4-8-2009

GENERALIZED LIQUID ASSOCIATION

Yen-Yi Ho
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, yho@jhsph.edu

Leslie Cope
The Johns Hopkins University School of Medicine, Oncology Bioinformatics

Thomas A. Louis
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Generalized Liquid Association

Yen-Yi Ho, Leslie Cope, Thomas A. Louis, Giovanni Parmigiani

Abstract

The analysis of interactions among a group of genes is fundamental to further our understanding of their biological interactions in a cell. Several studies suggested that the co-expression relationship of two genes can be modulated by a third controller gene. These controller genes and the corresponding modulated co-expressed gene pairs are the subjects of interests in this study. This described "controller-modulated genes" three-way interactions is referred as liquid association in the literature. Analysis of gene expression data has suggested that these interactions are present in many biological systems.

To quantify the magnitude of liquid association for a given gene triplet, we proposed a statistical measure named generalized liquid association (GLA). To estimate the value of GLA given the data, we propose two approaches: the direct and the model-based estimation approach. For the model-based approach, we introduce the conditional normal model (CNM). This is a generalization of the tri-variate normal distribution that allows us to characterize means, variances, as well as liquid association structures. We provide an approach based on generalized estimation equations to estimate the parameters in the CNM. We validate the proposed approaches through simulation studies and illustrate them in experimental data analysis. We also compare them with the three-product-moment measure suggested by Li in various settings and discuss related computational issues.

Key words: Higher-order interaction; Liquid association; Non-Gaussian multivariate distribution; Generalized estimating equations.

Introduction

Gene expression microarrays made it possible to measure the levels of thousands of RNA transcripts at the same time. With this high throughput technology, it is possible to study the interactions among genes and to further elucidate cellular biological networks. A natural approach to study gene interactions is to group genes with similar profiles across samples and to investigate these genes as functional modules [14]. However, research findings indicate that gene co-expression relationships often exist specifically under certain biological conditions [4], [2], [7]. Moreover, several studies suggested that sometimes the gene co-expression

relationship can be modulated by a third controller gene during a biological process [9], [10], [8], [17].

These controller genes and the corresponding modulated co-expressed gene pairs are the subjects of interests in this study. Because these "controller-modulated genes" triplets could provide further information about their functional interactions and the mechanism to turn-on or turn-off these interactions. One might also be interested in constructing genetic networks considering these three-way "controller-modulated genes" interactions besides the traditional two-way interactions [3]. This described interactions among "controller-modulated genes" triplets is referred as "liquid association" in the literature.

Prior biological knowledge provides evident support for the observed liquid association phenomenon. One example is from the regulation of the *Wnt* pathway and β-catenin [5]. When *Wnt* is present and binds to its receptor in the cell, the co-expression of β-catenin and transcription factors can be expected. However, when *Wnt* is absent, co-expression relationship would not be expected.

To identify these "controller-modulated genes" triplets through microarray experiments, we need to quantify the magnitude of liquid association given the data of a gene triplet. For this purpose, we propose a statistical measure, called generalized liquid association (GLA). In GLA, we used the conditional correlation coefficient to capture the co-expression relationship between two genes given the level of a third gene and measure the degree of modulation of conditional correlation by the third gene. To estimate the value of GLA given the data, we propose two approaches: the direct and the model-based estimation approach. For the model-based approach, we introduce the conditional normal model (CNM). The CNM describes the joint distribution of the three genes, while considering the means, variances and liquid association structures among them.

In this paper, we provide a generalized estimation equations (GEE)-based approach to estimate the parameters in the CNM. We illustrate the proposed approaches through simulation studies and experimental data analysis. We also compare the two approaches with the three-product-moment measure suggested by Li in various settings and discuss related computational issues. The organization of the paper is as follows. In Section 2, we briefly describe the three-product-moment measure proposed by Li, present related issues, and propose the GLA. In Section 3, we describe the CNM and its properties. In Section 4, we introduce the estimation and hypothesis testing procedures. In section 5, we analyze a *Saccharomyces cerevisiae* data set for illustration. Conclusion and highlight of future research directions are presented in Section 6. The proofs of theorems are in the Appendices.

1 Liquid Association

Li proposed the concept of liquid association and used the term "liquid", in contrast with "solid", to describe how the coexpression pattern of two genes, X_1 and X_2, changes according to the level of a third gene, X_3. Consider the random variables X_1, X_2 and X_3 to represent the expression levels of three genes. Standardize these random variables to have mean 0 and variance 1. Li uses $E(X_1X_2 \mid X_3)$ to measure the co-expression relationship between X_1 and

X_2 given the value of X_3, and derived the three-product-moment measure as follows:

$$\begin{aligned} g(X_3) &= E(X_1X_2 \mid X_3), \\ \mathrm{LA}(X_1X_2 \mid X_3) &= E(g'(X_3)) = E(X_1X_2X_3). \end{aligned}$$

In the above equation, $g'(x)$ denotes the derivative of $g(x)$ with respect to x. Li proposed a direct estimate for the liquid association given by $\sum_i \frac{X_{1i}X_{2i}X_{3i}}{n}$, where n is the total number of observations. However, when the conditional means and variances of X_1 and X_2 also depend on X_3, $E(X_1X_2 \mid X_3)$ depends on the conditional correlation as well as the conditional means and variances. Specifically,

$$E(X_1X_2 \mid X_3) = \rho(X_1, X_2 \mid X_3)\sigma_1(X_1 \mid X_3)\sigma_2(X_2 \mid X_3) + E(X_1 \mid X_3)E(X_2 \mid X_3),$$

where $\sigma_1^2(X_1 \mid X_3)$ and $\sigma_2^2(X_2 \mid X_3)$ are the conditional variances of X_1 and X_2 given X_3, respectively, and $\rho(X_1, X_2 \mid X_3)$ is the conditional correlation of X_1 and X_2 given X_3. As a result, $E(X_1X_2X_3)$ also depends on the conditional mean and variance of X_1 and X_2:

$$E(X_1X_2X_3) = E[X_3\rho(X_1, X_2 \mid X_3)\sigma_1(X_1 \mid X_3)\sigma_2(X_2 \mid X_3)] + E[X_3E(X_1 \mid X_3)E(X_2 \mid X_3)].$$

When the conditional means and variances also depend on X_3, then the three-product-moment measure, as originally defined by Li for standardized variables, no longer captures fully how the dependence between X_1 and X_2 is modulated by X_3. This can be demonstrated through the following example, where we show that $E(X_1X_2X_3) \neq 0$ even when the conditional correlation of X_1 and X_2 given X_3 does not change with X_3.

Example 1: Consider $E(X_1 \mid X_3) = E(X_2 \mid X_3) = 0$, $\sigma_1^2(X_1 \mid X_3) = \sigma_2^2(X_2 \mid X_3) = e^{-1/2+X_3}$. We assume the conditional correlation is constant. Specifically, let $\rho(X_1, X_2 \mid X_3) = 0.5$. It follows that $E(X_1X_2 \mid X_3) = 0.5e^{-1/2+X_3}$, and $E(X_1X_2X_3) = E[E(X_1X_2 \mid X_3)X_3] = 0.5E(e^{-\frac{1}{2}+X_3}X_3) = 0.5$, which is not zero.

To measure liquid association when the conditional means and variances also depend on X_3, we propose using $\rho(X_1, X_2 \mid X_3)$ as the the coexpression measure of X_1 and X_2 given X_3 instead of $E(X_1X_2 \mid X_3)$:

$$h(X_3) = \rho(X_1, X_2 \mid X_3).$$

Following this definition, assuming X_3 is distributed as N(0,1), the form of generalized liquid association (GLA) is:

$$\mathrm{GLA}(X_1, X_2 \mid X_3) = E[h'(X_3)] = E[\rho(X_1, X_2 \mid X_3)X_3], \tag{1}$$

where $h'(x)$ denotes the derivative of $h(x)$ with respect to x. GLA represents the expected value of the change of the conditional correlation with X_3. Using the example described above, the newly defined GLA measure is able to correctly conclude that the correlation of X_1 and X_2 does not change with X_3. That is, $\mathrm{GLA}(X_1, X_2 \mid X_3) = E[\rho(X_1, X_2 \mid X_3)\ X_3] = 0.5E(X_3) = 0$. A direct estimate for $\mathrm{GLA}(X_1X_2 \mid X_3)$ given the data is:

$$\frac{\sum_i^M \hat{\rho}_i \bar{X}_{3i}}{M}, \tag{2}$$

where M is a given number of grid points over X_3, $\hat{\rho}_i$ is the sample Pearson correlation coefficient of X_1 and X_2 using only those observations with X_3 in grid i, and $\bar{X}_{3i}$ is the mean of X_3 in grid i. In addition, based on equation (1), the upper bound of GLA can be achieved when $\rho(X_1, X_2 \mid X_3)X_3 = |X_3|$, that is:

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

Example 2: Assume X_1, X_2 and X_3 follow the tri-variate Clayton copula with standard normal marginals [12], [1]:

$$\begin{aligned} F(X_1, X_2, X_3) &= C(F_1(X_1), F_2(X_2), F_3(X_3)), \\ C(u_1, u_2, u_3) &= (u_1^{-\theta} + u_2^{-\theta} + u_3^{-\theta} - 2)^{-1/\theta} \end{aligned}$$

and also assume that F_1, F_2 and F_3 are all equal to the cumulative density function of a standard normal. A three-dimensional scatter plot of X_1, X_2 and X_3 with $\theta = 3$ is shown in Figure 1. In this example, the conditional means and variances of X_1, and X_2 depend on X_3, but not the conditional correlation of X_1 and X_2. We simulated 10,000 observations of X_1, X_2 and X_3 with various θ ranging from 0 to 3 and calculated the direct estimates of $\text{GLA}(X_1, X_2 \mid X_3)$ and $E(X_1X_2X_3)$ using the simulated data. As a result, we observe a large difference between the direct estimates of $\text{GLA}(X_1, X_2 \mid X_3)$ and $E(X_1X_2X_3)$ in Figure 2. In this example, the $\text{GLA}(X_1, X_2 \mid X_3)$ estimate correctly captures that fact that X_3 does not modulate that relationship between X_1 and X_2, while $E(X_1X_2X_3)$ is misled by the conditional means and variances.

We now present a theorem giving conditions for the equivalence between GLA and the three-product-moment measure. We provide a detailed proof in the Appendix A.
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) $E(X_1 \mid X_3) \perp X_3$ and $E(X_2 \mid X_3) \perp X_3$,
(iii) $\sigma_1(X_1 \mid X_3) \perp X_3$ and $\sigma_2(X_2 \mid X_3) \perp X_3$,
If $E[h'(X_3)]$ and $E(X_1X_2X_3)$ exist, then $GLA(X_1, X_2 \mid X_3) = E(X_1X_2X_3)$.

In the above theorem, (i) (ii), and (iii) are sufficient conditions for $\text{GLA}(X_1, X_2 \mid X_3) = E(X_1X_2X_3)$. Next, we also propose a model-based approach for GLA which allows us to consider conditional mean, variance and liquid association simultaneously. In the following sections, we introduce the CNM and compare the performances of the direct estimate and model-based approaches.

2 The Conditional Normal Model

Consider three standardized random variables, X_1, X_2 and X_3 with mean 0 and variance 1 and assume that X_3 is distributed as N(0,1). Given X_3, we assume that $(X_1, X_2)'$ follows a

bivariate normal distribution with conditional mean $(\mu_1, \mu_2)'$ and variance Σ, that is:

$$\begin{aligned} X_3 &\sim N(0,1) \\ \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \mid X_3 &\sim N(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma). \end{aligned}$$

where $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. The conditional mean $(\mu_1, \mu_2)'$ and covariance matrix Σ are further specified as:

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

The marginal variances of X_1 and X_2 are both 1, therefore the parameters in CNM need to satisfy the relations: $e^{\alpha_3 + \frac{\beta_3^2}{2}} + \beta_1^2 = 1$ and $e^{\alpha_4 + \frac{\beta_4^2}{2}} + \beta_2^2 = 1$ and the following inequalities: $|\beta_1| < 1, |\beta_2| < 1, \alpha_3 + \frac{\beta_3^2}{2} < 0$, $\alpha_4 + \frac{\beta_4^2}{2} < 0$. In this model, we choose the log link function for the variances (σ_1^2, σ_2^2) and the rescaled Fishers Z-transformation for the correlation, ρ, to ensure that the variances are positive and the correlation is within (-1, 1). The choice of the link functions in CNM can also be modified for specific problems. Interested readers can refer to [11], [13] for more discussion on this topic. Finally, the joint distribution of X_1, X_2 and X_3, $CN(X_1, X_2, X_3)$, can be expressed as the product of the conditional distribution of X_1, X_2 given X_3, $f(X_1, X_2 \mid X_3)$, and the marginal distribution of X_3, $f(X_3)$:

$$CN(X_1, X_2, X_3) = f(X_1, X_2 \mid X_3) f(X_3).$$

The LA pattern of CNM can be demonstrated by examining the change of conditional correlation of X_1 and X_2, given various levels of X_3. In Figure 3, we simulate 10^5 observations from the CNM with $\beta_5 = 0.5$ and $\alpha_3 = \alpha_4 = \alpha_5 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ and plot a panel of conditional distributions of X_1 and X_2 given various levels of X_3. In these figures, we observe that ρ increases with X_3, as the GLA is 0.236 in this simulation example.

The magnitude of GLA in the CNM is controlled by the values of α_5 and β_5. The relation of β_5 and GLA can be written as:

$$\text{GLA} = \int_{X_3} \rho(X_3)\ X_3 N(X_3; 0, 1)\ dX_3 = \int_{X_3} \frac{e^{\alpha_5 + \beta_5 X_3} - 1}{e^{\alpha_5 + \beta_5 X_3} + 1}\ X_3 N(X_3; 0, 1)\ dX_3.$$

This functional relation is plotted in Figure 4. It can be observed that GLA increases monotonically with β_5 and GLA has the same sign as β_5.

The CNM is generally not symmetric in index. For example, $\text{CN}(X_1, X_2, X_3) \neq \text{CN}(X_1, X_3, X_2)$. Using the data simulated previously from $\text{CN}(X_1, X_2, X_3)$, we observed that the conditional

correlation of X_1 and X_3 increases according to the level of X_2 as shown in Figure 5. The relation between $\text{GLA}(X_1, X_2 \mid X_3)$ and $\text{GLA}(X_1, X_3 \mid X_2)$ under $CN(X_1, X_2, X_3)$ with fixed $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \alpha_3 = \alpha_4 = \alpha_5 = 0$, and varied β_5 is shown in Figure 6. In this plot, we observed that $\text{GLA}(X_1, X_2 \mid X_3)$ is not equivalent to $\text{GLA}(X_1, X_3 \mid X_2)$ in this simplified settings. This implies that in general GLA is not invariant to permutation of indexes in CNM.

3 Estimation and Hypothesis Testing

We used the three-estimating-equations (3EE) approach by Yan [16] to estimate the parameters in the CNM. Three estimating equations were constructed for the mean, variances and correlation coefficient parameters. The standard errors of the parameter estimates were obtained through the sandwich estimators, referred to as san.se. Wald tests were performed for each parameter estimate, to examine whether the parameters significantly differ from zero. To verify the 3EE approach, we simulated data with 100 observations with true $\beta_5 = 0.5$ and repeated the simulation 1000 times. In this simulation, the 80 % confidence interval covered the true β_5 766 times (coverage rate = 76.6 %).

We consider four approaches to test for the existence of liquid association. The first approach is based on the CNM, the second and the third approaches are based on the direct estimates of GLA and $E(X_1X_2X_3)$, respectively. The fourth is the approach suggested by Li [9].

When $\beta_5 = 0$ in the CNM, the correlation of X_1, X_2 does not depend on X_3, hence GLA is 0. Thus, to test for the existence of LA in the CNM, the null hypothesis can be formulated as:

$$H_0 : \beta_5 = 0.$$

The proposed statistic GEEb_5, based on the Wald test, can be written as: $\text{GEEb}_5 = \frac{\hat{\beta}_5}{SE(\hat{\beta}_5)}$, where $SE(\hat{\beta}_5)$ is the standard error of $\hat{\beta}_5$ which can be computed through the sandwich estimator using GEE. Also $\hat{\beta}_5$ can be estimated using the 3EE approach as described previously. Notice that GEEb_5 is applicable when the conditional means and variances also depend on the third gene, because the CNM would be able to account for changes of conditional means and variances with X_3.

In more general settings that the CNM, the null hypothesis of no liquid association can also be written as:

$$H_0 : \text{GLA}(X_1, X_2 \mid X_3) = 0$$

Motivated by this we defined a second test statistic as:

$$\text{sGLA} = \frac{\widehat{\text{GLA}}}{SE(\widehat{\text{GLA}})},$$

where $\widehat{\text{GLA}}$ can be calculated using equation (2), and $SE(\widehat{\text{GLA}})$ is the standard error of $\widehat{\text{GLA}}$ and can be estimated using bootstrap. To obtain the distribution of sGLA under the null

hypothesis, we consider two procedures. First, we obtained the null distribution of sGLA by simulating data from the conditional normal model with $\beta_5 = 0$. The second procedure is to permute the observations of X_3 and treat the permuted data as samples under the null hypothesis [9].

When conditions in theorem 1 hold, we can also use $E(X_1X_2X_3)$ to test for the existence of liquid association. A third test statistics is written as:

$$\text{sLA} = \frac{\widehat{E(X_1X_2X_3)}}{SE(\widehat{E(X_1X_2X_3)})}.$$

where $\widehat{E(X_1X_2X_3)} = \sum_i (X_{1i}X_{2i}X_{3i})/n$ and $SE(\widehat{E(X_1X_2X_3)})$ can be calculated using bootstrap samples. Finally, a fourth testing procedure based on the direct estimate is proposed by Li [9]; the test statistics can be written as $\text{uLA} = \widehat{E(X_1X_2X_3)}$. The uLA is an unstandardized measure.

4 Simulation Studies

To evaluate the performance of the four test statistics described above, we perform power analyses under three scenarios. In the first scenario, we simulated 300 observations from CNM with all parameters fixed at 0 except β_5. For the model-based test statistic, we evaluated the power of GEEb_5 under two different null distribution approaches: (1) the CNM with all parameters equal to 0, denoted as 'CN', and (2) the null distribution obtained permuting X_3 denoted as 'perm'. For sGLA, we used bootstrap with 100 repetitions to calculate $SE(\widehat{\text{GLA}})$. We evaluated the power of sGLA under the same null distribution approaches as described above. Similar procedures were performed for sLA. We repeated the procedure 100 times to get the alternative and null distribution of the four test statistics, GEEb_5, sGLA, sLA and uLA, and obtained the corresponding power.

In the second simulation scenario, we generated data by fixing all the parameters to 0, except $\beta_1 = \beta_2 = 0.5$, and then varied β_5 values. In this simulation scheme, the conditional means depends on the value of X_3. We used the same procedure as described above to obtain the power of the three test statistics and the corresponding 95 % confidence intervals.

We also investigated the robustness of the model-based test statistics when data are not generated from a CNM. A recent study suggested using copulas to model the dependence structure in gene expression data [6]. Hence in the third scenario, we simulated data by assuming that X_1 and X_2 follow a T copula with standard normal margins and 1 degree of freedom. In addition, we assumed that the correlation of X_1 and X_2 in the T copula depends on the level of X_3 through the following relation: $\log(\frac{1+\rho}{1-\rho}) = \beta X_3$. An example of the conditional distributions of X_1 and X_2 given X_3 in this scenario is shown in Figure 9.

The results of our simulation studies are presented in Figures 8 and 10. For scenario 1, the power of GEEb5, sGLA, sLA and uLA are almost the same under the two null distributions, as shown in Figure 8 (left). For the second simulation scenario, shown in Figure 8 (right), we observed that the model-based statistic GEEb5 outperforms sLA. This is because that

the model-based estimator is able to account for the dependence of conditional means, and gains better power to detect liquid association.

In addition, the uLA testing procedure proposed by Li demonstrated elevated type I error rate in scenario 2 as indicated by the fact that the corresponding curve is above 0.05 at $\beta_5 = 0$. This is because permuting X_3 leaves it completely independent of the other 2 variables rather than selectively destroying the dependence between $\rho(X_1, X_2)$ and X_3. This leads to an under-estimation of the variance of $E(\widehat{X_1X_2X_3})$. If one applies the unstandardized test statistic proposed by [9], anti-conservative results might occur if the conditional means or variances of X_2 or X_1 depend on X_3. The results from the third simulation scenario are shown in Figure 10, we find that the GEEb_5 demonstrates robustness even when X_1 and X_2 are not generated from bivariate normal distribution. GEEb_5 achieves similar power as sGLA in this scenario.

With regard to the issue of computational time, it took approximately 0.65 times longer to obtain the estimates and their standard errors from the CNM compared to the time required by calculating the direct-estimate, sGLA, and its bootstrap standard error (based on 100 bootstrap iterations).

5 Experimental Data Analysis

We now use the proposed CNM and estimation procedure to analyze the *Saccharomyces cerevisiae* cell-cycle dataset described in [15]. This dataset is available at http://genome-www.stanford.edu/cellcycle. The dataset contains the RNA abundance measures for 6,178 genes under 73 conditions. After removing genes missing more than 30 % of measurements, we identify 150 genes with variances larger than 0.5 in the remaining 5,721 genes.

We first performed the normal quantile transformation for each gene as described in Li [9] so that marginally each gene expression intensity is symmetric. In addition, we also standardized each gene so that they have mean 0 and variance 1. We calculated $\hat{\text{GLA}}(X_1, X_2 \mid X_3)$, $\hat{\text{GLA}}(X_1, X_3 \mid X_2)$ and $\hat{\text{GLA}}(X_2, X_3 \mid X_1)$ for each 551,300 gene triplet combinations generated from the 150 genes. We identified 11 gene triplets with at least one $|\hat{\text{GLA}} > 0.5|$ and listed them in Table 1. According to the three direct GLA estimation results for a given triplet, we noticed that the values GLAs are not exactly the same. The results from experimental data analysis might suggest that GLA is not symmetric in index.

We ordered the gene triplets so that GLA is the largest of the three and applied CNM to the 11 gene triplets. The results are shown in Table 2. Among these 11 triplets, the conditional means and/or variances of triplet # 1, 2, 3, 4, 6, 7, and 8 depend on the third gene. We show several scatter plot examples of the conditional mean and variances depending on the third gene in Figure 11. In these cases, the original three-product-moment measure, $E(X_1X_2X_3)$, can be misled by the conditional mean and variances as discussed in Section 2. On the other hand, the proposed CNM is able to more accurately quantify GLA by modeling conditional means and variances together. As examples, in triplet # 1, a noticeable difference was observed two estimates, $\widehat{\text{GLA}} = 0.534$ (95 % bootstrap C.I.:

0.311, 0.608) and $E(\widehat{X_1X_2X_3}) = 0.465$ (95 % bootstrap C.I.: 0.284, 0.631). In addition, in triplet # 8, $\widehat{\text{GLA}} = 0.522$ (95 % bootstrap C.I.: 0.367, 0.595) and $E(\widehat{X_1X_2X_3}) = 0.686$ (95 % bootstrap C.I.: 0.451, 0.878). Comparing table 1 and 2, we observed the model-based estimates (GEEb5) reached similar conclusions compared with the direct estimates, sGLA. This suggests that the experimental data does not deviate dramatically from the conditional normal model assumptions.

#	X_1	X_2	X_3	$(X_2, X_3 \mid X_1)$	$(X_1, X_3 \mid X_2)$	$(X_1, X_2 \mid X_3)$	$E(X_1 \hat{X}_2 X_3)$	sGLA*	P value*
1	PSA1	PIN4	GPH1	0.303	0.444	0.534	0.465	7.271	<0.001
2	HTA1	PIN4	GPH1	0.326	0.396	0.514	0.434	5.196	0.001
3	MRH1	YDR225W	YDR033W	0.348	0.434	0.504	0.614	5.799	<0.001
4	PHO12	YDR225W	SCW11	0.316	0.335	0.517	0.702	7.102	<0.001
5	PHO12	YDR225W	DSE4	0.376	0.399	0.54	0.61	7.654	<0.001
6	WSC4	YDR225W	CDA2	0.396	0.408	0.502	0.605	7.744	<0.001
7	OLE1	MSH6	GLK1	-0.397	-0.436	-0.513	-0.672	-8.307	<0.001
8	MF(ALPHA)2	HSP12	WSC4	0.446	0.49	0.522	0.686	6.228	<0.001
9	OLE1	AXL2	GAD1	-0.388	-0.438	-0.562	-0.6	-9.914	<0.001
10	CSI2	OLE1	GAD1	-0.341	-0.344	-0.51	-0.548	-8.591	<0.001
11	OLE1	CLN2	GAD1	-0.326	-0.389	-0.509	-0.5	-9.486	<0.001

Table 1: A list of 11 triplets with at least one $|\widehat{\text{GLA}} > 0.5|$.
*The sGLA and p value are calculated using the order of the triplet that yields the maximum GLA estimate.

#	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\alpha_5}$	$\hat{\beta}_5$	GEEb5	P value
1	0.170	0.330*	0.133	0.192	2.219	4.682	2.262	0.024
2	0.009	0.294*	0.394*	0.175	1.008	2.133	2.634	0.008
3	-0.114	0.262*	0.165	0.021	0.482	1.937	3.027	0.002
4	-0.132	0.346*	0.007	0.093	0.340	1.881	4.003	<0.001
5	-0.079	0.086	0.180	0.030	-0.011	1.760	3.521	<0.001
6	-0.446*	0.202	0.005	-0.017	-0.199	2.453	3.834	<0.001
7	0.201*	0.185	-0.692*	-0.375*	-0.096	-2.343	-3.209	0.001
8	-0.161	-0.415*	0.100	-0.196	0.207	1.693	4.233	<0.001
9	0.004	-0.081	-0.077	0.003	0.011	-3.199	-3.676	<0.001
10	-0.093	-0.009	0.087	-0.106	0.311	-2.440	-3.436	0.001
11	0.020	0.077	-0.090	0.143	0.418	-3.187	-3.749	<0.001

Table 2: Estimates of CNM parameters using the 11 triplets with highest GLA.

* P value < 0.05.

6 Conclusion

In this article, we proposed methodologies for formal statistical inference on modulating correlations, that is on bivariate correlation that vary with the level of a third variable. Our discussion includes both exploratory and model based approaches. Our work was motivated by the pioneering work of Li [9] on Liquid Association, a concept that we generalized here. Among other advantages, our generalization overcomes a potential limitation of the three-product-moment estimator, which arises when the means and variances of the two variables studies, in addition to their correlation, depends on the third variable. This effect was observed in a number of gene triplets in the experimental data analysis.

Specifically, we proposed a generalized metric for Liquid Association (GLA), and two estimation procedures: the direct and the model-based estimation approaches. To address GLA within a model-based approach, we introduced the conditional normal model (CNM). This model quantifies liquid association, while also accommodating for dependence in the conditional means and variances. We provided a GEE-based estimation procedure to estimate the parameters in the CNM and introduced the model-based statistic, GEEb5, and the sGLA statistic based on the direct-estimate, to test for the existence of liquid association.

We first compared the power of GEEb5 and sLA through simulation studies. When the conditional means depend on the third gene, GEEb5 outperformed sLA. A disadvantage of sLA, the three-product-moment-based approach, is that it is only applicable when the conditions in theorem 1 hold. To investigate the robustness of GEEb5, we compare GEEb5 and sGLA when the data were not generated from the CNM. The GEEb5 as a model-based statistic has commensurable power compared with sGLA, the direct-estimate based approach. Finally, a fourth statistic we examined is the unstandardized statistic proposed by Li, denoted as uLA. The results from the simulation studies suggested that in some situations, uLA showed elevated type-I error rate and is likely to give anti-conservative conclusions in real data applications.

Although the model-based estimate demonstrated satisfying performance in simulations, even when the data were not generated from CNM, the model assumptions in the CNM might be a potential concern in real data applications, so we compared the model-based estimates on gene expression data with those from direct-estimates. As the results in Table 1 and 2 show, the model-based estimates were comparable to direct estimates, suggesting that the experimental data does not deviate dramatically from the conditional normal model assumptions.

Finally, a direction for future research is to extend the CNM to higher dimensions (more than three genes). One important issue is that the covariance matrix Σ is not guaranteed to be positive definite if we consider more than three genes. The second topic for future research is to incorporate the CNM into Bayesian network modeling so that we can consider not only pairwise interactions but also three-way interactions in network analysis.

References

[1] Kjersti Aas. Modelling the dependence structure of financial assets: A survey of four copulas. Technical report, SAMBA, 2004.

[2] Marcel Dettling, Edward Gabrielson, and Giovanni Parmigiani. Searching for differentially expressed gene combinations. *Genome Biol*, 6(10):R88, 2005.

[3] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, Feb 2004.

[4] Yen-Yi Ho, Leslie Cope, Marcel Dettling, and Giovanni Parmigiani. Statistical methods for identifying differentially expressed gene combinations. *Methods Mol Biol*, 408:171–191, 2007.

[5] A. Kikuchi. Regulation of beta-catenin signaling in the wnt pathway. *Biochem Biophys Res Commun*, 268(2):243–248, Feb 2000.

[6] Jong-Min Kim, Yoon-Sung Jung, Engin A Sungur, Kap-Hoon Han, Changyi Park, and Insuk Sohn. A copula method for modeling directional dependence of genes. *BMC Bioinformatics*, 9:225, 2008.

[7] Yinglei Lai, Baolin Wu, Liang Chen, and Hongyu Zhao. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics*, 20(17):3146–3155, Nov 2004.

[8] K-C. Li and S. Yuan. A functional genomic study on nci's anticancer drug screen. *Pharmacogenomics J*, 4(2):127–135, 2004.

[9] Ker-Chau Li. Genome-wide coexpression dynamics: theory and application. *Proc Natl Acad Sci U S A*, 99(26):16875–16880, Dec 2002.

[10] Ker-Chau Li, Ching-Ti Liu, Wei Sun, Shinsheng Yuan, and Tianwei Yu. A system for enhancing genome-wide coexpression dynamics study. *Proc Natl Acad Sci U S A*, 101(44):15561–15566, Nov 2004.

[11] P. Mccullagh and J. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC, August 1989.

[12] Roger B. Nelsen. *An introduction to copulas*. Springer, New York, 1999.

[13] M. C. Paik. Parametric variance function estimation for nonnormal repeated measurement data. *Biometrics*, 48(1):19–30, Mar 1992.

[14] G. Sherlock. Analysis of large-scale gene expression data. *Curr Opin Immunol*, 12(2):201–205, Apr 2000.

[15] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, Dec 1998.

[16] Jun Yan and Jason Fine. Estimating equations for association structures. *Stat Med*, 23(6):859–74; discussion 875–7,879–80, Mar 2004.

[17] Jiexin Zhang, Yuan Ji, and Li Zhang. Extracting three-way gene interactions from microarray data. *Bioinformatics*, 23(21):2903–2909, Nov 2007.

Appendix A

Proof of Theorem 1

Using Stein's lemma, with $X_3 \sim N(0,1)$

$$\begin{aligned} \mathrm{GLA}(X_1, X_2 \mid X_3) &= E[h'(X_3)] = E[\rho(X_1, X_2 \mid X_3)X_3]. \\ \rho(X_1, X_2 \mid X_3) &= \frac{E(X_1X_2 \mid X_3 = x_3) - E(X_1 \mid X_3)E(X_2 \mid X_3)}{\sigma_1(X_1 \mid X_3)\sigma_2(X_2 \mid X_3)}, \end{aligned}$$

Notice that the marginal variance for example of X_2 equals to:

$$\sigma_2(X_2) = Var[E(X_2 \mid X_3)] + E[\sigma_2(X_2 \mid X_3)].$$

Given that $E(X_2 \mid X_3) \perp X_3$ and $\sigma_2(X_2 \mid X_3) \perp X_3$, we have $Var[E(X_2 \mid X_3)] = 0$, and $\sigma_2(X_2 \mid X_3) = \sigma_2(X_2) = 1$. Similarly, $\sigma_1(X_1 \mid X_3) = \sigma_1(X_1) = 1$.

$E[\rho(X_1, X_2 \mid X_3)X_3] = E[X_3E(X_1X_2 \mid X_3 = x_3)] - E(X_1 \mid X_3)E(X_2 \mid X_3)E(X_3) = E(X_1X_2X_3)$.

Appendix B

When $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \alpha_3 = \alpha_4 = \alpha_5 = 0$, $\mathrm{GLA}(X_1, X_3 \mid X_2)$ can be expressed as follows:

$$\mathrm{GLA}(X_1, X_3|X_2) = \int_{X_2} \frac{X_2(X_2 - \mu_2)\cdot A(\beta_5)}{\sqrt{[B(\beta_5)(X_2 - \mu_2)^2 + C(\beta_5)(X_2 - \mu_2) + D(\beta_5]}} f(X_2)dX_2,$$

where $A(\beta_5), B(\beta_5), C(\beta_5)$ and $D(\beta_5)$ are functions of β_5, and they are constant with respect to X_2. $A(\beta_5) = \int_{x_3} \frac{e^{\alpha_5+\beta_5X_3-1}}{e^{\alpha_5+\beta_5x_3+1}}(X_3 - \mu_3)N(X_3; 0, 1)dX_3$. We use $N(X_3; 0, 1)$ to denote the standard normal density for random variable X_3.
$B(\beta_5) = \int_{X_3} N(X_3; 0, 1)(\frac{e^{\alpha_5+\beta_5X_3-1}}{e^{\alpha_5+\beta_5x_3+1}})^2dX_3 - [\int_{X_3} N(X_3; 0, 1)\frac{e^{\alpha_5+\beta_5X_3-1}}{e^{\alpha_5+\beta_5x_3+1}}dX_3]^2$,
$C(\beta_5) = \int_{X_3} N(X_3; 0, 1)\cdot 2\frac{e^{\alpha_5+\beta_5X_3-1}}{e^{\alpha_5+\beta_5x_3+1}}dX_3 - 2\cdot[\int_{X_3} N(X_3; 0, 1)dX_3][\int_{X_3} N(X_3; 0, 1)\frac{e^{\alpha_5+\beta_5X_3-1}}{e^{\alpha_5+\beta_5x_3+1}}dX_3]$,
and $D(\beta_5) = \int_{X_3} N(X_3; 0, 1)(1-(\frac{e^{\alpha_5+\beta_5X_3-1}}{e^{\alpha_5+\beta_5x_3+1}})^2)dX_3 + \{\int_{X_3} N(X_3; 0, 1)dX_3 - [\int_{X_3} N(X_3; 0, 1)dX_3]^2\}$.
And $f(X_2)$ is the marginal distribution of X_2. A detailed calculation is as follows.
By definition

$$\mathrm{corr}(X_1, X_3|X_2) = \frac{E(X_1X_3|X_2) - E(X_1|X_2)E(X_3|X_2)}{\sqrt{\sigma_1^2(X_1|X_2)\sigma_3^2(X_3|X_2)}}.$$

where $\mathrm{corr}(X_1, X_3|X_2)$ denotes conditional correlation between X_1, X_3 given X_2. $\sigma_1^2(X_1|X_2)$ is the variance of X_1 given X_2, and $\sigma_3^2(X_3|X_2)$ is the variance of X_3 given X_2. To derive $\mathrm{corr}(X_1, X_3|X_2)$, we first calculated the distributions of X_2 given X_3, the distribution of X_1 given X_2 and the distribution of X_1 given X_2 and X_3, $f(X_2|X_3), f(X_1|X_2)$, and $f(X_1|X_2, X_3)$. Given that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \alpha_3 = \alpha_4 = \alpha_5 = 0$, hence $\mu_1 = \mu_2 = \mu_3 = 0$

and $\sigma_1 = \sigma_2 = \sigma_3 = 1$.

$$\begin{aligned}
f(X_2|X_3) &= \int_{X_1} f(X_1, X_2|X_3)dX_1 = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}(u^2+v^2-2\rho uv)}dv. \\
\text{where } u &= \frac{X_2-\mu_2}{\sigma_2}, v = \frac{X_1-\mu_1}{\sigma_1}. \\
\text{And } u^2+v^2-2\rho uv &= (v-\rho u)^2 + u^2(1-\rho^2). \\
f(X_2|X_3) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{\frac{-u^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}(v-\rho u)^2}dv. \\
&\quad \text{integrate a normal with mean } \rho u \text{ and variance } (1-\rho^2). \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{\frac{-u^2}{2}} \sqrt{2\pi}\sqrt{1-\rho^2}, \\
&= \frac{1}{\sqrt{2\pi}\sigma_2} e^{\frac{-(X_2-\mu_2)^2}{2\sigma_2}}, \\
&= N(X_2; \mu_2, \sigma_2^2), \\
&= N(X_2; 0, 1).
\end{aligned}$$

$f(X_3|X_2)$ can be written as follows:

$$\begin{aligned}
f(X_3|X_2) &= \frac{f(X_2, X_3)}{f(X_2)} \\
&= \frac{\int_{X_1} f(X_1, X_2|X_3)dX_1 f(X_3)}{\int_{X_3}\int_{X_1} f(X_1, X_2|X_3)dX_1 f(X_3)dX_3} \\
&= \frac{N(X_2; 0, 1)N(X_3; 0, 1)}{N(X_2; 0, 1)} \\
&= N(X_3; 0, 1)
\end{aligned}$$

And $f(X_1|X_2, X_3)$ can be calculated as:

$$f(X_1|X_2, X_3) = \frac{f(X_1X_2|X_3)}{f(X_2|X_3)} = N(X_1; \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2), \sigma_1^2(1-\rho^2)).$$

$$
\begin{aligned}
\mathrm{cov}(X_1, X_3|X_2) &= E(X_1 \cdot X_3|X_2) - E(X_1|X_2)E(X_3|X_2) \\
&= \int_{X_3}\int_{X_1} X_1 X_3 f(X_1, X_3|X_2) dX_1 dX_3 - E(X_1|X_2)E(X_3|X_2) \\
&= \int_{X_3} X_3 \int_{X_1} [X_1 f(X_1|X_2, X_3) dX_1] f(X_3|X_2) dX_3 - E(X_1|X_2)E(X_3|X_2) \\
&= \int_{X_3} X_3 [\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2)] f(X_3|X_2) - E(X_1|X_2)E(X_3|X_2) \\
&= [\mu_1 \int_{X_3} X_3 f(X_3|X_2) dX_3] + [(X_2 - \mu_2)\frac{\sigma_1}{\sigma_2}\int_{X_3} \frac{e^{\alpha+\beta X_3} - 1}{e^{\alpha+\beta X_3} + 1} X_3 f(X_3|X_2) dX_3] \\
&- E(X_1|X_2)E(X_3|X_2), \\
E(X_1|X_2) &= \int_{X_1} X_1 f(X_1|X_2) dX_1 \\
&= \int_{X_1} X_1 [\int_{X_3} f(X_1|X_3, X_2) f(X_3|X_2) dX_3] dX_1 \\
&= \int_{X_3}\int_{X_1} \{X_1 N[X_1; \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2), \sigma_1^2(1 - \rho^2)] dX_1\} f(X_3|X_2) dX_3 \\
&= \int_{X_3} [\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2)] N(X_3; \mu_3, \sigma_3^2) dX_3 \\
&= \mu_1 + \frac{\sigma_1}{\sigma_2}(X_2 - \mu_2)\int_{X_3} \frac{e^{\alpha+\beta X_3} - 1}{e^{\alpha+\beta X_3} + 1} N(X_3; \mu_3, \sigma_3^2) dX_3.
\end{aligned}
$$

$$
\begin{aligned}
E(X_3|X_2) &= \int_{X_3} X_3 f(X_3|X_2) dX_3 \\
&= \int_{X_3} X_3 N(X_3; \mu_3, \sigma_3^2) dX_3 \\
&= \mu_3 \\
\mathrm{cov}(X_1, X_3|X_2) &= \mu_1\mu_3 + [(X_2 - \mu_2)\frac{\sigma_1}{\sigma_2}\int_{X_3} \frac{e^{\alpha+\beta X_3} - 1}{e^{\alpha+\beta X_3} + 1} X_3 N(X_3; \mu_3, \sigma_3^2) dX_3], \\
&- \mu_3[\mu_1 + \frac{\sigma_1}{\sigma_2}(X_2 - \mu_2)\int_{X_3} \frac{e^{\alpha+\beta X_3} - 1}{e^{\alpha+\beta X_3} + 1} N(X_3; \mu_3, \sigma_3^2) dX_3] \\
&= (X_2 - \mu_2)\frac{\sigma_1}{\sigma_2}\int_{X_3} \frac{e^{\alpha+\beta X_3} - 1}{e^{\alpha+\beta X_3} + 1}(X_3 - \mu_3) N(X_3; \mu_3, \sigma_3^2).
\end{aligned}
$$

$$
\mathrm{cov}(X_1, X_3|X_2) = (X_2 - \mu_2)\frac{\sigma_1}{\sigma_2}\int_{X_3} \frac{e^{\alpha+\beta X_3} - 1}{e^{\alpha+\beta X_3} + 1}(X_3 - \mu_3) N(X_3; \mu_3, \sigma_3^2) = A(\beta_5)(X_2 - \mu_2).
$$

where $A(\beta_5) = \int_{X_3} \frac{e^{\alpha+\beta X_3} - 1}{e^{\alpha+\beta X_3} + 1}(X_3 - \mu_3) N(X_3; \mu_3, \sigma_3^2) dX_3$ is a constant with respect to X_2.
In the denominator of $\mathrm{corr}(X_1 X_3|X_2)$, we derive $\mathrm{var}(X_1|X_2)$ and $\mathrm{var}(X_2|X_2)$ as follows.

$$
V(X_1|X_2) = E(X_1^2|X_2) - [E(X_1|X_2)]^2.
$$

$$
\begin{aligned}
E(X_1^2|X_2) &= \int_{X_1} X_1^2 f(X_1|X_2)dX_1 = \int_{X_1} X_1^2[\int_{X_3} f(X_1|X_3,X_2)f(X_3|X_2)dX_3]dX_1. \\
&= \int_{X_3} N(X_3;0,1)\int_{X_1} X_1^2 N[X_1;\mu^*,\sigma^{*2}]dX_1dX_3. \\
\text{where } \mu^* &= \mu_1+\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2), \sigma^{*2}=\sigma_1^2(1-\rho^2). \\
E(X_1^2|X_2) &= \int_{X_3} N(X_3;0,1)\{\sigma_1^2(1-\rho^2)+[\mu_1+\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)]^2dX_3\}, \\
&= \int_{X_3} N(X_3;0,1)\sigma_1^2(1-\rho^2)dX_3+\int_{X_3} N(X_3;0,1)[\mu_1+\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)]^2dX_3. \\
&= \int_{X_3} N(X_3;0,1)\sigma_1^2(1-\rho^2)dX_3+\int_{X_3} N(X_3;0,1)[\mu_1^2+2\mu_1\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2) \\
&+ \rho^2\frac{\sigma_1^2}{\sigma_2^2}(X_2-\mu_2)^2]dX_3. \\
&= \int_{X_3} N(X_3;0,1)\sigma_1^2(1-\rho^2)dX_3+\int_{X_3}\mu_1^2N(X_3;0,1)dX_3 \\
&+ (X_2-\mu_2)\int_{X_3} N(X_3;0,1)[2\mu_1\rho\frac{\sigma_1}{\sigma_2}]dX_3 \\
&+ (X_2-\mu_2)^2\int_{X_3} N(X_3;0,1)(\rho^2\frac{\sigma_1^2}{\sigma_2^2})dX_3.
\end{aligned}
$$

$$
\begin{aligned}
[E(X_1|X_2)]^2 &= [\int_{X_1} X_1 f(X_1|X_2)dX_1]^2 = \{\int_{X_1}[\int_{X_3} f(X_1|X_3,X_2)f(X_3|X_2)dX_3]dX_1\}^2 \\
&= [\int_{X_3} N(X_3;0,1)(\int_{X_1} f(X_1|X_2,X_3)f(X_3|X_2)dX_1)dX_3]^2, \\
&= [\int_{X_3} N(X_3;0,1)(\int_{X_1} N[X_1;\mu^*,\sigma^{*2}]dX_1)dX_3]^2 \\
&= \{\int_{X_3} N(X_3;0,1)[\mu_1+\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)]dX_3\}^2 \\
&= \{\int_{X_3} N(X_3;0,1)\mu_1 dX_3 + \int_{X_3} N[X_3]\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)dX_3\}^2 \\
&= [\int_{X_3} N(X_3;0,1)\mu_1 dX_3]^2 + (X_2-\mu_2)\cdot 2[\int_{X_3}[N(X_3;0,1)\mu_1 dX_3][\int_{X_3} N(X_3;0,1)\rho\frac{\sigma_1}{\sigma_2}dX_3]] \\
&+ (X_2-\mu_2)^2[\int_{X_3} N(X_3;0,1)\rho\frac{\sigma_1}{\sigma_2}dX_3]^2. \\
V(X_1|X_2) &= E(X_1^2|X_2) - [E(X_1|X_2)]^2 \\
&= \int_{X_3} N(X_3;0,1)\sigma_1^2(1-\rho^2)dX_3 + \{\int_{X_3}\mu_1^2 N(X_3;0,1)dX_3 - [\int_{X_3}\mu_1 N(X_3;0,1)dX_3]^2\} \\
&+ (X_2-\mu_2)\{\int_{X_3} N(X_3;0,1)\cdot 2\mu_1\rho\frac{\sigma_1}{\sigma_2}dX_3 \\
&- 2\cdot[\int_{X_3}\mu_1 N(X_3;0,1)dX_3][\int_{X_3} N(X_3;0,1)\rho\frac{\sigma_1}{\sigma_2}dX_3]\} \\
&+ (X_2-\mu_2)^2\{\int_{X_3} N(X_3;0,1)\rho^2\frac{\sigma_1^2}{\sigma_2^2}dX_3 - [\int_{X_3} N(X_3;0,1)\rho\frac{\sigma_1}{\sigma_2}dX_3]^2\} \\
&= B(\beta_5)(X_2-\mu_2)^2 + C(\beta_5)(X_2-\mu_2) + D(\beta_5).
\end{aligned}
$$

where $B(\beta_5), C(\beta_5)$ and $D(\beta_5)$ are functions of β_5 and are constant with respect to X_2.

$$
\begin{aligned}
\text{corr}(X_1,X_3|X_2) &= \frac{\text{cov}(X_1,X_3|X_2)}{\sqrt{V(X_1|X_2)V(X_3|X_2)}} \\
\text{We have} \quad \text{cov}(X_1,X_3|X_2) &= (X_2-\mu_2)A(\beta_5) \\
V(X_1|X_2) &= B(\beta_5)(X_2-\mu_2)^2 + C(\beta_5)(X_2-\mu_2) + D(\beta_5). \\
\text{corr}(X_1,X_3|X_2) &= \frac{(X_2-\mu_2)\cdot A(\beta_5)}{\sqrt{\sigma_3^2[B(\beta_5)(X_2-\mu_2)^2 + C(\beta_5)(X_2-\mu_2) + D(\beta_5)]}}
\end{aligned}
$$

Finally, $\text{GLA}(X_1, X_3 \mid X_2)$ can be expressed as:

$$
\begin{aligned}
\text{GLA}(X_1,X_3|X_2) &= E[\text{corr}(X_1,X_3 \mid X_2)X_2], \\
&= \int_{X_2} \frac{X_2(X_2-\mu_2)\cdot A(\beta_5)}{\sqrt{\sigma_3^2[B(\beta_5))(X_2-\mu_2)^2 + C(\beta_5))(X_2-\mu_2) + D(\beta_5)]}} f(X_2)dX_2.
\end{aligned}
$$

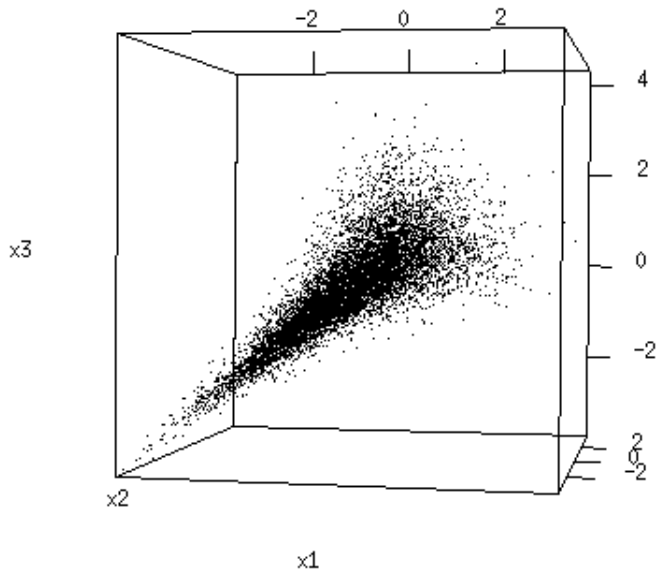


Figure 1: A three-dimensional scatter plot of data generated from Clayton copula with standard normal marginals ($\theta = 3$).

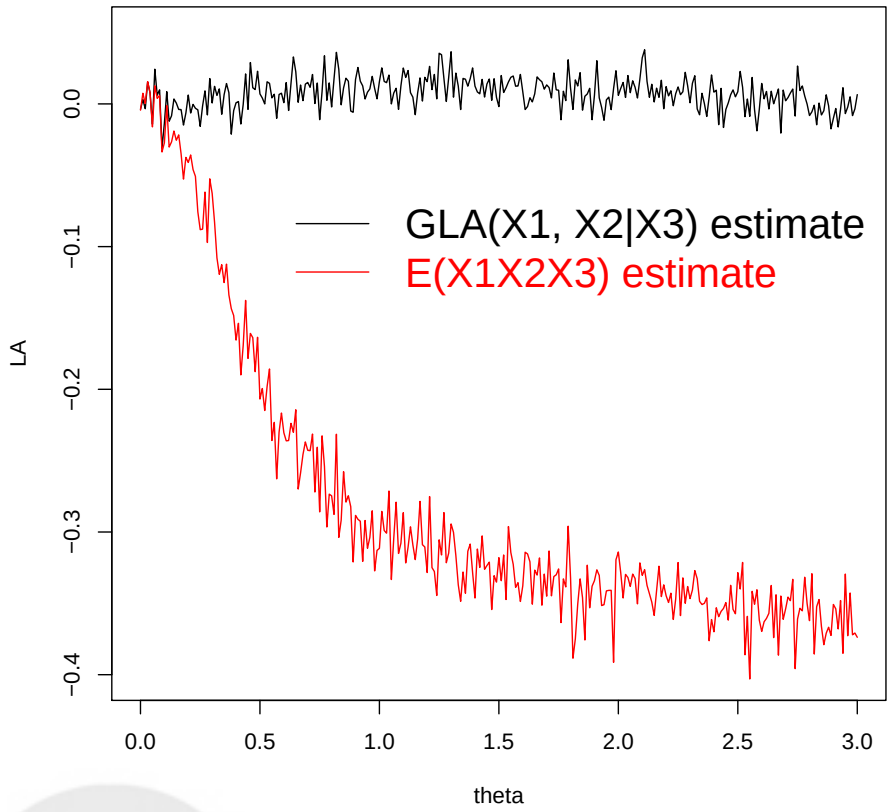


Figure 2: Differences between $\mathrm{GLA}(X_1, X_2 \mid X_3)$ and $E(X_1X_2X_3)$ estimates. Fluctuations reflect simulation noise.

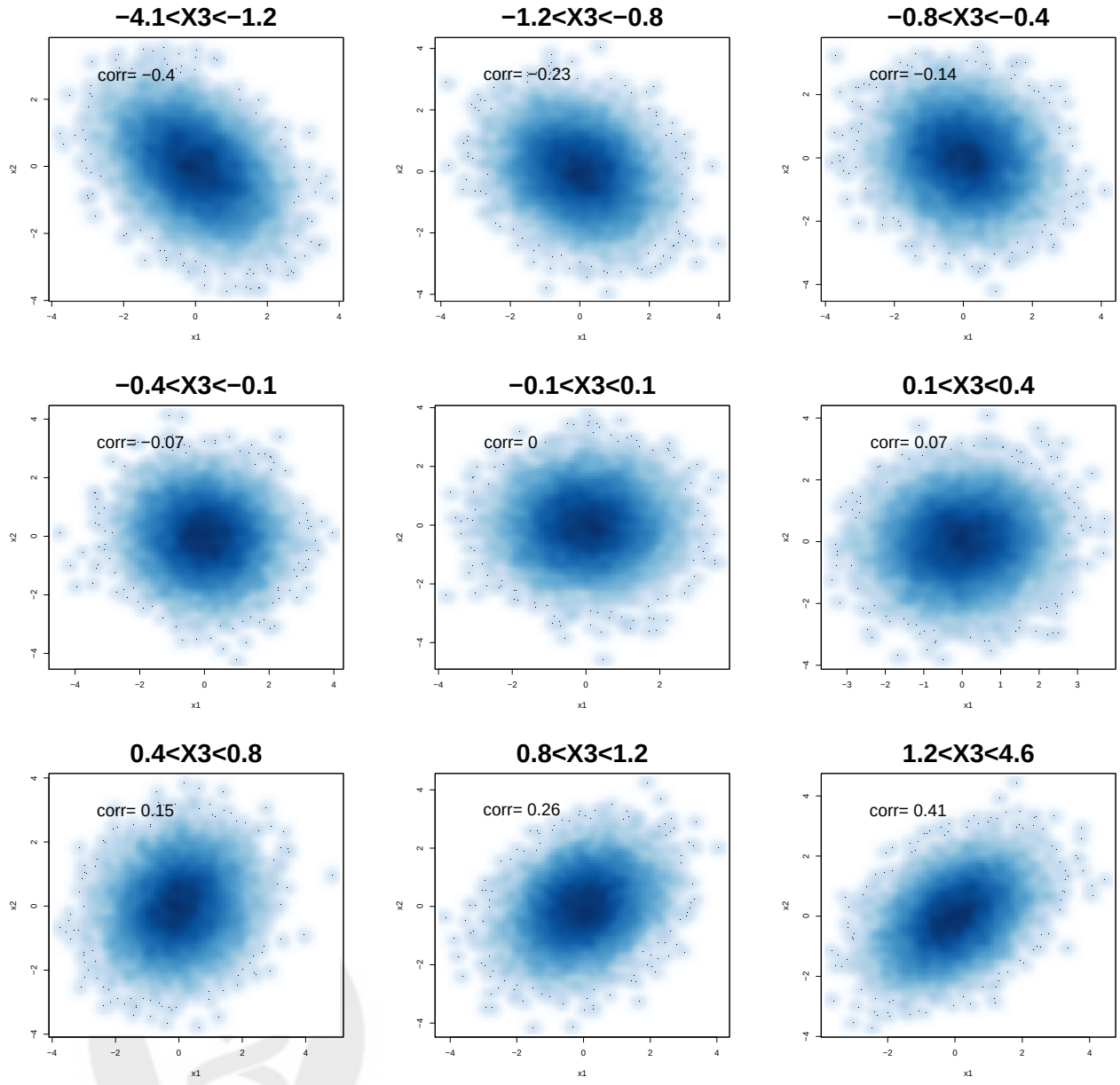


Figure 3: Conditional distributions of $(X_1, X_2|X_3)$ for varying X_3. Here $\beta_5 = 0.5$, and $\alpha_3 = \alpha_4 = \alpha_5 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

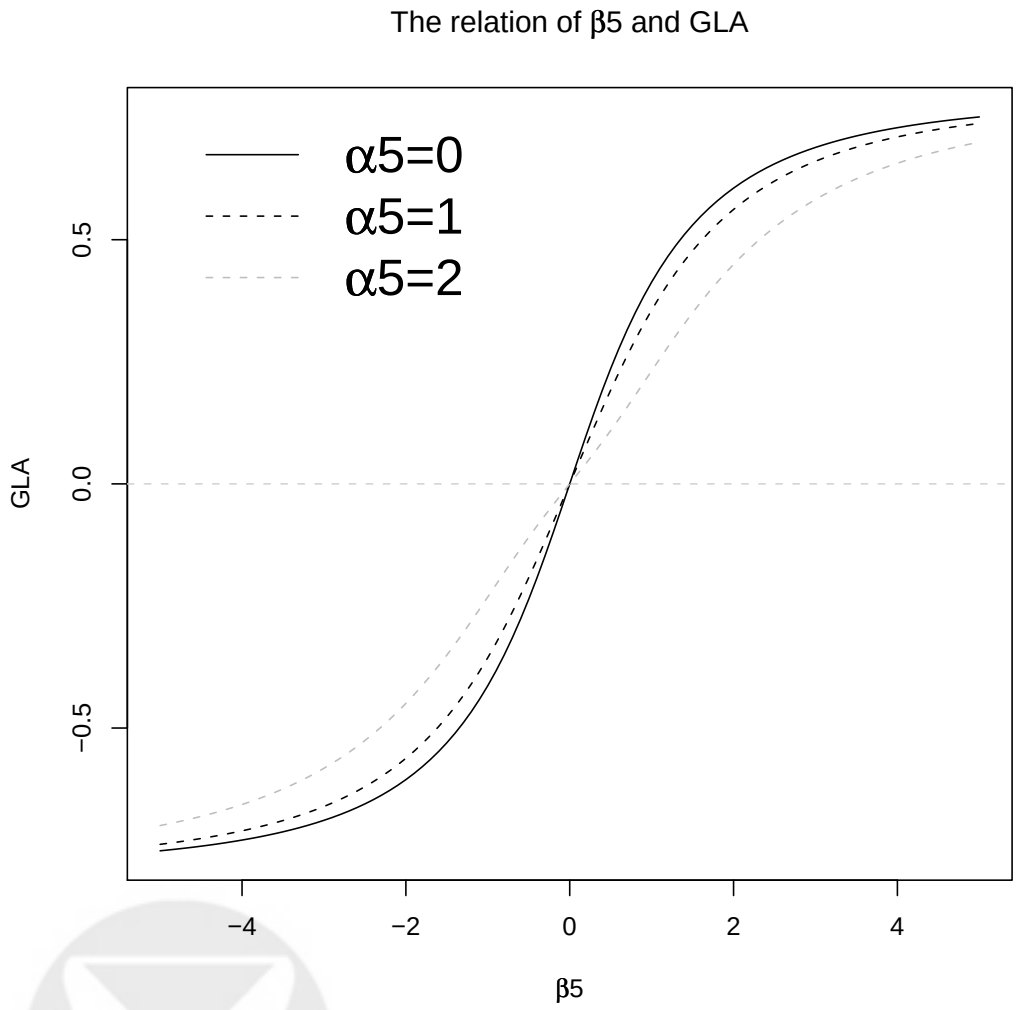


Figure 4: The relation of β_5 and GLA

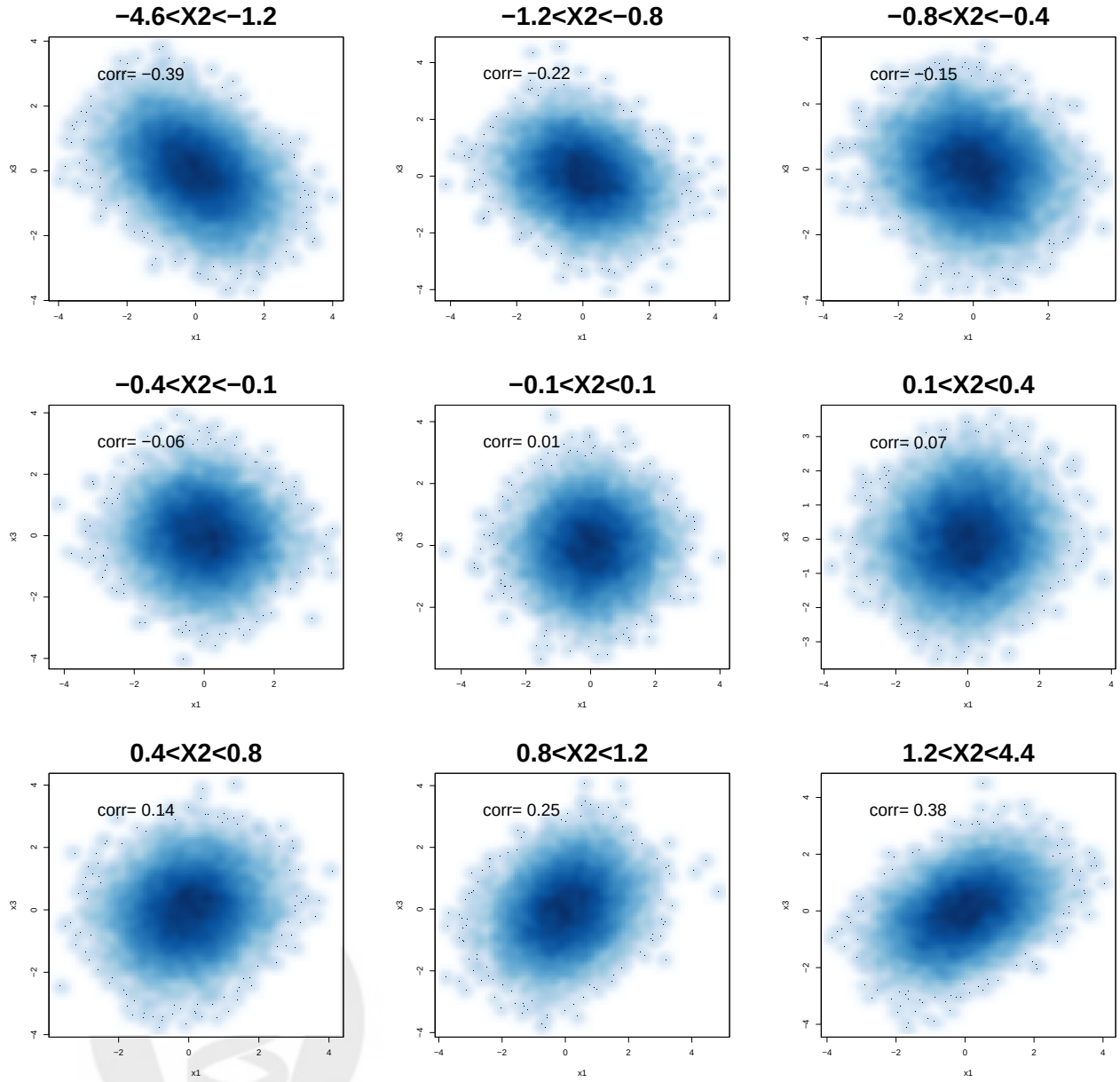


Figure 5: Conditional distributions of $(X_1, X_3 \mid X_2)$ for varying X_2.

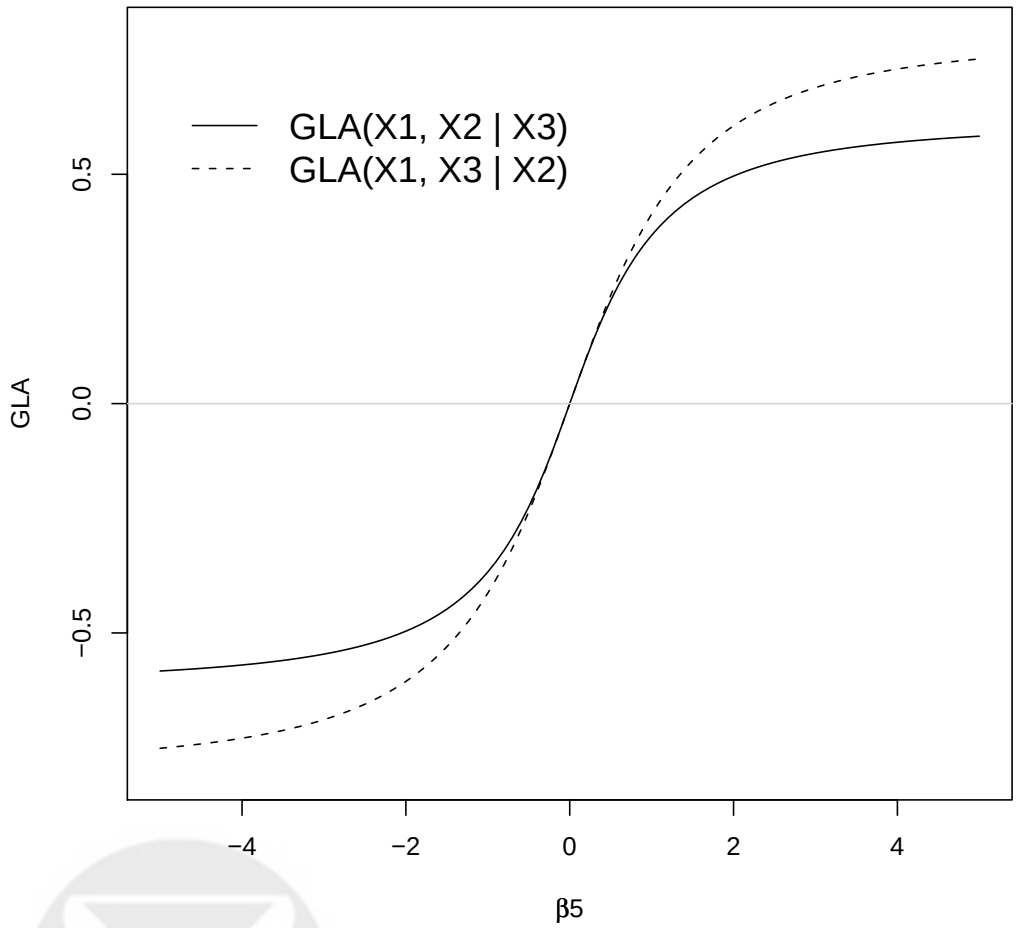


Figure 6: GLA$(X_1, X_2 \mid X_3)$ and GLA$(X_1, X_3 \mid X_2)$ in $CNM(X_1, X_2, X_3)$ with $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \alpha_3 = \alpha_4 = \alpha_5 = 0$, and varied β_5.

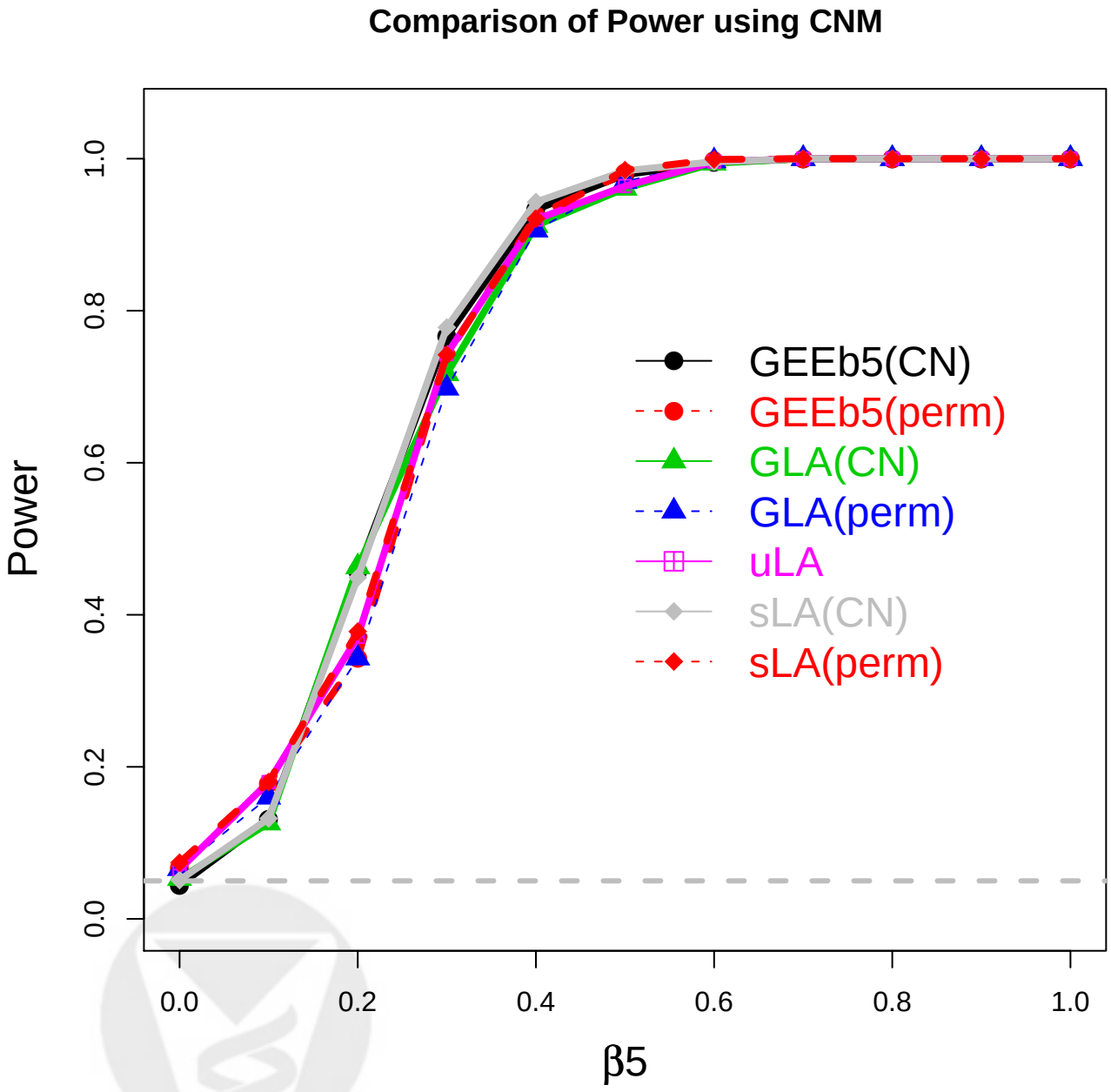


Figure 7: The power of the test statistics under scenarios 1.

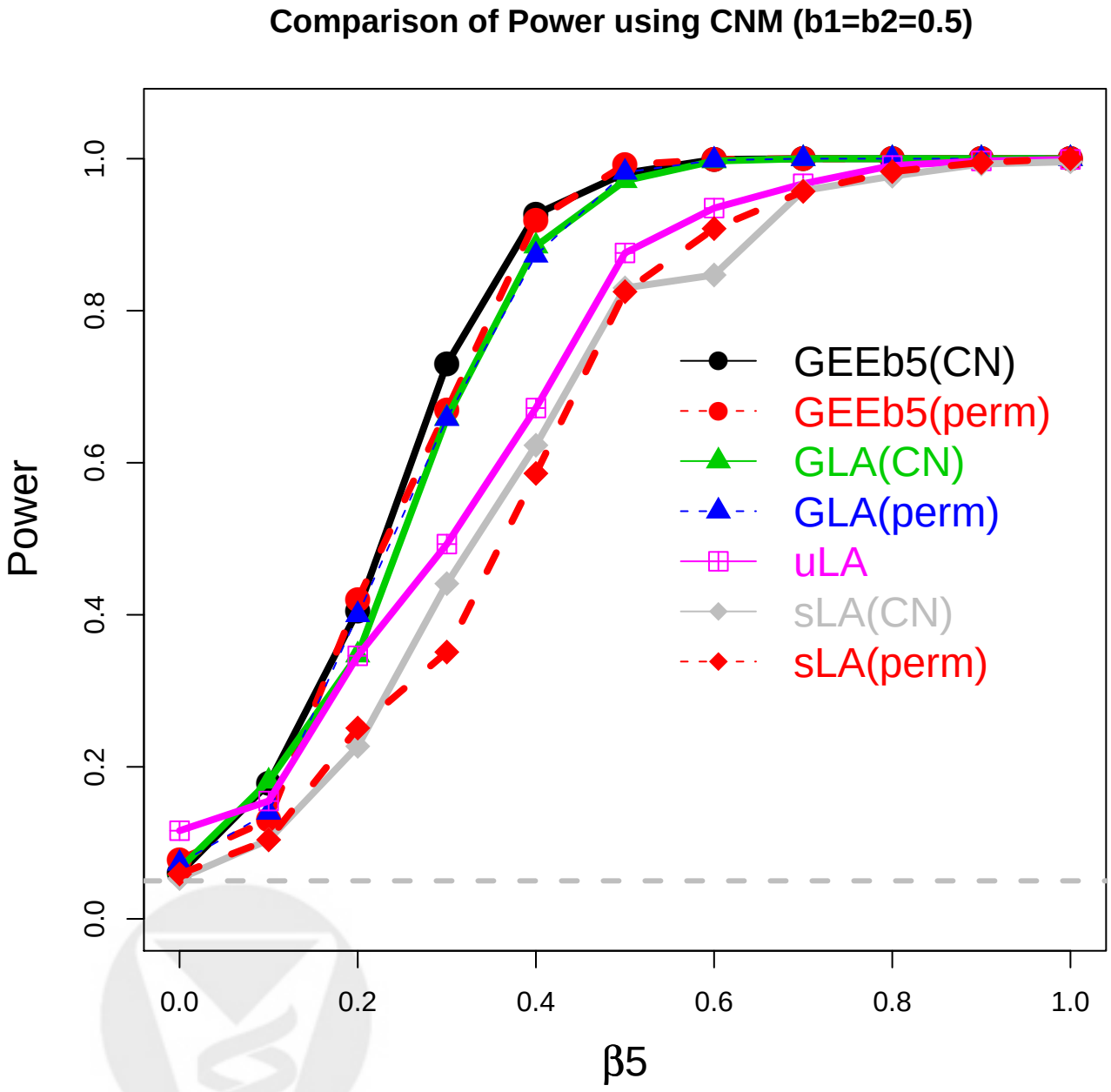


Figure 8: The power of the test statistics under scenarios 2.

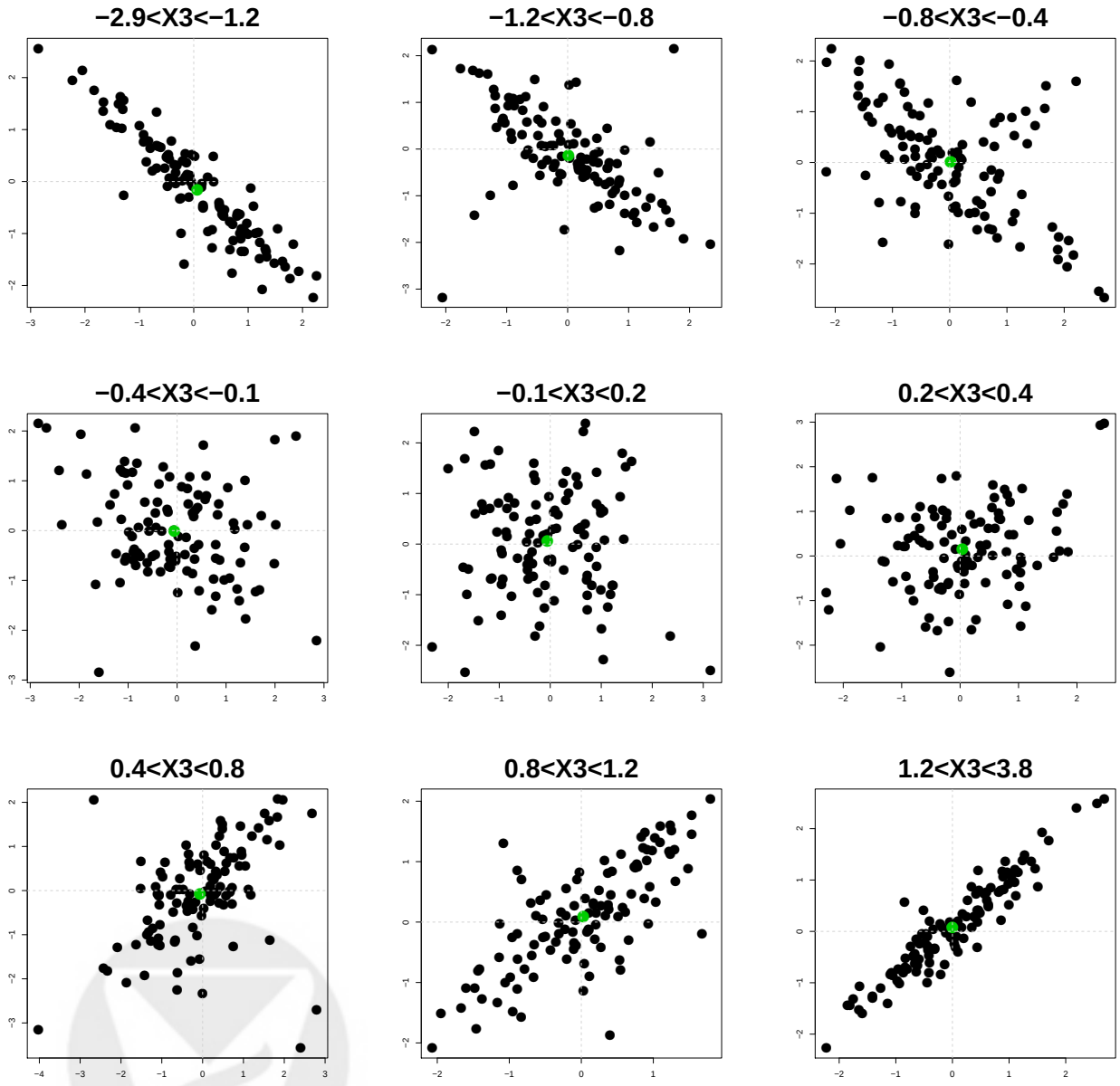


Figure 9: Conditional distributions of $(X_1, X_2|X_3)$ for varying X_3. Here X_1 and X_2 follows T copula with standard normal margins with 1 degree of freedom, and $\beta = 2$.

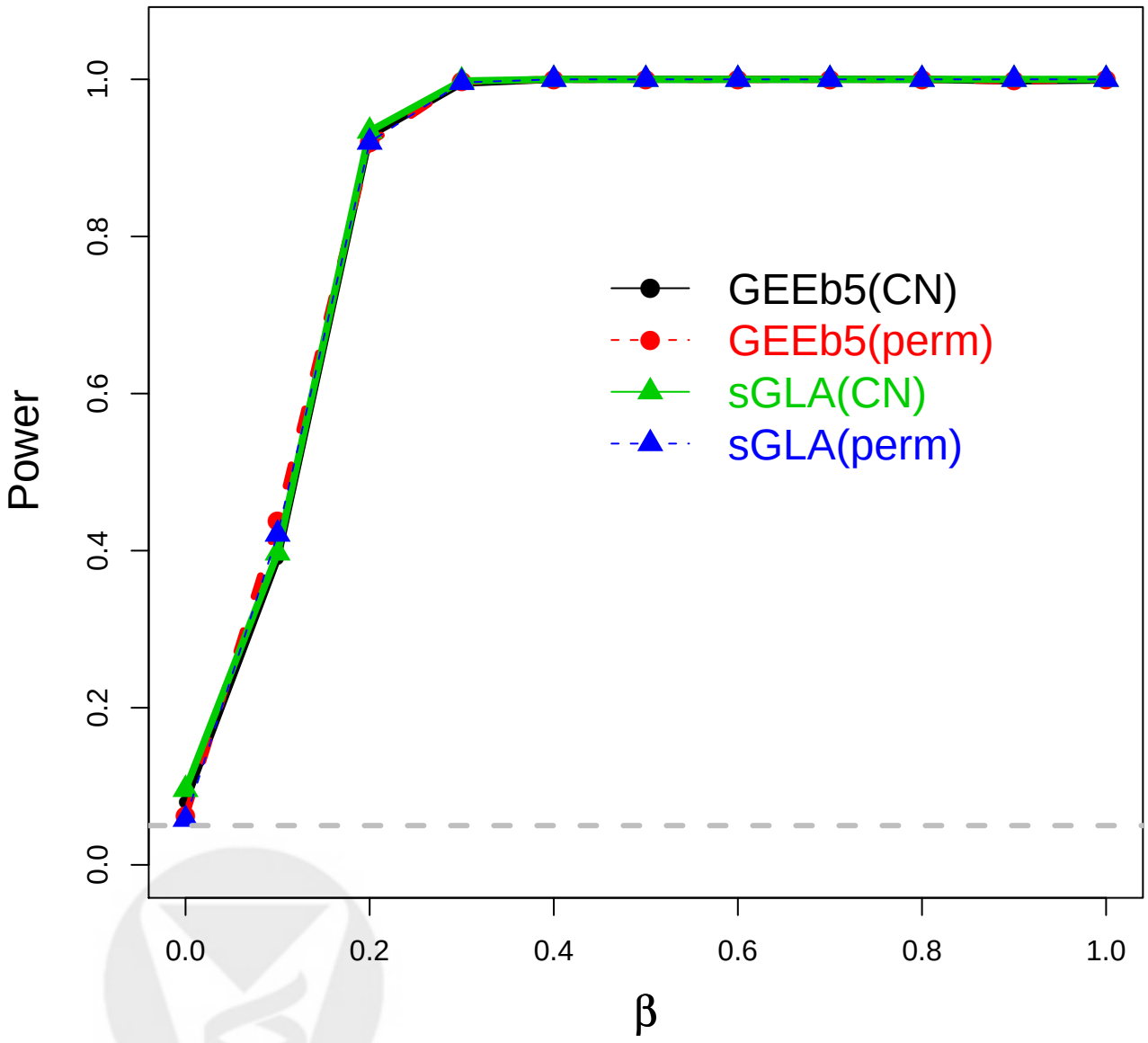
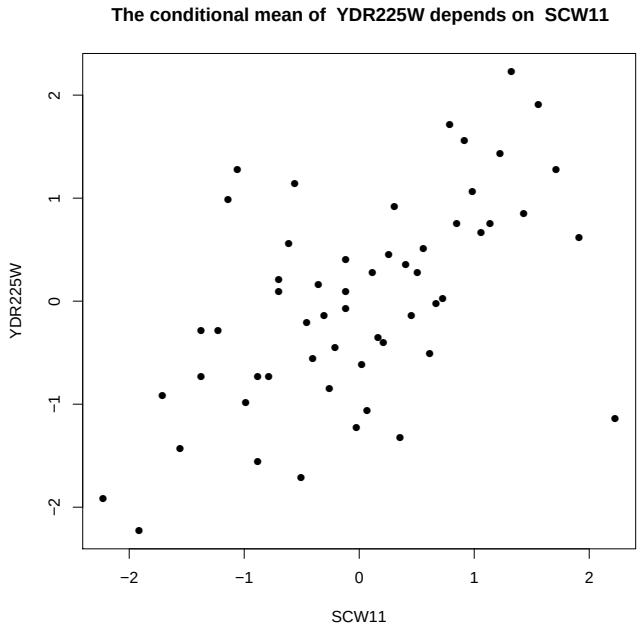


Figure 10: The power of $GEEb_5$ and sGLA under T copula with 1 degree of freedom (scenario 3).



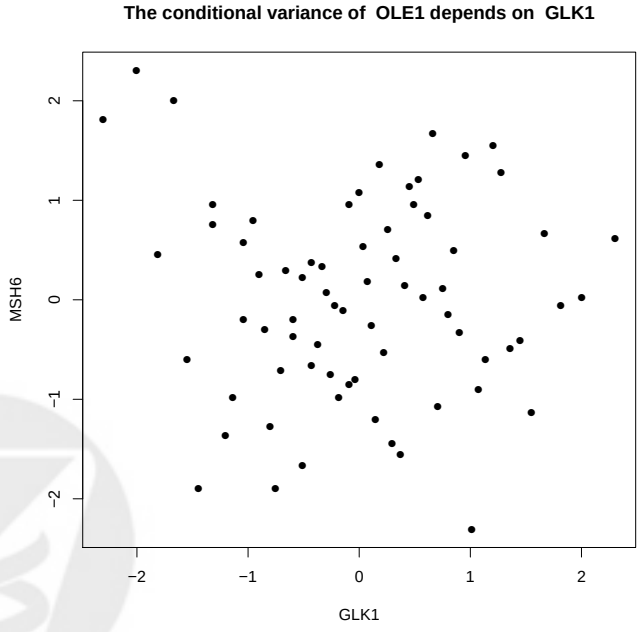


Figure 11: Examples from triplet # 4 and 7 that show how conditional means and variances can depend on the third gene (x-axis).