Biometrics 72, 1358–1368
December 2016
DOI: 10.1111/biom.12516

PCAN: Probabilistic Correlation Analysis of Two Non-normal Data Sets

Roger S. Zoh,[1,*] Bani Mallick,[2] Ivan Ivanov,[3] Veera Baladandayuthapani,[4] Ganiraju Manyam,[4] Robert S. Chapkin,[5] Johanna W. Lampe,[6] and Raymond J. Carroll[2]

[1]Department of Epidemiology and Biostatistics, Texas A&M University, College Station, Texas, U.S.A.
[2]Department of Statistics, Texas A&M University, College Station, Texas, U.S.A.
[3]Department of Veterinary Medicine and Biomedical Sciences, Texas A&M University, Texas, U.S.A.
[4]The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Houston, Texas, U.S.A.
[5]Program in Integrative Nutrition and Complex Diseases, Texas A&M University, Texas, U.S.A.
[6]Department of Epidemiology, University of Washington and the Fred Hutchinson Cancer Research Center Seattle, Washington, U.S.A.
email: rszoh@sph.tamhsc.edu

Summary. Most cancer research now involves one or more assays profiling various biological molecules, e.g., messenger RNA and micro RNA, in samples collected on the same individuals. The main interest with these genomic data sets lies in the identification of a subset of features that are active in explaining the dependence between platforms. To quantify the strength of the dependency between two variables, correlation is often preferred. However, expression data obtained from next-generation sequencing platforms are integer with very low counts for some important features. In this case, the sample Pearson correlation is not a valid estimate of the true correlation matrix, because the sample correlation estimate between two features/variables with low counts will often be close to zero, even when the natural parameters of the Poisson distribution are, in actuality, highly correlated. We propose a model-based approach to correlation estimation between two non-normal data sets, via a method we call Probabilistic Correlations ANalysis, or PCAN. PCAN takes into consideration the distributional assumption about both data sets and suggests that correlations estimated at the model natural parameter level are more appropriate than correlations estimated directly on the observed data. We demonstrate through a simulation study that PCAN outperforms other standard approaches in estimating the true correlation between the natural parameters. We then apply PCAN to the joint analysis of a microRNA (miRNA) and a messenger RNA (mRNA) expression data set from a squamous cell lung cancer study, finding a large number of negative correlation pairs when compared to the standard approaches.

Key words: Canonical correlation analysis; Correlation; Generalized linear models; Poisson regression; RNA-sequencing

1. Introduction

We develop methods to analyze data concerning squamous cell lung cancers from The Cancer Genome Atlas consortium (https://tcga-data.nci.nih.gov/tcga/). MicroRNAs (miRNA) refer to highly conserved, short non-coding RNAs which have important roles in many biological processes such as cellular differentiation, apoptosis, cell proliferation, and development. They regulate protein production by repressing their putative messenger RNA (mRNA) targets. Hence, high expression of a miRNA is often associated with reduced expression of its gene targets. A single miRNA can have many putative mRNA targets and a single mRNA can be regulated by many miRNAs. The role of miRNA and mRNA interaction in many disease-related regulatory pathways, is well established (Shah et al., 2011).

In our analysis, the data are fragment (read) counts from next generation sequencing using the Illumina Hiseq platform. The data set contains 50 selected miRNAs based on prior biological knowledge and 66 mRNAs of interests, jointly obtained from 150 independent samples. Of the 66 mRNAs considered, 23 have an average number of read counts between 1 and 2, which can be considered as lowly expressed, relative to the average number of reads for the other genes. Standard correlation estimates based on low counts tend to severely underestimate strong negative correlations, which are expected in the expression between a miRNA and its target mRNA. Our main goal is to quantify the dependency between the miRNA read counts and the mRNA read counts, even when the counts are small. More discussion of the data is given in Section 3.2, which also contains the results of analysis of the data.

By way of background, the decreasing cost of DNA sequencing makes affordable for more biologists to run multiple assays profiling different biological molecules/platforms, such as microRNA (miRNA) and messenger RNA (mRNA), on the same sample. The primary objective when collecting these various data sets on the same sample is to identify possible sets of variables or features active in explaining the

dependence within and across data sets. It is well known that genes work in a complex network in connection with other genes and biological molecules. Thus, a better understanding of disease progression requires the characterization of these pathways or networks. For example, Shah et al. (2011) characterize the repressive effect of some miRNAs on their putative target messenger mRNAs under various treatment regimes, see also Ren et al. (2009). Canonical correlation analysis is a standard statistical approach used to uncover the relationship between two blocks of variables. Because in most genomic data sets, few variables or features are expected to be correlated, regularized canonical correlation analysis has also been extensively considered (see González et al., 2008; Lê Cao et al., 2009; Witten et al., 2009). However, in most of the references provided on the canonical correlation analysis, the data generating model is often not clearly specified.

Tipping and Bishop (1999) and Bach and Jordan (2006) provided a probabilistic foundation for principal component analysis and the canonical correlation analysis. Their work established a link between standard canonical correlation analysis and a latent factor model assuming a conditional normal distribution for the block of variables (data sets), and also paved the way for the development of a Bayesian version of canonical correlation analysis through the latent variable machinery. Various authors have addressed modeling issues of block variables in the Bayesian setting assuming various conditional distributions for the observed data, see for example Klami and Kaski (2007), Archambeau and Bach (2008), Virtanen et al. (2011) for a review.

We consider the joint analysis or correlation estimation between two genomic block variables associated with data from one of the next-generation (Next-Gen) sequencing platforms. Next-Gen genomic data sets raise two main modeling issues: (i) they are integer data for which the assumptions of the standard canonical correlation analysis are violated and hence may perform poorly; and (ii) a large proportion of these counts are very small and standard correlation estimate based on these counts can sometimes be misleading, because a correlation of zero between two count variables does not necessarily imply independence. In fact, Whitt (1976) and Shin and Pasupathy (2007) show that the correlation between two Poisson variables with low counts is restricted in a much narrower interval than $(-1, 1)$, even in the case where the natural parameters are perfectly correlated. To see this, consider two variables $\mathbf{u} \sim \text{Pois}\{\exp(\lambda_1)\}$ and $\mathbf{v} \sim \text{Pois}\{\exp(\lambda_2)\}$, where $\lambda_1 \sim \text{Normal}(0, 1)$ and $\lambda_2 = 1 - \lambda_1$. Although $\text{corr}(\lambda_1, \lambda_2) = -1$, the $\text{corr}(\mathbf{u}, \mathbf{v})$ is ≈ -0.31, which is a weak correlation. Further, assume now that $\lambda_2 = 1 - .01\lambda_1$ and the correlation is $\text{corr}(\mathbf{u}, \mathbf{v})$ is -0.013, which is a much weaker correlation estimate, although $\mathbf{u}$ and $\mathbf{v}$ are perfectly negatively correlated. Because miRNAs and their putative targets mRNA are expected to be negatively correlated, and correlation based on the raw counts, as computed by current standard approaches, may fail to reveal the strength of the association and cause the investigator to dismiss its presence. Hence, we propose to measure the strength of association at the natural parameter level instead.

Thus, we consider the estimation of the dependency between two Next-Gen genomic data sets, especially for small counts, in the form of correlation using a novel Bayesian factor model. The article is organized as follows: we discuss the model along with estimation in Section 2. Simulation results and data analysis are given in Section 3. Discussion and concluding remarks are given in Section 4. Web-based Supplementary Material includes additional figures and tables related to the data analysis in Section 3.2, the technical details of the MCMC sampling, and R and Rstan (Stan Development Team, 2013) programs for running the analyses.

2. Method

2.1. *Model*

Although this work was primarily motivated by the need to estimate the correlation between features/genes in two next-generation gene expression data sets, we propose a more general model in a generalized linear model (GLM) framework. Let $\mathbf{x}_{\cdot j} = (x_{1j}, \cdots, x_{pj})^{\mathrm{T}}$ be a vector of length p and $\mathbf{y}_{\cdot j} = (y_{1j}, \cdots, y_{qj})^{\mathrm{T}}$, another vector of length q, both denoting co-occurring data vectors obtained from measurements on the jth individual/sample. Let x_{ij} represent the observed value for the ith feature/variable measured on the jth individual in a set of p measured features/variables (e.g., mRNA). Let y_{kj} denote the observed value for the kth feature measured also on the jth individual in another set of q measured characteristics (e.g., miRNA), with $i = 1, \cdots, p$, $j = 1, \cdots, N$, $k = 1, \cdots, q$, and N denotes the sample size. We write $\mathbf{X} = (\mathbf{x}_{\cdot 1}, \cdots, \mathbf{x}_{\cdot N})$ and $\mathbf{Y} = (\mathbf{y}_{\cdot 1}, \cdots, \mathbf{y}_{\cdot N})$. Throughout this article, we use boldface for vector values and matrices. Also, for a matrix $\mathbf{U}$, $\mathbf{U}_{i\cdot}$ and $\mathbf{U}_{\cdot j}$ denote the ith row or jth column of $\mathbf{U}$, respectively; $U_{\cdot\cdot}$ is used to denote a scalar variable. Each individual data vector is assumed to follow a conditional exponential family distribution and we consider a generalized linear model (GLM).

Let $\mathbf{F}_x(\cdot)$ and $\mathbf{F}_y(\cdot)$ be distribution functions with density/mass function from the natural parameter exponential family. A vector $\mathbf{x} \in \Re^m$ has a conditional distribution member of the exponential family if $\mathbf{f}(\mathbf{x} \mid \boldsymbol{\theta}) = c(\mathbf{x}) g(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{m} h(x_i)\theta_i\right\}$. We model both variable blocks (data matrices) individually as

$$
\begin{aligned}
&X_{ij} \mid \theta_{ij} \sim \mathbf{F}_x(\theta_{ij}), \quad Y_{kj} \mid \lambda_{kj} \sim \mathbf{F}_y(\lambda_{kj}),\\
&\mathbf{h}_1(\theta_{ij}) = \mu_{\theta_i} + \mathbf{A}_{i\cdot}\mathbf{Z}_j + \epsilon_{ij}, \quad \mathbf{h}_2(\lambda_{kj}) = \mu_{\lambda_k} + \mathbf{B}_{k\cdot}\mathbf{Z}_j + \eta_{kj},\\
&\epsilon_{ij} \sim \mathrm{f}_\epsilon(\epsilon_{ij}), \quad \eta_{kj} \sim \mathrm{f}_\eta(\eta_{kj}), \quad \mathbf{Z}_j \sim \mathrm{f}_z(\mathbf{Z}_j).
\end{aligned} \tag{1}
$$

The parameter vector $\boldsymbol{\theta} \in \Re^m$ represents the natural parameters; $c(\mathbf{x})$, $g(\boldsymbol{\theta})$ are both non-negative functions. We obtain the natural family member by choosing $h(x) = x$, the canonical link. In model (1), the functions $\mathbf{h}_1$ and $\mathbf{h}_2$ are referred to as link functions. We assume that the link functions $\mathbf{h}_1$ and $\mathbf{h}_2$ are known smooth and invertible, as typically assumed in the generalized linear model framework (see McCulloch, 2006). The parameters μ_{θ_i} and μ_{λ_k} represent the mean of the natural parameters associated with the ith and kth feature in data sets $\mathbf{X}$ and $\mathbf{Y}$, respectively. The error terms ϵ_{ij} and η_{kj} are assumed independently distributed. In most applications,

f_ϵ and f_η are chosen to be the normal density with mean 0; the matrix $\mathbf{A}$, of dimension $p \times d$, and the matrix $\mathbf{B}$, of dimension $q \times d$, denote the weights (loading factors) matrices associated with the latent vector $\mathbf{Z}_j = (Z_{1j}, \cdots, Z_{dj})^{\mathrm{T}}$.

Model (1) is similar to the model considered in Virtanen et al. (2011), however we consider an extra level of stochasticity on the natural parameters. This allows for dependency between vectors of natural parameters as seen below. The shared vector of latent variables $\mathbf{Z}_j$ allows for dependency between variables in a specific data set. It also allows for dependency between variables in separate data sets that share a common vector of latent variables, i.e., data vector that belongs to the same sample. In model (1), if we assume normal distributions for $\mathbf{F}_x$, $\mathbf{F}_y$, f_ϵ, and f_η we recover the probabilistic canonical correlation model proposed by Bach and Jordan (2006). Hence, model (1) is an extension of that model. Our primary focus is estimation of dependency between two data sets while keeping the number of parameters to be estimated at a minimum.

Model (1) has few appealing characteristics, especially when used to model Next-Gen sequencing data sets. To illustrate these properties, as in Next-Gen sequencing we assume that entries of the data matrices $\mathbf{X}$ and $\mathbf{Y}$ are counts and we consider the special case of (1), wherein

$$\begin{aligned}
&X_{ij} \,|\, \theta_{ij} \sim \text{Poisson}\{\delta_{xj} \exp(\theta_{ij})\}, \quad Y_{kj} \,|\, \lambda_{kj} \sim \text{Poisson}\{\delta_{yj} \exp(\lambda_{kj})\},\\
&\theta_{ij} = \mu_{\theta_i} + \mathbf{A}_{i\cdot}\mathbf{Z}_j + \epsilon_{ij}, \quad \lambda_{kj} = \mu_{\lambda_k} + \mathbf{B}_{k\cdot}\mathbf{Z}_j + \eta_{kj},\\
&\epsilon_{ij} \sim \mathrm{f}_\epsilon(\epsilon_{ij}), \quad \eta_{kj} \sim \mathrm{f}_\eta(\eta_{kj}), \quad \mathbf{Z}_j \sim \text{Normal}(0, \mathrm{I}_d);
\end{aligned} \tag{2}$$

where I_d denotes the $d \times d$ identity matrix. δ_{xj} and δ_{yj} are known as sequencing or library size normalization factor as they adjust for the potential disproportional number of reads in different samples. Note that δxj and δyj are assumed fixed and are not estimated with the other model parameters. They can be estimated using methods proposed by (Anders and Huber, 2010; Robinson and Oshlack, 2010). We discuss statistical packages available to estimate δ_{xj} and δ_{yj} in more detail in Section 3.2. The random noises ϵ_{ij} and η_{kj} in (2) are assumed independent and normally distributed with mean 0 and variances σ^2_θ and σ^2_λ, respectively. The latent vector $\mathbf{Z}_j$ has length d and assumed to have a multivariate normal with mean vector 0 and an identity covariance matrix. The link functions $\mathbf{h}_1$ and $\mathbf{h}_2$ in (1) are chosen to be the log function $[\log\{\exp(x)\} = x]$, the canonical link for the Poisson distribution.

Prior to computing the marginal mean and covariances, we first define $\boldsymbol{\theta} = (\theta_{ij})$ and $\boldsymbol{\lambda} = (\lambda_{kj})$, for $i = 1, \cdots, p$, $k = 1, \cdots, q$, and $j = 1, \cdots, N$. Note that from (2), the vector $(\theta_{1j}, \cdots, \theta_{pj}, \lambda_{1j}, \cdots, \lambda_{qj})^{\mathrm{T}}$ has a multivariate normal distribution with mean $\boldsymbol{\mu}_\theta = (\mu_{\theta_1}, \cdots, \mu_{\theta_p}, \mu_{\lambda_1}, \cdots, \mu_{\lambda_q})^{\mathrm{T}}$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{A}\mathbf{A}^{\mathrm{T}} + \sigma^2_\theta \mathrm{I} & \mathbf{A}\mathbf{B}^{\mathrm{T}} \\ \mathbf{B}\mathbf{A}^{\mathrm{T}} & \mathbf{B}\mathbf{B}^{\mathrm{T}} + I\sigma^2_\lambda \end{pmatrix}. \tag{3}$$

Using properties of conditional expectation, the unconditional (marginal) mean and covariances are

$$\begin{aligned}
E(X_{ij}) = \mu_{x_{ij}} &= \delta_{xj} \exp\left\{\mu_{\theta_i} + (\mathbf{A}_{i\cdot}\mathbf{A}_{i\cdot}^{\mathrm{T}} + \sigma^2_\theta)/2\right\}\\
\text{var}(X_{ij}) &= \mu_{x_{ij}} + \left\{\exp\left(\mathbf{A}_{i\cdot}\mathbf{A}_{i\cdot}^{\mathrm{T}} + \sigma^2_\theta\right) - 1\right\}\mu^2_{x_{ij}}\\
\text{cov}(X_{ij}, X_{lj}) &= \mu_{x_{ij}}\mu_{x_{lj}}\left\{\exp\left(\mathbf{A}_{i\cdot}\mathbf{A}_{l\cdot}^{\mathrm{T}}\right) - 1\right\},\\
\text{cov}(X_{ij}, X_{im}) &= 0 \;\text{ for }\; j \neq m.\\
\text{cov}(X_{ij}, Y_{kj}) &= \mu_{x_{ij}}\mu_{y_{kj}}\left\{\exp\left(\mathbf{A}_{i\cdot}\mathbf{B}_{j\cdot}^{\mathrm{T}}\right) - 1\right\},\\
\text{cov}(X_{ij}, Y_{km}) &= 0 \quad \text{for} \quad j \neq m.
\end{aligned} \tag{4}$$

We make the following remarks about the quantities in (4):

- Although we assumed a conditional Poisson distribution for each entry of the data matrix $\mathbf{X}$ and $\mathbf{Y}$, the unconditional distribution has a larger variance than that of the Poisson. We can easily verify that $\exp\left(\mathbf{A}_{i\cdot}\mathbf{A}_{i\cdot}^{\mathrm{T}} + \sigma^2_\theta\right) - 1 > 0$. This is desirable, especially when modeling Next-Gen sequencing data that tends to be over-dispersed (see Lund et al., 2012; McCarthy et al., 2012).
- No constraints are imposed on the elements of the weight matrix $\mathbf{A}$. The marginal correlation between two variables X_{ij} and X_{lj} has the sign of the dot product of their respective row vectors in the weight matrix $\mathbf{A}$, i.e., $\mathbf{A}_{i\cdot}\mathbf{A}_{l\cdot}^{\mathrm{T}}$.
- Similar quantities are obtained from the data set $\mathbf{Y}$ by replacing μ_θ, $\mathbf{A}$, and σ^2_θ with μ_λ, $\mathbf{B}$, and σ^2_λ respectively.

The covariance and correlation between θ_{ij} and λ_{kj}, for any sample j are also obtained as

$$\begin{aligned}
\text{cov}(\theta_{ij}, \lambda_{kj}) &= \mathbf{A}_{i\cdot}\mathbf{B}_{k\cdot}^{\mathrm{T}},\\
\text{corr}(\theta_{ij}, \lambda_{kj}) &= \frac{\mathbf{A}_{i\cdot}\mathbf{B}_{k\cdot}^{\mathrm{T}}}{\sqrt{\mathbf{A}_{i\cdot}\mathbf{A}_{i\cdot}^{\mathrm{T}} + \sigma^2_\theta}\sqrt{\mathbf{B}_{k\cdot}\mathbf{B}_{k\cdot}^{\mathrm{T}} + \sigma^2_\theta}},\\
\text{cov}(\theta_{ij}, \lambda_{km}) &= 0 \quad \text{for} \quad j \neq m,\\
\text{corr}(\theta_{ij}, \lambda_{km}) &= 0 \quad \text{for} \quad j \neq m.
\end{aligned} \tag{5}$$

Since our problem originated from the need to assess dependencies between two Next-Gen data, we focus our discussion on the model proposed in (2). Inference in model (2) reduces to estimating the correlation/covariance matrix of the natural parameters. However, model (2) allows for two forms for covariance (correlation) estimation. The first one, in equation (4), is based on the marginal correlation. The second approach is based on the correlation between natural parameters of the Poisson distribution in Equation (5). As discussed in Section 1, the marginal correlation approach can result in correlation estimates constrained to a much narrower interval than the $(-1, 1)$ interval (see Whitt, 1976; Shin and Pasupathy, 2007; Yahv and Shmueli, 2011), especially in the presence of low count data. In this manuscript, we demonstrate that the correlation estimates based on the natural parameters perform better in capturing the dependency between two Next-Gen data sets when compared to the marginal correlation approach.

2.2. *Identifiability*

In model (2), the covariance matrix in equation (3) is identified up to a sign and a rotation. To see this, consider a rotation matrix $\mathbf{R}$ where $\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{I}$. For any choices of matrix $\mathbf{A}$, then $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{A}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} = \mathbf{A}^{*}(\mathbf{A}^{*})^{\mathrm{T}}$. Various approaches have been proposed to make the model identifiable and here we adopt the method proposed by Geweke and Zhou (1996). We impose a lower triangular structure for the matrices $\mathbf{A}$ and $\mathbf{B}$ and further impose that the elements of the diagonal are non-negative to remove the identifiability related to the sign. We now turn to the choice of d. Given the constraints imposed on the elements of $\mathbf{A}$ and $\mathbf{B}$, we only need to estimate $pd - d(d-1)/2 + p$ as compared to $p(p+1)/2$ parameters that are needed to estimate the covariance matrix of the data $\mathbf{X}$. Also, we only need to estimate $qd - d(d-1)/2$ parameters compared to the $q(q+1)/2 + q$ when estimating the covariance matrix of the data $\mathbf{Y}$. By imposing that $pd - d(d-1)/2 + p \leq p(p+1)/2$ and $qd - d(d-1)/2 \leq q(q+1)/2$, we can derive an upper bound for the values of parameter d given values of p and q. In reality, the bound on the possible values of d will not matter much since the relevant values of d will tend to be small as pointed out by Lopes and West (2004).

2.3. *Prior Specification*

Estimation of the parameters of model (2) is done using a Bayesian approach. We consider the following automatic relevance determination priors (Mackay, 1994) as conjugate priors for the elements of the weight matrices $\mathbf{A}$ and $\mathbf{B}$ as

$$\begin{aligned}
a_{ij} &\sim \text{Normal}(a_{ij} \,|\, 0, \boldsymbol{\tau}_j^{-1}) \text{ if } i < j;\\
a_{jj} &\sim \text{Normal}(a_{ij} \,|\, 0, \boldsymbol{\tau}_j^{-1})\mathbf{1}(\mathbf{a_{jj}} > \mathbf{0}) \text{ if } \mathbf{i} = \mathbf{j};\\
b_{kj} &\sim \text{Normal}(b_{kj} \,|\, 0, \boldsymbol{\tau}_j^{-1}) \text{ if } k < j;\\
b_{jj} &\sim \text{Normal}(b_{kj} \,|\, 0, \boldsymbol{\tau}_j^{-1})\mathbf{1}(\mathbf{b_{jj}} > \mathbf{0}) \text{ if } \mathbf{i} = \mathbf{j}, \qquad (6)
\end{aligned}$$

and $\boldsymbol{\tau}_j \,|\, a_0, b_0 \sim \text{Gamma}(a_0, b_0)$, where $i = 1, \cdots, p$, $j = 1, \cdots, d$, and $k = 1, \cdots, q$. Note that $\mathbf{1}$ is an indicator function and for an event A, $\mathbf{1}(\mathbf{A}) = \mathbf{1}$ if A is true and 0 otherwise. The hierarchical prior in (6) assumes that each column of the weight matrices $\mathbf{A}$ and $\mathbf{B}$ has the same prior. This allows for the sharing of information across data sets and also does an automatic column size selection of the matrices $\mathbf{A}$ and $\mathbf{B}$ (Mackay, 1994; Bishop, 2007). The sharing of information provides more stable estimates of the elements of the weight matrices $\mathbf{A}$ and $\mathbf{B}$ by artificially inflating the amount of information used to obtain a parameter estimate (Congdon, 2006). This is desirable, especially in a small sample size relative to the number of variable settings. The hyper-parameters a_0 and b_0 are assumed known. We also assume default conjugate priors for the remaining parameters in the model as

$$\begin{aligned}
\boldsymbol{\mu}_\theta \,|\, \nu_x &\sim \textstyle\prod_{i=1}^{p} \text{Normal}(0, \kappa_{x_i}),\\
\boldsymbol{\mu}_\lambda \,|\, \nu_y &\sim \textstyle\prod_{k=1}^{q} \text{Normal}(0, \kappa_{y_k}),\\
\sigma_\theta^2 \,|\, \nu_\theta, \mathrm{s}_\theta^2 &\sim \text{Inv-}\chi^2(\nu_\theta, \mathrm{s}_\theta^2), \quad \sigma_\lambda^2 \,|\, \nu_\lambda, \mathrm{s}_\lambda^2 \sim \text{Inv-}\chi^2(\nu_\lambda, \mathrm{s}_\lambda^2);
\end{aligned} \qquad (7)$$

where $\boldsymbol{\mu}_\theta = (\mu_{\theta_1}, \cdots, \mu_{\theta_p})^{\mathrm{T}}$ and $\boldsymbol{\mu}_\lambda = (\mu_{\lambda_1}, \cdots, \mu_{\lambda_q})^{\mathrm{T}}$. The parameters κ_x, κ_y, s_θ^2, s_λ^2, ν_θ, and ν_λ are assumed to be known. The combined priors (6) and (7) induce a prior distribution for the individual correlation parameters defined in (5).

2.4. *Posterior Sampling*

Inference in a Bayesian analysis is based on the posterior distribution. Given model (2), priors in (6) and (7), the posterior distribution is proportional to:

$$\begin{aligned}
&P(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{A}, \mathbf{B}, \mathbf{Z}, \sigma_\theta^2, \sigma_\lambda^2 \,|\, \text{Data}) \propto \ell(\text{Data} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{A}, \mathbf{B}, \mathbf{Z}, \sigma_\theta^2, \sigma_\lambda^2)\\
&\times \prod_{i=1}^{d} \Big\{ \tau_i^{p/2} \exp(-.5\tau_i A_{\cdot i}^{\mathrm{T}} A_{\cdot i}) \times \tau_i^{q/2} \exp(-.5\tau_i B_{\cdot i}^{\mathrm{T}} B_{\cdot i}) \tau_i^{a_0 - 1}\\
&\times \exp(-\tau_i / b_0) \Big\}\\
&\times \Big\{ \textstyle\prod_{j=1}^{N} \exp(-.5\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j) \Big\} \exp\Big\{ -\nu_\theta \mathbf{s}_\theta^2 / (2\sigma_\theta^2) \Big\} \sigma_\theta^{-2(1+\nu_\theta/2)}\\
&\times \exp\Big\{ -\nu_\lambda \mathbf{s}_\lambda^2 / (2\sigma_\lambda^2) \Big\} \sigma_\lambda^{-2(1+\nu_\lambda/2)}. \qquad (8)
\end{aligned}$$

The posterior distribution in (8) is difficult to directly simulate from. However, the choice of conjugate priors allows us to derive the full conditional posterior distributions for all model parameters. A Gibbs sampler is then used to update the parameters in a Markov chain Monte Carlo (MCMC). We derive the explicit form of the full conditional posterior for the model parameters in Web Appendix A. Our R code implementation exploits the package Rstan for fast computation.

3. Results

3.1. *Simulation*

We investigate the properties of the correlation estimates compared to other standard correlation approaches, under various assumed correlation matrices on the natural parameters. We assume that $\delta xj = \delta yj = 1$, for $j = 1, \cdots, N$. We simulated 50 data sets assuming for each data set $p = 10$, $q = 30$, and $N = 50$ subjects. We consider five correlation matrices for the natural parameters: (i) the identity correlation matrix ($d = 0$); (ii) a correlation matrix obtained assuming $d = 5$ for a given value of the weight matrices $\mathbf{A}$ and $\mathbf{B}$ using (3); (iii) a correlation matrix for the case where $d = 10$ also assuming equation (3); (d) an arbitrary correlation matrix with a specific dependence structure where the first two variables in the data set $\mathbf{X}$ are only strongly positively correlated with the variables 2–6 in $\mathbf{Y}$ ($d = NA(\text{Pos})$) through their natural parameters with correlation of .862; and (e) an arbitrary correlation matrix with a specific dependence structure where the first two variables in the data set $\mathbf{X}$ are only strongly negatively correlated with the variables 2–6 in $\mathbf{Y}$ ($d = NA(\text{Neg})$) through their natural parameters with correlation of $-.862$. We then fit model (2) to each of the $N = 50$ simulated data assuming various values of d, compute the posterior mean covariance/correlation matrix. We also compute the Frobenius and Stein losses for each of the estimated correlation matrices, where the Frobenius loss incurred by estimating the matrix $\mathbf{V}$ by $\mathbf{U}$, both $p \times p$ matrices, is defined as $\sum_{i,j}(\mathbf{U}_{ij} - \mathbf{V}_{ij})^2$ and the Stein loss is defined as $\text{diag}(\mathbf{V}^{-1}\mathbf{U}) - \det(\mathbf{V}^{-1}\mathbf{U}) - p$.

Table 1

Summary of the Stein losses when estimating the true correlation structure for the natural parameters. Here, d true is the value of d assumed for the true correlation matrix; d represents the value of d assumed by PCAN when fitting the model. Here d = 0 (identity matrix); d = 5 (correlation matrix obtained assuming d = 5); d = 10(correlation matrix obtained assuming d = 10); d = NA (arbitrary correlation matrix). Here Pearson is the sample correlation; Pearson(log) is the Pearson correlation based on the log of the counts after we added 1 to all the values and Spearman denotes the Spearman correlation estimator. Numbers in parentheses are standard errors.*

d	d*	PCAN	Pearson	Pearson (Log)	Spearman
0	10	2.46 (0.21)	41.19 (4.10)	44.50 (6.14)	47.78 (7.06)
5	10	12.72 (4.44)	66.40 (4.34)	41.39 (4.27)	27.25 (2.52)
10	10	24.48 (1.97)	108.77 (6.12)	36.53 (3.9)	27.79 (2.61)
NA(POS)	10	34.09 (.97)	65.15 (6.14)	64.04 (6.30)	64.66 (6.11)
NA(NEG)	10	34.75 (1.40)	65.15 (6.14)	65.80 (4.61)	64.90 (4.63)

Web Table 1 summaries the results of the simulation. As is clear from the terminology, smaller losses are preferred. There is a very obvious feature in these results, namely that assuming d larger than the truth leads to smaller losses, while choosing d smaller than the truth leads to large losses. Assuming $d = 10$ is clearly overall the safest course of action in this simulation.

We estimate correlations using other standard correlation estimation approaches. Since $N > p + q$, standard correlation estimates are valid and can be used to obtain estimates of the underlying correlation matrix. We consider: (i) the Pearson (sample based) correlation based on the raw data; (ii) the Pearson correlation based on the log transformed of the data after we add 1 to the observed counts; and (c) the Spearman correlation. The three correlation estimation approaches were applied to each of the 50 data sets simulated as previously described. We report the summary of the Stein loss incurred in estimating the true correlation matrices using the three standard approaches, see Table 1.

The PCAN approach, assuming $d = 10$, yields smaller Stein losses estimates compared to the default approaches considered overall. However, we note that the sample correlation approach yields the largest error in each case considered. The Spearman correlation approach seems to perform similarly to the PCAN approach when the true correlation matrix is obtained from equation 3; for arbitrary correlation matrix, however, all the default correlation estimation approaches tend to perform similarly, judging by the Stein loss. From the Frobenius norm, we also observe that the Pearson correlation gives the largest error in each scenario. However, the errors estimates for the Pearson(log) and the Spearman approach seem very similar to the error estimates under the PCAN approach, except when the true correlation is the identity matrix ($d = 0$). We report the results of the Frobenius lost in Web Table 1.

In addition to estimating the loss incurred in estimating the true underlying correlation, we also look at the plot of the first two loading factors on each of the variables. We expect that variables active (important) in describing the correlation across data sets will tend to have higher loading compared to non-important variables. We also take a look at the first two canonical vectors when using the traditional canonical correlation analysis approach. To compute the canonical vectors, we consider the R packages of González et al. (2008) for extended canonical correlation analysis and Witten et al. (2009) for penalized canonical correlation analysis to obtain estimates of the canonical vectors over the simulated data sets. For data simulated assuming the identity correlation matrix for the natural parameters, we obtain the plot of the estimated first and second canonical vectors along with their approximated individual 95% confidence bands for the PCAN and the extended canonical correlation analysis approaches in Figures 1 and 2. The same plot for the penalized canonical correlation analysis approach is reported Web Figure 1. The estimated means of the canonical vectors and factor loadings are zero or very close to zero, which is consistent with the fact that the true correlation matrix between all variables is the identity. Both the PCAN and the canonical correlation approaches are in agreement. This suggests that no correlation at the natural parameters level also translates into no correlation between the counts.

Now, we assume a true correlation matrix with a defined structure as above ($d = NA(\text{Pos})$). Because the true correlation matrix depicts strong positive correlation (.862) between the first two variables (natural parameters) in the data matrix $\mathbf{X}$ and the variables 2–6 in the data matrix $\mathbf{Y}$ and 0 elsewhere, we expect to observe non-zero or significantly large coefficients/loadings associated with these variables compared to other variables (or features). We show the plots of the estimated loading factors (Figure 3) and the plots of the canonical coefficients for both canonical correlation approaches (Web Figures 2 and 3).

The canonical weights estimated from the extended canonical correlation analysis approach yields large but non-significant coefficients on the true active variables. The PCAN and penalized canonical correlation analysis methods, however, yield very similar results and assign significant large coefficients to the correct (active) variables. Assuming now that the true correlation matrix also depicts strong negative correlation (−.862) between the first two variables (natural parameters) in the data matrix $\mathbf{X}$ and the variables 2–6 in the data matrix $\mathbf{Y}$ and 0 elsewhere ($d = NA(\text{Neg})$), we expect larger weights for these variables. The PCAN approach assigns larger weights to the variables driving the dependency between both data sets, although these weights are not significant (see Web Figure 4). However, estimates of the canonical weights obtained using the extended canonical correlation analysis approach give very similar results as in the case of $d = NA(\text{Pos})$, but the weights tend to be much closer to zero, similar to the case of the identity matrix ($d = 0$) (see Figure 1 and Web Figure 5). This suggests that strong negative

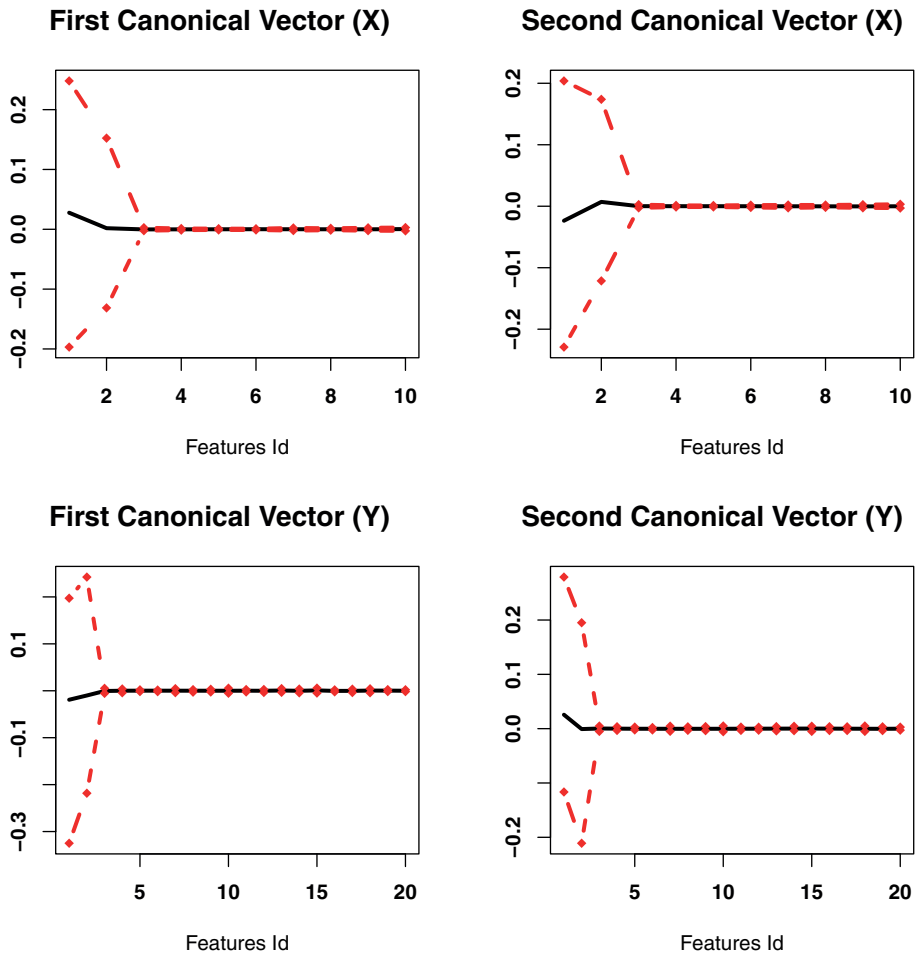


Figure 1. Plot of the estimated mean (solid line) of the canonical weights coefficients for each feature/variable along with their 95% confidence interval. Estimates are obtained using the extended canonical correlation analysis approach. The data are simulated assuming an identity correlation matrix for the natural parameters. $\mathbf{X}(50 \times 10)$ and $\mathbf{Y}(50 \times 20)$ are both simulated matrices of counts.

correlation at the natural parameters levels does not necessarily translates into strong negative correlation between the observed counts as they tend to be shrunken to zero, when correlation is computed using the standard approaches. We noted that the penalized canonical approach identifies the true variables in all cases (see Web Figure 6), but performance tend to decrease as we increase the number of variables and hold the sample size fixed.

As stated before, model (1) can be extended to many distributions member of the exponential family. In addition to the case of Poisson distribution, we also considered the case of Binomial data sets, with the logit link function. We compare the results of the correlation estimates using PCAN against correlation estimates obtained from the Pearson and Spearman approaches based on the logit (or log odd) of the estimated proportion of successes. The results are reported in Web Tables 2 and 3. Overall, we found that PCAN performs better than the Pearson and Spearman methods.

3.2. *Case Study*

Here, we analyze the squamous cell lung cancer data described in Section 1.

We considered $N = 150$ matched miRNA and the mRNA samples for our analysis. Here, we are interested in uncovering a potential relationship between lowly expressed mRNA and a given subset of miRNA. We select $p = 50$ miRNAs and $q = 66$ mRNAs. Of the 66 mRNA selected, 23 represent mRNAs with the lowest expression average across the 150 samples (average number of read counts between 1 and 2) and the remaining 43 were selected from a set of reported up/down-regulated mRNAs in squamous cell lung carcinoma (Shi et al., 2011). To estimate the dependency between both miRNA and mRNA, we used model (2). The values of δ_{xj} and δ_{yj} were computed using the calcNormFactors function in the edgeR package in R (Robinson et al., 2010) and were estimated equal to 1. This suggested that the mRNA and miRNA samples were sequenced at relatively similar depth (total number of reads).

We fit model (2) assuming the priors provided in the Web Appendix B to the data for values of $d = 2, 5, 10$. We ran three separate MCMC chains with different starting values for 20, 000 iterations, and monitored them for proper mixing. We discarded 10, 000 iterations as burn-in and inference was based on the 30, 000 remaining iterations. We then estimate

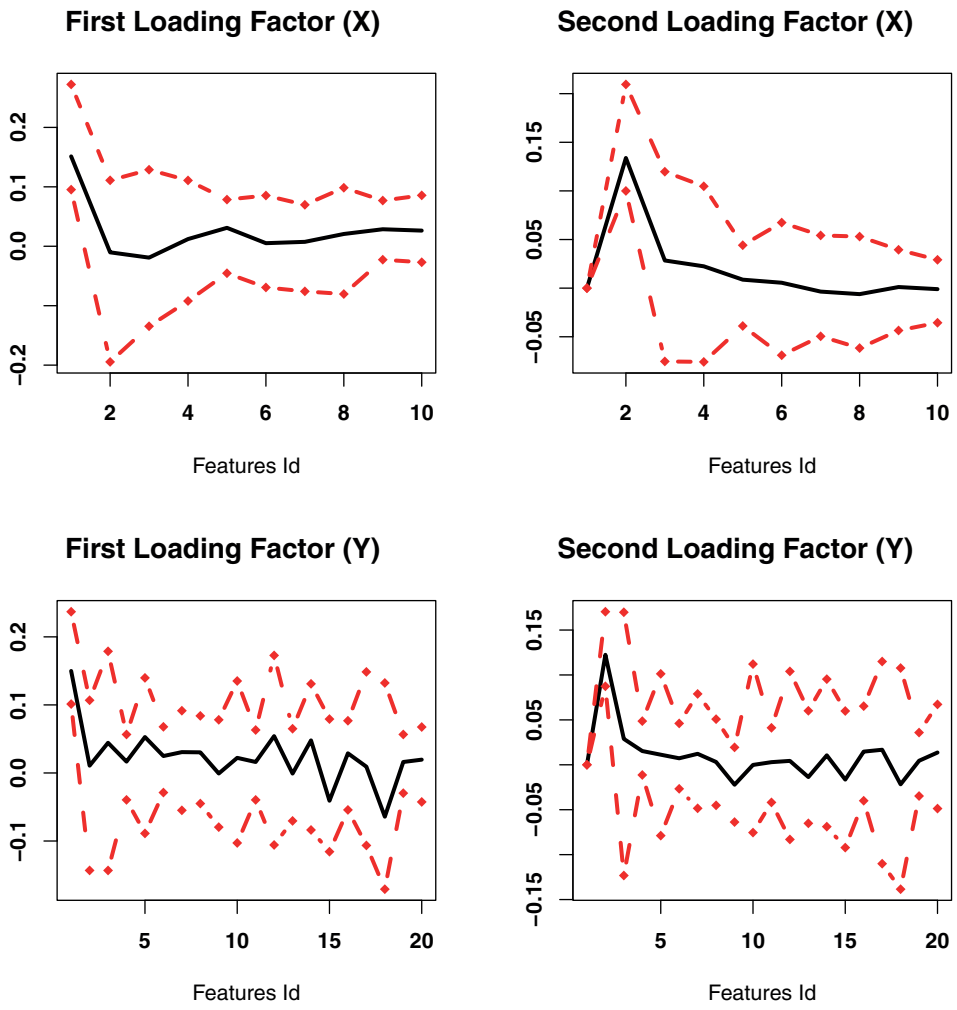


Figure 2. Plot of the estimated posterior mean (solid line) of the loadings on each natural parameters along with their 95% credible intervals. Estimates are obtained using the PCAN approach. The data are simulated assuming an identity correlation matrix for the natural parameters. $\mathbf{X}(50 \times 10)$ and $\mathbf{Y}(50 \times 20)$ are both simulated matrices of counts.

each element of the correlations between the miRNA and the mRNA, along with the bounds of the 95% credible intervals. We display, as a heat map, the posterior mean estimates of each correlation estimates in Figure 4.

The posterior correlation mean heat map identifies interesting groups of miRNA–mRNA interaction. We have for example, miR-205-5p are negatively correlated with the genes SLCO2A1, PECAM1, PTPRB, STARD8, which were found to be down-related mRNA (genes) in squamous cell lung carcinoma (see Shi et al., 2011); suggesting that SLCO2A1, PECAM1, PTPRB, STARD8 could be potential direct targets of miRNA-205-5p. These genes targets were subsequently validated using targetHub (`http://app1.bioinformatics.mdanderson.org/tarhub/_design/basic/index.html`). In addition, these correlations estimates are also found to be significant at the 5% level (see Web Table 5). We also generate a heat map depicting correlation estimates found to be significant (see Web Figure 6.) Approximately 47% of the correlation estimates are significant, and just 41(or 1%) of these correlation are greater than .34 in absolute value. Focusing on the correlations with point estimates above .34 in absolute value obtained from the PCAN approach (see Web Table 5), we observed that the pairs involving both miR-205 and miR-375 show up with genes found to be down-regulated in squamous cell lung cancer. We also note that for each of these mRNAs, both correlations with miR-205-5p and miR-375 are high with almost perfect opposite signs. This is consistent with literature suggesting that hsa-miR-205-5p produced a high diagnostic accuracy between squamous cell (SQ) and adenocarcinoma (AC) and are reported to be highly expressed in squamous cell lung cancer when compared to normal cells (Wei et al., 2014). Also, We also found the largest negative correlation is −.62 (PGC and miR-205-5p), which is consistent with the fact that of all the downregulated mRNAs selected PGC was reported as having the largest fold change, although PGC has not been predicted as a target of miR-205-5p. In addition, Huang et al. (2012) reported that miR-205-5p was found to be up regulated when miR-375 was down regulated

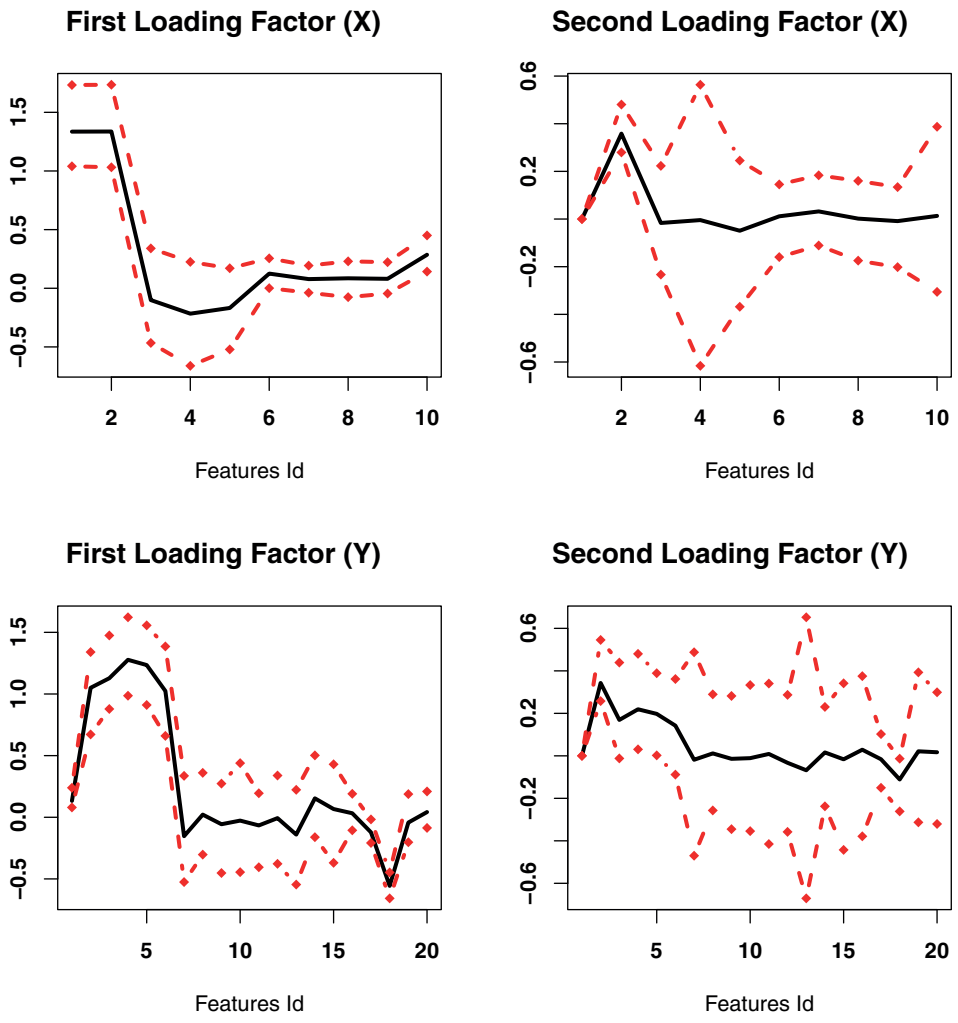


Figure 3. Plot of the estimated posterior mean (solid line) of the loadings on each natural parameters along with their 95% credible intervals. Estimates are obtained using the PCAN approach. The data are simulated assuming a correlation matrix with known structure for the natural parameters (NA(Pos)). $\mathbf{X}(50 \times 10)$ and $\mathbf{Y}(50 \times 20)$ are both simulated matrices of counts.

in the squamous cell, also explaining why both miR-205-5p and miR-375 show up with almost perfect opposite signs.

We also estimate the correlation using two other approaches: (i) Pearson/Spearman correlation based on the raw counts; (ii) Pearson/Spearman correlation based on the log2 transform of the counts, after adding one. The correlation estimates are computed and we report the p-value of the test, after correcting for multiple comparisons. The computation is done using the R function corr.test, available in the R package psych (Revelle, 2015). We use as cut-off 5% for the corrected p-values and correlation estimates above .34 in absolute value (see Web Tables 6-9). Figure 5 shows the venndiagram depicting the overlap between the pairs miRNA–mRNA correlation idenfied by all three approaches. We observe that there is a no pairs or very few pairs jointly identified by all three approaches. However, we obtain a single pair (PECAM1 - hsa-miR-205-5p) jointly identified by all three approaches when the Pearson and Spearman correlation are based on the log2 transformation of the data (see Figure 5b). It is worth restating here that we have a particular interest in estimation of negative correlation between miRNA and mRNA, as they may indicate that the mRNA is a target of the miRNA. Figure 5c,d show the venndiagram of significant negative correlations obtained by all three approaches. Although there is a greater overlap between the Pearson and Spearman approach, most of the pairs identified were not predicted by targetHub. In fact, the standard approach tend to miss few miRNA and mRNA pairs that are predicted by targetHub (i.e., miR-217 - TP63, miR-205-5p - SLC02A1, miR-205-5p - PTPRB) but these pairs were identified by PCAN as significant negative correlations. The pairs miRNA and mRNA, common to all three approaches, have similar estimates and are between $-.37$ and $-.35$.

In genomics, normalized counts are sometimes preferred when compared to raw counts and various normalization approaches have been proposed; see e.g., (Dillies et al., 2013) for a discussion of the current normalization approaches in

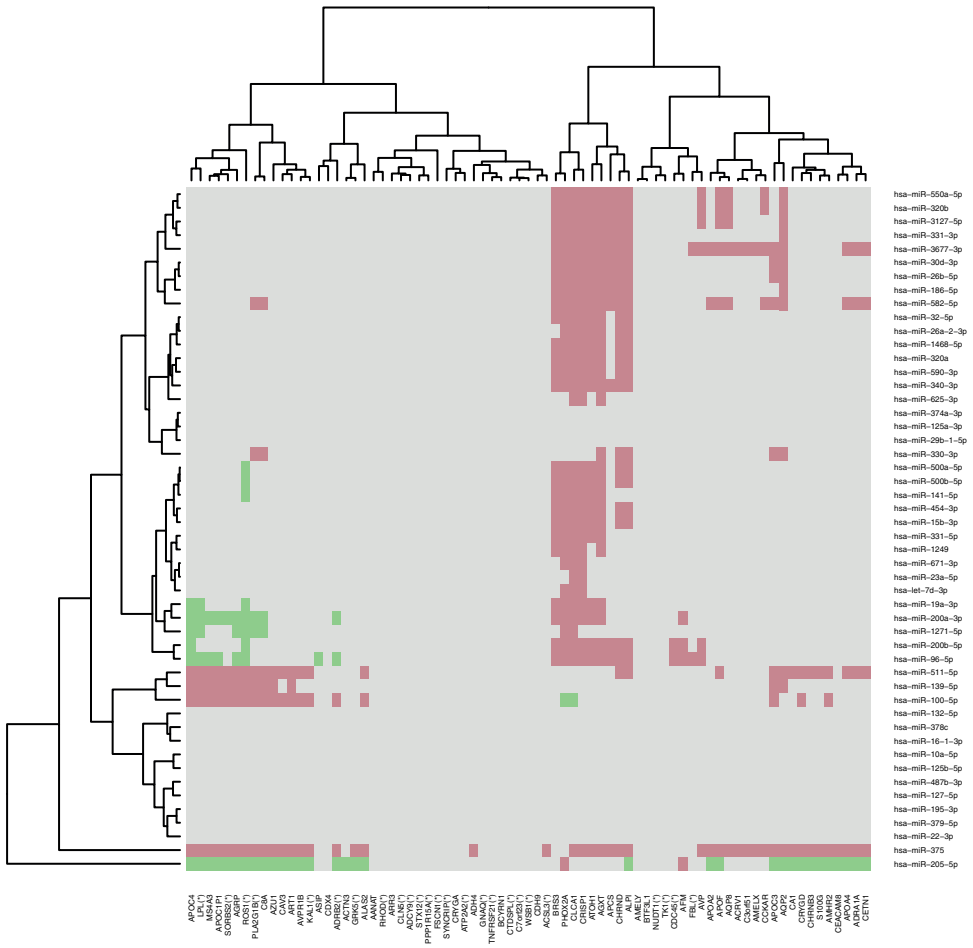
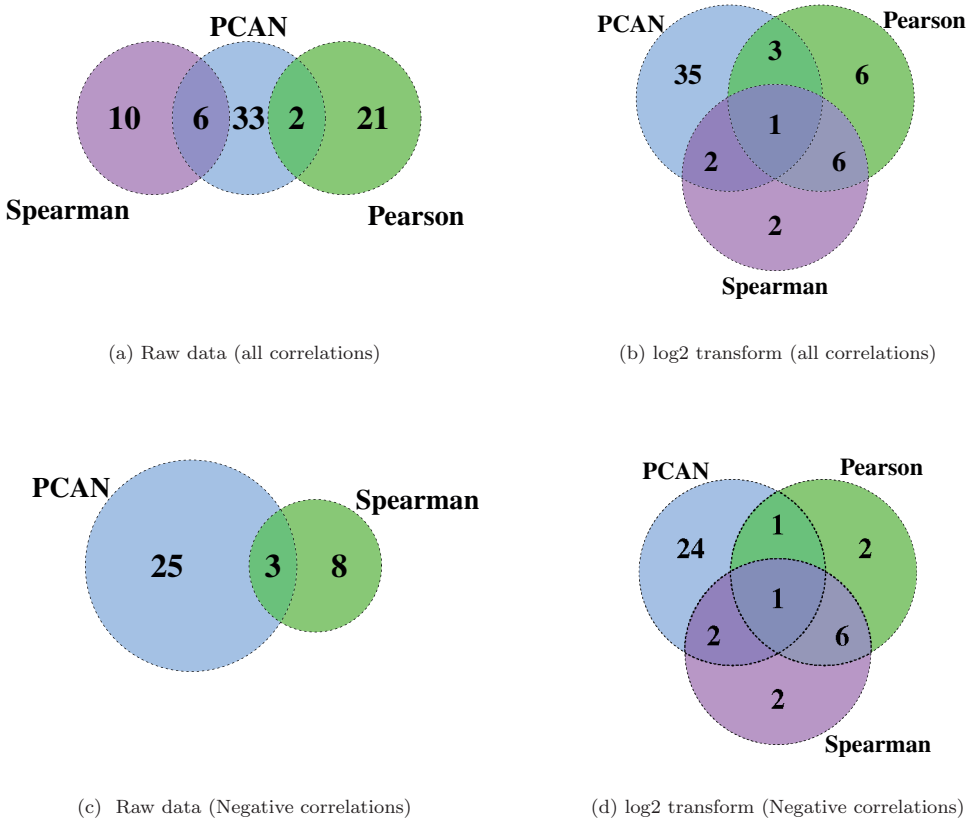


Figure 4. Heatmap of the posterior mean ($d = 2$) correlation estimates between the miRNA and mRNA. Lightest color represents correlation estimates between −0.2 and 0.2. Darker color represents correlation estimates less than −0.2. Darkest color represents correlation estimates greater than 0.2. This figure appears in color in the electronic version of this article.

RNAseq data. We consider the Reads Per Kilobase per Million Mapped reads (RPKM) approached proposed by (Mortazavi et al., 2008). We compute the normalized RPKM values for both miRNA and mRNAs and based the estimation of correlation (Pearson/Spearman) on their RPKM and log2 of the RPKM values. We repeat the same procedure as for the correlation computed based on the raw counts. We report the significant Pearson/Spearman correlation estimates based on RPKM and log2 RPKM values in Web Tables 10–13. Pearson/Spearman correlations based on the RPKM/log2 RPKM yields many significant pairs when compared to the same correlation based on the raw counts. This is not too surprising since the RPKM expression values follow by a log2 transformation tend to make the distribution of the expressions value closer to symmetric distribution (see, e.g., Rahmatallah et al., 2014. We observe a greater overlap between the PCAN approach and the Pearson/Spearman correlation based on the RPMK/log2 RPKM values (see Web Figure 7). We obtain an even greater overlap when we lower our cutoff point from .34 to say .24. However, PCAN tended to estimate low, although not significantly, correlations for some of the common miRNA and mRNA pairs identified, which all tended to involve the miRNA hsa-miR-205-5p. This is consistent with the behavior of PCAN which tend to focus on few important miRNA and mRNA and shrink the correlation of the remaining pairs to zero. Unlike the Spearman/Pearson correlation estimates, estimation of the correlations between miRNAs and mRNAs affect the estimates of correlations between miRNA and mRNA in the PCAN approach. For example, miR-205-5p and miR-375 pair is found to be significantly negatively correlated in all approaches (correlation estimates around −.37); and PCAN also finds both miR-205-5p and miR-375 to be associated with the same mRNAs with almost opposite correlation. For example, we have PGC - miRNA-205-5p and PGC - miRNA-375 with correlation estimates, respectively −.62 and −.69. The Pearson/Spearman approach identify the pair PGC - miRNA-375 (.59) but not the pair PGC - miRNA-205-5p (−.15). All of these illustrate the difference between the PCAN and the standard approaches.

4. Discussion and Conclusions

We have proposed a new probabilistic model for the estimation of the correlations based on two non-normal data sets, with emphasis on Next-Gen data. Next-Gen data sets are



(a) Raw data (all correlations)

(b) log2 transform (all correlations)

(c) Raw data (Negative correlations)

(d) log2 transform (Negative correlations)

Figure 5. Venn diagram summarizing the significant miRNA and mRNA correlation estimates identified under the PCAN, and Pearson/Spearman approaches based both on the raw data and the log2 transformation of the data as described in the case study section.

counts data, with often very low counts, for which standard approaches of computing correlation lead to small correlation estimates. In this article, we propose considering correlation as describing the dependency between natural parameters of the data generating model, rather than the correlation between the counts. Correlation estimated based on the dependence between the natural parameters are preferred compared to the correlation based on the raw counts for the following reasons. In the case of count data with low counts, weak correlation estimates obtained at the natural parameters level are directly associated with weak correlations estimates based directly on the counts data sets. This explains the clear agreement between the extended canonical correlation analysis approach and the PCAN approach when the true underlining correlation is the identity (see Figure 1 and 2). However, strong (negative) correlations at the natural parameters level do not necessarily translate into strong correlation between the observed counts. This explains the discrepancy between the extended canonical correlation analysis method and PCAN when identifying active variables on Figure 3 and Web Figure 2. We also found that strong correlation estimates based on the raw counts are also associated with strong correlation at the natural parameters. Finally, because PCAN considers the estimation of dependency within and between two data sets as a joint task, the results obtained by PCAN tend to be more interpretable when compared to other standard correlation estimation approaches.

5. Supplementary Materials

Web Appendices, Tables, Figures referenced in Sections 2–3, and the R code implementing our approach are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

Roger S. Zoh was supported by the grant U01CA162077 awarded to Drs. Chapkin and Lampe. Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030). Mallick's research was supported in part by NSF grant DMS-0914951. The authors would like to thank the Whole Systems Genomics Initiative at Texas A&M University for their computing support. The authors would like to sincerely thank the associated editor and reviewers for their comments and suggestions which help significantly improve the presentation of the article.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* **11**, R106.

Archambeau, C., and Bach, F. (2009). Sparse Probabilistic Projections. In Advances in Neural Information Processing Systems 21 22nd Annual Conference on Neural Information Processing Systems 2008 (pp. 73-80). NY: Curran Associates Inc. Red Hook.

Bach, F. R. and Jordan, M. I. (2006). A probability interpretation of canonical correlation analysis. Technical Report 688.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York, NY: Springer, 1 edition.

Congdon, P. (2006). *Bayesian Statistical Modelling*. West Sussex, England: John Wiley & Sons, 2 edition.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics* **14**, 671–683.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies* **9**, 557–587.

González, I., Déjean, S., Martin, P. G., and Baccini, A. (2008). Cca: An r package to extend canonical correlation analysis. *Journal of Statistical Software* **23**, 1–14.

Huang, W., Hu, J., Yang, D.-W., Fan, X.-T., Jin, Y., Hou, Y.-Y., et al. (2012). Two microrna panels to discriminate three subtypes of lung carcinoma in bronchial brushing specimens. *American Journal of Respiratory and Critical Care Medicine* **186**, 1160–1167.

Klami, A. and Kaski, S. (2007). Local dependent components. In *Proceedings of the 24th International Conference on Machine Learning*, 425–432, ICML '07, New York, NY: ACM.

Lê Cao, K.-A., González, I., and Déjean, S. (2009). integrOmics: An R package to unravel relationships between two omics datasets. *Bioinformatics* **25**, 2855–2856.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–68.

Lund, S. P., Nettleton, D., McCarthy, D. J., and Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*, **11**, 8.

Mackay, D. J. C. (1994). Bayesian methods for backpropagation networks. In *Models of Neural Networks III*, E. Domany, J. L. van Hemmen, and K. Schulten, (eds), chapter 6, 211–254. New York, NY: Springer.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Research* **40**, 4288–4297.

McCulloch, C. E. (2006). *Generalized linear mixed models*. Hoboken, NJ: Wiley Online Library.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods* **5**, 621–628.

Rahmatallah, Y., Emmert-Streib, F., and Glazko, G. (2014). Comparative evaluation of gene set analysis approaches for rna-seq data. *BMC Bioinformatics* **15**, 397.

Ren, J., Jin, P., Wang, E., Marincola, F. M., and Stroncek, D. F. (2009). Microrna and gene expression patterns in the differentiation of human embryonic stem cells. *Journal of Translational Medicine* **7**, 20.

Revelle, W. (2015). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.5.8.

Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology* **11**, R25.

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**, 139–140.

Shah, M. S., Schwartz, S. L., Zhao, C., Davidson, L. A., Zhou, B., Lupton, J. R., et al. (2011). Integrated microrna and mrna expression profiling in a rat colon carcinogenesis model: Effect of a chemo-protective diet. *Physiological Genomics* **43**, 640.

Shi, I., Sadraei, N. H., Duan, Z.-H., and Shi, T. (2011). Aberrant signaling pathways in squamous cell lung carcinoma. *Cancer Informatics* **10**, 273.

Shin, K. and Pasupathy, R. (2007). A method for fast generation of bivariate poisson random vectors. In *Simulation Conference, 2007 Winter*, 472–479. NJ: IEEE Press Piscataway.

Stan Development Team (2014). Stan: A c++ library for probability and sampling, version 1.3.0. http://mc-stan.org/

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis. *Journal of Royal Statistical Society B* **61**, 611–622.

Virtanen, S., Klami, A., and Kaski, S. (2011). Bayesian cca via group sparsity. In *ICML*, 457–464, NY: New York.

Wei, H., Yi, J., Yunfeng, Y., Chunxue, B., Ying, W., Hongguang, Z., et al. (2014). Validation and target gene screening of hsa-mir-205 in lung squamous cell carcinoma. *Chinese Medical Journal* **127**, 272–278.

Whitt, W. (1976). Bivariate distributions with given marginals. *The Annals of Statistics* **4**, 1280–1289.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

Yahv, I. and Shmueli, G. (2011). On generating multivariate poisson data in management science applications. *Applied Stochastic Models in Business Industry* **28**, 91–102.

Received July 2014. *Revised December* 2015.
Accepted February 2016.