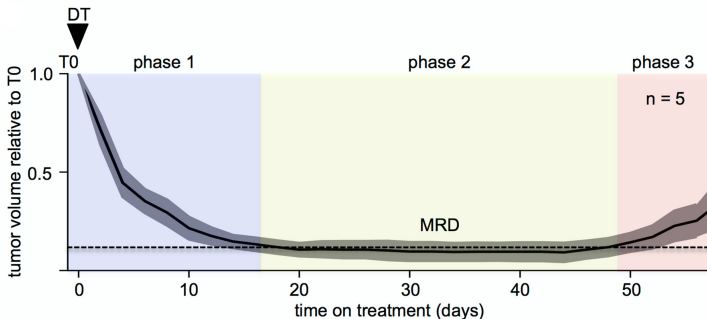


Figure adapted from Villanueva et al. (2011)

scRNA-seq Data

- BRAF mutant patient-derived xenograft (PDX) melanoma cohorts Rambow et al. (2018).
- Once the PDX tumors grew to comparable size, mice were treated with concurrent RAF/MEK-inhibition
- The data contain information for 57,445 transcripts from 675 melanoma cells from all phases.
- The three phases are: drug-sensitive, minimum residual disease (MRD), drug-resistance



The ZERo-inflated Negative binomial dynamic CORrelation (ZENCO) model

- Let X_{ij} denote the transcript counts for the i -th gene in the j -th cell and \mathbf{X}_i represents the gene expression count for the i -th gene. The distribution of \mathbf{X}_i is modelled as:

$$\mathbf{X}_i \sim \begin{cases} \text{Poisson}(\lambda_0), & \text{with probability } p_i; \\ \text{NB}(\mu_i, \phi_i), & \text{with probability } 1 - p_i. \end{cases}$$

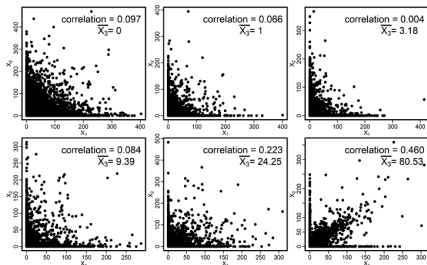
- p_i is the dropout rate of \mathbf{X}_i and is modelled as a function of μ_i : $p = \frac{e^{(b_0 + b_1 \mu)}}{1 + e^{(b_0 + b_1 \mu)}}$.

Poisson-Gamma mixture with random effects

- The correlation of a gene pair: \mathbf{X}_1 and \mathbf{X}_2 can be observed when both genes are observed in the j -th cell.
- Poisson-Gamma mixture

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

- Integrate out u_{ij} , $X_{ij} \sim \text{NB}(\mu_i, \phi_i = \frac{1}{\alpha_i})$
- u_{ij} can be considered as the cell-specific random effect



Modeling correlation structure in count data

- Let the latent variable $\mathbf{Z}_j = (Z_{1j}, Z_{2j})'$ be a bivariate normal variable that

$$\mathbf{Z}_j \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_j \\ \rho_j & 1 \end{bmatrix}\right).$$

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

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

- Plug-in \mathbf{Z}_j into u_{ij} , we have

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

where $F_{\alpha_i}(\cdot)$ is the cumulative distribution function of a $\text{Gamma}(\alpha_i, \alpha_i)$ distribution with $\alpha_i = 1/\phi_i$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

- The joint distribution of \mathbf{X}_1 and \mathbf{X}_2 can be specified using:

$$x_{ij} \sim \begin{cases} \text{Poisson}(\lambda_0), & \text{with probability } p_i; \\ \text{Poisson}[F_{1/\phi_i}^{-1}\{\Phi(z_{ij})\}\mu_i], & \text{with probability } 1 - p_i. \end{cases}$$

Simulation Analyses

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

- Under the hypotheses:

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

Table: Coverage probability (CP) of 95% credible interval (CI) and interval lengths based on 1,000 MCMC simulations ($\tau_0 = 0.01$, $\tau_1 = 0.05$)

		Without Zero-inflation		With Zero-inflation	
		CP	CI length	CP	CI length
$N = 200$	τ_0	0.997	0.455	1.000	0.541
	τ_1	0.170	0.042	0.942	0.111
$N = 500$	τ_0	0.985	0.288	1.000	0.342
	τ_1	0.009	0.022	0.950	0.064
$N = 1,000$	τ_0	0.955	0.204	1.000	0.242
	τ_1	0.000	0.014	0.951	0.043

Simulation Analyses

Table: Mean square errors (MSE) and mean bias errors (MBE) based on 1,000 MCMC simulations ($\tau_0 = 0.01$, $\tau_1 = 0.05$). MBE = $\frac{1}{N} \sum_{i=1}^N (\hat{\beta}_i - \beta)$.

		Without Zero-inflation		With Zero-inflation	
		MSE	MBE	MSE	MBE
$N = 200$	τ_0	0.008	0.044	0.001	-0.008
	τ_1	0.002	-0.039	0.001	-0.001
$N = 500$	τ_0	0.006	0.051	0.000	-0.008
	τ_1	0.002	-0.040	0.000	0.001
$N = 1,000$	τ_0	0.005	0.051	0.000	-0.009
	τ_1	0.002	-0.041	0.000	0.001

Power Comparison to existing methods

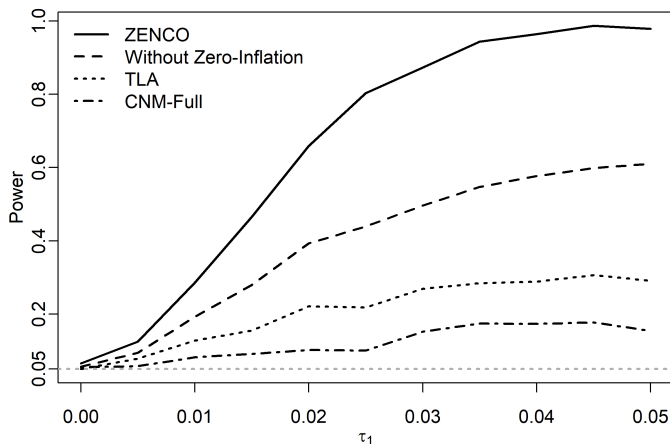


Figure: Power curves comparing various methods. Both TLA and CNM-Full approaches are Gaussian-based models Li (2002a); Ho et al. (2011a).

Experimental Data Analysis

- We use BRAF gene expression count as \mathbf{X}_3 and screen all gene-pair combinations in the KEGG melanoma pathway.

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

Gene 1	Gene 2	$\tau_1(P1)$	$\tau_1(P3)$	$\Delta\tau_1$
PDGFC	FGFR1	0.084 (0.045,0.120)	0.000 (-0.006,0.007)	-0.084
BAX	POLK	0.053 (0.023,0.085)	0.000 (-0.007,0.005)	-0.054
AKT1	ARAF	-0.024 (-0.046,-0.004)	0.019 (0.000,0.039)	0.043
AKT1	MAPK1	0.004 (-0.008,0.015)	0.043 (0.020,0.060)	0.039
AKT3	MAP2K2	0.033 (0.017,0.048)	-0.003 (-0.010,0.002)	-0.037
AKT1	BAK1	-0.027 (-0.053,-0.004)	0.008 (-0.003,0.030)	0.035
MAP2K2	FGFR1	0.031 (-0.001,0.081)	-0.003 (-0.009,0.003)	-0.033
BAX	MDM2	0.032 (0.005,0.059)	-0.001 (-0.007,0.005)	-0.033
AKT1	AKT2	0.003 (-0.009,0.014)	0.031 (0.003,0.050)	0.029
MAP2K2	BAX	0.035 (-0.006,0.075)	0.006 (-0.003,0.016)	-0.029

Integrating Single-Cell Multi-Omics Data

- Multi-omics data from single-cell experiments often contain information complement to each other
- Dynamic association: the association between two types of -omics data can change depending other data modality (Li, 2002b; Ho et al., 2011b; Chen et al., 2011)
- For example: (chromatin accessibility, gene expression) | DNA methylation
- The statistical challenge is that different data types often have distinct marginal distribution characteristics

Motivating Example

Dynamic association: the association Y_1 (gene expression) and Y_2 (chromatin accessibility) can be mediated by a third variable, X (DNA methylation).

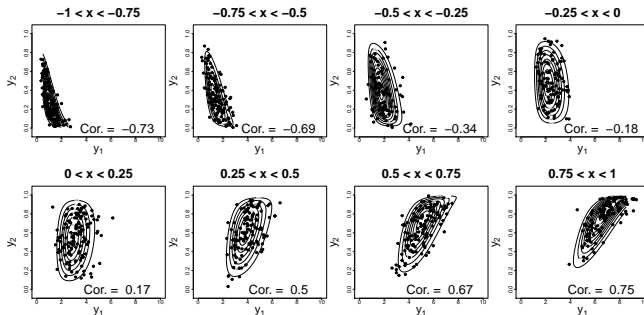


Figure: Simulated example of dynamic association

Flexible Copula-Based Model

$$\mathbf{Z}_i \sim N_2 \left(\mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{R}_i = \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix} \right) \quad (3)$$

$$\rho_i = \text{corr}(Z_{i1}, Z_{i2}) = \frac{\exp(\mathbf{x}_i' \boldsymbol{\tau}) - 1}{\exp(\mathbf{x}_i' \boldsymbol{\tau}) + 1}, \quad (4)$$

For both discrete and continuous marginals, the general form of the joint CDF of \mathbf{Y} can be written as:

$$F_{\mathbf{Y}}(\mathbf{y}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\tau}, \mathbf{x}_i) = \Phi_{\boldsymbol{\tau}} \left(\Phi^{-1}[F_1(y_{i1}; \boldsymbol{\theta}_1, \mathbf{x}_i)], \Phi^{-1}[F_2(y_{i2}; \boldsymbol{\theta}_2, \mathbf{x}_i)] \right), \quad (5)$$

where $\Phi_{\boldsymbol{\tau}}$ is the joint CDF of \mathbf{Z}_i , and Φ^{-1} represents the inverse CDF of $N(0, 1)$.

Conclusion

- Single-cell data often exhibit characteristics unique to specific data modality.
- Our proposed Poisson-Gamma, and copula-based models can accommodate a wide range of marginal distributions.
- In addition to modeling the marginal means, our models consider covariate-dependent correlation structure.
- The expression level of BRAF was considered as the modulator variable \mathbf{X}_3 . In other applications, \mathbf{X}_3 can be easily modified to represent other conditions such as tumor status, degree of inflammation, or cell types, ...etc.

Acknowledgement

Research reported in this talk is supported by National Cancer Institute of the National Institutes of Health under award number 1R21CA264353.

References I

- Chen, J., Xie, J., and Li, H. (2011). A penalized likelihood approach for bivariate conditional normal models for dynamic co-expression analysis. *Biometrics*, 67(1):299–308.
- Gunderson, T. and Ho, Y.-Y. (2014). An efficient algorithm to explore liquid association on a genome-wide scale. *BMC bioinformatics*, 15(1):371.
- Ho, Y.-Y., Parmigiani, G., Louis, T. A., and Cope, L. M. (2011a). Modeling liquid association. *Biometrics*, 67(1):133–141.
- Ho, Y.-Y., Parmigiani, G., Louis, T. A., and Cope, L. M. (2011b). Modeling liquid association. *Biometrics*, 67(1):133–141.
- Li, K.-C. (2002a). Genome-wide coexpression dynamics: theory and application. *Proceedings of the National Academy of Sciences*, 99(26):16875–16880.
- Li, K.-C. (2002b). Genome-wide coexpression dynamics: theory and application. *Proceedings of the National Academy of Sciences*, 99(26):16875–16880.

References II

- 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.
- Villanueva, J., Vultur, A., and Herlyn, M. (2011). Resistance to BRAF inhibitors: unraveling mechanisms and future treatment options. *Cancer research*, 71(23):7137–7140.
- Yang, Z. and Ho, Y.-Y. (2021). Modeling dynamic correlation in zero-inflated bivariate count data with applications to single-cell RNA sequencing data. *Biometrics*.
- 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.

Search Strategies

- For a given pair of genes ($\mathbf{X}_1, \mathbf{X}_2$), screen the whole-genome to identify a third modulator gene.
- For a given modulator variable (\mathbf{X}_3), screen the whole-genome to identify a pair of genes that are modulated by \mathbf{X}_3 ($\binom{m}{2}$, m is the total number of genes).
- If no prior information about \mathbf{X}_3 or ($\mathbf{X}_1, \mathbf{X}_2$) is available, screen the relevant pathways or the whole genome to identify potential gene triplets ($\binom{m}{3}$).
- When the number of genes under considerations is large (for example $\approx 20,000$). Pre-screening is beneficent such as Gunderson and Ho (2014) or the screening statistic (ζ) introduced in Yu (2018).