

A Statistical Framework to Infer Functional Gene Associations from Multiple Biologically Interrelated Microarray Experiments

¹Siew Leng Teng, ²X. Jasmine Zhou, ^{3*}Haiyan Huang

¹Division of Biostatistics, University of California, Berkeley, USA; ²Program in Molecular and Computational Biology, University of Southern California Los Angeles, USA; ³Department of Statistics, University of California, Berkeley, USA
slteng@stat.berkeley.edu, xjzhou@usc.edu, hhuang@stat.berkeley.edu

September 2007

Abstract

A major task in understanding biological processes is to elucidate relationships between genes involved in the underlying biological pathways. Microarray data from an increasing number of biologically interrelated experiments now allows for more complete portrayals of functional gene relationships in the pathways. In current studies of gene relationships, the existence of expression dependencies attributable to the biologically interrelated experiments, however, has been widely ignored. When not accounted for, these (experimental) dependencies can result in inaccurate inferences of functional gene relationships, and hence incorrect biological conclusions. This article contributes a framework to provide a model and an estimation procedure for inferring gene relationships when there are two-way dependencies in the gene expression matrix (the gene-wise and experiment-wise dependencies). The main aspect of the framework is the use of the Kronecker product covariance matrix to model the gene-experiment interactions. The resulting novel gene co-expression measure, named Knorm correlation, can be understood as a natural extension of the widely used Pearson coefficient. Compared to Pearson approach, the Knorm correlation has a much smaller estimation variance and is asymptotically consistent with the Pearson coefficient. We demonstrated the advantages of the Knorm correlation in both simulation studies and real datasets applications. The Knorm correlation estimation procedure is implemented in the R package `knorm` that is freely available from the Bioconductor website.

KEY WORDS: Co-expression measure; Experimental dependency; Functional gene relationships; Gene-experiment interactions; Kronecker product.

* Corresponding author.

1 Introduction

A major task in understanding biological processes is to elucidate relationships between genes involved in the underlying biological pathways. Of particular interest are the functional gene relationships that arise as genes respond variedly to different but biologically interrelated experiments. These experiments are typically designed to trigger cellular changes by the use of different reagents, physiological conditions or their combinations with different time points (e.g. Sabet et al., 2004; Wu et al., 2005; Cromer et al., 2004; Lund et al., 2005). As such, the gene relationships are context-specific.

In this article, we address a lesser studied but critical issue in the gene relationship inference: the presence of experiment dependencies in gene expression data. We define experiment dependencies as dependencies in gene expressions *between* experiments due to the similar or related cellular states induced by the experiments. The presence of experiment dependencies is natural, and they *co-exist* with the gene relationships (or gene dependencies). Evident in a yeast dataset, Figure 1 shows stronger dependencies in normalized gene expressions between experiments 3, 4 and 7 than that between experiment 1 and the remaining experiments. This is due to similar cellular changes induced by histone H3 mutations in experiments 3, 4 and 7 but *not* in experiment 1. Figure 1 also illustrates the varied levels of dependencies in gene expressions between the experiments, depending on the extent and type of histone mutations being introduced. These experiment dependencies have been similarly observed in other real datasets used to study context-specific gene relationships, e.g. Cromer et al. (2004), Lund et al. (2005) and Wu et al. (2005). The negative impact of having experiment dependencies is that they introduce redundancies in data that can overwhelm the important signals and thereafter lead to inaccurate estimates of gene dependencies. In particular, the correlation coefficient measure (Pearson coefficient), widely used to infer gene relationships in many important studies (e.g. Eisen et al. 1998; Hanisch et al., 2002; Kim et al. 2001; Li, 2002; Zhou et al. 2005), suffers from an increased estimation variance and almost a random sign as experiment dependencies go unaccounted for (see Figure 2). This undesirable effect consequently contributes to a higher false positive rate of functional gene relationships identified by the Pearson coefficient in real datasets (see Tables 1 and 2). Therefore there is a need to adjust for these experiment dependencies to increase accurate inferences and improve biological conclusions, especially in data from biologically interrelated experiments where experiment dependencies are naturally strong.

Another fundamental issue we face in inferring gene relationships is the complex data structure in real datasets. A typical dataset used in pathway studies consists of replicates of gene expressions in each experiment and not expression matrix replicates. The number of replicates is often small (e.g. 2–3) and can be different for each experiment. The data structure is further complicated by the presence of both biological and nuisance variations in a gene expression.

Motivated by the above issues, this article attempts to address the following important questions: *How do we model and estimate the **experiment** dependencies from the complex data structure? How would the gene correlation measure be adjusted for experiment dependencies?* For the first question, we present a framework consisting of a statistical model, and a practical estimation procedure. This model uses a linear additive model with random gene, experiment and gene-experiment interaction effects, and a Kronecker product covariance matrix to model both experiment and gene dependencies in gene expressions. The model also delineates biological and nuisance variations in gene expressions. For the second question, we derive a new measure (named Knorm correlation) that adjusts for

experiment dependencies by weighting the gene expression in each experiment proportional to the partial correlation of that experiment. The Knorm correlation has a smaller estimation variability than the Pearson coefficient, especially when there are only a few replicates available for each experiment. The Knorm correlation simplifies to the Pearson coefficient when experiments are uncorrelated (i.e. experiment dependencies are absent). In addition to the model in our framework, we further provide a practical estimation procedure to mitigate issues in real dataset applications. This procedure uses bootstrapping to re-construct gene expression matrices from the replicates observed in each experiment, and a gene sub-sampling and covariance shrinkage technique to stabilize and reduce estimation errors in Knorm correlations due to the high dimensionality of real datasets.

We note that although two-way dependencies in matrix data has been studied in areas like repeated measurements (e.g. Timm, 1980; de Munck et al., 2002), the modeling of co-existing gene and experiment dependencies is still relatively unexplored in conjunction with the complex and high dimensional gene expression data in pathway studies. Other existing works that model specific spatial dependencies between experiments, e.g. fourier series approach by Spellman et al. (1998) and autoregressive models by Ramoni et al. (2002), require specific assumptions that may not be generally satisfied by typical datasets in pathway studies (e.g. as illustrated in Figure 1). Our work, on the other hand, is motivated to model the generally monotonic and varied experiment dependencies *from* the complex data structure observed in many real datasets used in pathway studies. We demonstrate the advantages of Knorm correlation over the existing methods by both analytically discussing its properties and numerically comparing it with other methods in applications (to simulation and real datasets).

The article is organized as follows. Section 2 introduces and elaborates on our framework, Knorm correlation and practical estimation procedure. Real datasets used in our analyses are described in Section 3. Section 4 presents the application results of Knorm correlation in simulation studies and real datasets. The Knorm correlation reports higher percentages of functionally related GO (Gene Ontology) annotated gene pairs in real datasets. Using the yeast dataset as an illustrative example, Section 5 provides an empirical justification of an assumption in our model. Finally, Section 6 discusses some practical and technical issues encountered in practice.

2 Statistical Framework

2.1 Statistical Model

Let X_{ijk} represents the gene expression for gene i in the k th replicate of experiment j , $i = 1, \dots, p$, $j = 1, \dots, n$, $k = 1, \dots, n_j$ where n_j represents the number of replicates for experiment j . Following the data structure, we introduce two random effects: gene effect and experiment effect. We postulate that a gene effect is a *random* effect that consists of three components:

- (i) a *fixed* component \mathbf{G} that measures the average gene expression level. This component depends on the gene only and is *independent* of the experiments.
- (ii) a *random* component that accounts for the **nuisance variations** arising from sources such as measurement errors. This component explains changes in gene expressions that are *independent* of the experiments.

- (iii) a *random* component that accounts for **biological variations** in gene expressions. These variations are triggered as the gene responds to the various experiments. This component then represents the “*gene-experiment interaction effects*” and would dominate the random component in (ii) if the gene expresses differently in the experiments.

The postulation also applies to the experiment effect. Putting together the above components, a linear additive model of gene and experiment effects simplifies to:

$$X_{ijk} = u_i + v_j + \gamma_{ij}^{GE} + \varepsilon_{ijk}, \quad i = 1, \dots, p, j = 1, \dots, n, k = 1, \dots, n_j, \quad (1)$$

where u_i and v_j are fixed components representing the individual gene and experiment effects respectively, γ_{ij}^{GE} denotes the random gene–experiment interaction effect, and ε_{ijk} is a random term representing all nuisance variations in X_{ijk} . The nuisance terms ε_{ijk} are *i.i.d.* with zero means and are independent of the interaction terms γ_{ij}^{GE} , $i = 1, \dots, p, j = 1, \dots, n, k = 1, \dots, n_j$. For an appropriately constructed $p \times n$ gene expression matrix \mathbf{X} (further elaborated in Section 2.3), we can rewrite our model in the following matrix representation

$$\mathbf{X} = \mathbf{G} + \mathbf{E} + \mathbf{\Gamma}^{GE} + \boldsymbol{\varepsilon}, \quad (2)$$

where $\mathbf{G} = (u_1 \dots u_p)^T \cdot \mathbf{1}^T$ ($\mathbf{1}$ is the unit column vector), $\mathbf{E} = \mathbf{1} \cdot (v_1 \dots v_n)$, $\mathbf{\Gamma}^{GE} = (\gamma_{ij}^{GE})_{j=1, \dots, n}^{i=1, \dots, p}$ is a zero-mean random matrix with elements representing the gene-experiment effects, $\boldsymbol{\varepsilon}$ is a zero-mean random matrix with elements representing *i.i.d.* normal noises, and $\mathbf{\Gamma}^{GE}$ and $\boldsymbol{\varepsilon}$ are independent of each other.

Using covariance matrices $\boldsymbol{\Sigma}^G$ and $\boldsymbol{\Sigma}^E$ to represent gene and experiment dependencies respectively, we represent the covariance matrix of $\mathbf{\Gamma}^{GE}$ by $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^E$: the Kronecker product of $\boldsymbol{\Sigma}^G$ and $\boldsymbol{\Sigma}^E$. With negligible nuisances, $\text{vec}(\mathbf{X}^T)$ follows a multivariate normal distribution with mean $\mathbf{G} + \mathbf{E}$ and a covariance matrix $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^E$.

The use of the Kronecker product covariance matrix can be understood in the following ways. First, $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^E$ can be interpreted as *a natural extension of the covariance matrix* $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^I$, where $\boldsymbol{\Sigma}^I$ denotes the identity matrix. The covariance matrix $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^I$ is the current widely used dependency structure for multivariate normal models that leads to the Pearson coefficients between genes. Secondly, $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^E$ can be interpreted as a dependency structure between the interaction terms caused by *a dual projection of a matrix of i.i.d. random variables with unit variances onto the eigenspaces determined by $\boldsymbol{\Sigma}^G$ and $\boldsymbol{\Sigma}^E$* . That is, $\mathbf{\Gamma}^{GE}$ can be represented as

$$\mathbf{\Gamma}^{GE} = (\mathbf{U}\mathbf{D}^{1/2}) \boldsymbol{\Lambda} (\mathbf{P}^{1/2}\mathbf{V}^T), \quad (3)$$

where $\boldsymbol{\Lambda}$ is a matrix of *i.i.d.* random variables with unit variances, \mathbf{P} is a diagonal matrix with diagonal elements being the eigenvalues of $\boldsymbol{\Sigma}^E$, and the eigenvectors of $\boldsymbol{\Sigma}^E$ make up the columns of \mathbf{V} (i.e. $\boldsymbol{\Sigma}^E = \mathbf{V} \mathbf{P} \mathbf{V}^T$), \mathbf{D} is a diagonal matrix with diagonal elements being the eigenvalues of $\boldsymbol{\Sigma}^G$, and the eigenvectors of $\boldsymbol{\Sigma}^G$ make up the columns of \mathbf{U} (i.e. $\boldsymbol{\Sigma}^G = \mathbf{U} \mathbf{D} \mathbf{U}^T$). In brief, \mathbf{P} , \mathbf{V} , \mathbf{D} and \mathbf{U} are from the singular value decompositions of $\boldsymbol{\Sigma}^E$ and $\boldsymbol{\Sigma}^G$. Consequently the covariance matrix of $\mathbf{\Gamma}^{GE}$ is $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^E$. Therefore, $\mathbf{\Gamma}^{GE}$ can be understood as a random matrix derived by “projecting a matrix of *i.i.d.* random variables with unit variances onto the eigenspaces determined by $\boldsymbol{\Sigma}^G$ and $\boldsymbol{\Sigma}^E$ ”. Following equation (3),

$$\mathbf{X} = E(\mathbf{X}) + (\mathbf{U}\mathbf{D}^{1/2}) \boldsymbol{\Lambda} (\mathbf{P}^{1/2}\mathbf{V}^T) + \boldsymbol{\varepsilon}. \quad (4)$$

That is, with negligible nuisances,

$$\boldsymbol{\Lambda} = \mathbf{D}^{-1/2} \mathbf{U}^T (\mathbf{X} - E(\mathbf{X})) \mathbf{V} \mathbf{P}^{-1/2}, \quad (5)$$

which implies that after removing the fixed components $E(\mathbf{X})=\mathbf{G}+\mathbf{E}$ from \mathbf{X} , projecting $(\mathbf{X}-E(\mathbf{X}))$ onto the gene and experiment eigenspaces (determined by $\boldsymbol{\Sigma}^G$ and $\boldsymbol{\Sigma}^E$ respectively) removes the two-way dependencies among the elements in \mathbf{X} and results in a matrix of independent random variables. When the covariance matrices are singular, the pseudo-inverses of \mathbf{P} and \mathbf{D} can be used for projection (Penrose, 1995). This will achieve a similar projection effect; the elements in the resulting matrix $\boldsymbol{\Lambda}$ are either independent random variables or zeros, with the number of zeros determined by the ranks of $\boldsymbol{\Sigma}^G$ and $\boldsymbol{\Sigma}^E$.

Thus when the elements in $\boldsymbol{\Lambda}$ are *i.i.d.* $N(0,1)$ random variables, from (4) with negligible nuisances, $\text{vec}(\mathbf{X}^T)$ would be multivariate normally distributed with mean $\mathbf{G} + \mathbf{E}$ and covariance matrix $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^E$. A detailed proof is provided in the Appendix.

2.2 Parameter Estimation and Knorm Correlation

For model (2) to be identifiable, we assume $(\mathbf{E})_{ij}=E_j = 0$ and that each experiment effect is associated with mean zero and a unit variance. These assumptions are reasonable with gene expression datasets as the normalized gene expressions have the same mean and variance in each experiment (e.g. RMA by Irizarry et al. 2003). Without loss of generality, we set these parameters to 0 and 1 respectively. These identifiability conditions, however, can be different when different datasets are considered and should be determined based on the nature of the data and the purpose of analysis. For the remaining part of the article, we will use the experiment correlation matrix \mathbf{R}^E to represent the experiment dependencies.

Following from the distribution of $\text{vec}(\mathbf{X}^T)$, the Maximum Likelihood Estimators (MLEs) of $\boldsymbol{\Sigma}^G$, \mathbf{R}^E , and $\boldsymbol{\mu}$, conditional upon the remaining parameters, can be shown to be

$$\hat{\boldsymbol{\Sigma}}^G = \frac{1}{n} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T) (\mathbf{R}^E)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T)^T, \quad (6)$$

$$\hat{\mathbf{R}}^E = \frac{1}{p} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T)^T (\boldsymbol{\Sigma}^G)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T), \quad (7)$$

$$\hat{\boldsymbol{\mu}} = \frac{\mathbf{X} (\mathbf{R}^E)^{-1} \mathbf{1}}{\mathbf{1}^T (\mathbf{R}^E)^{-1} \mathbf{1}}, \quad (8)$$

where $\mathbf{1}$ is a unit column vector. The gene correlation matrix \mathbf{R}^G can be estimated as

$$\hat{\mathbf{R}}^G = \mathbf{W}^{-1/2} \hat{\boldsymbol{\Sigma}}^G \mathbf{W}^{-1/2} = \frac{1}{n} \mathbf{W}^{-1/2} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T) (\mathbf{R}^E)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T)^T \mathbf{W}^{-1/2}, \quad (9)$$

where \mathbf{W} is a diagonal matrix with the same diagonal elements in $\hat{\boldsymbol{\Sigma}}^G$. The MLE of $\boldsymbol{\mu}$ in equation (8), conditional on $\boldsymbol{\Sigma}^G$ and \mathbf{R}^E , is unbiased and consistent. Similarly, the MLEs of $\boldsymbol{\Sigma}^G$ and \mathbf{R}^E , conditional upon the remaining two parameters, are also consistent estimators. The reader is referred to the Appendix for detailed derivations of equations (6)–(8).

Knorm correlation. The *Knorm correlation* is defined by equation (9). The Knorm correlation has the several appealing interpretations:

- (i) When experiments are uncorrelated, i.e. $\mathbf{R}^E = \mathbf{\Sigma}^I$, the Knorm correlation reduces to the Pearson coefficient. This follows from the previous argument that the covariance matrix $\mathbf{\Sigma}^G \otimes \mathbf{\Sigma}^E$ is a natural extension of $\mathbf{\Sigma}^G \otimes \mathbf{\Sigma}^I$. A model with the latter covariance matrix assumes absence of experiment dependencies (i.e. uncorrelated experiments), and leads to the Pearson coefficient.
- (ii) The Knorm correlation is the simple correlation between transformed gene expression profiles. The transformation is achieved through a projection onto the eigenspace of \mathbf{R}^E that removes the dependencies in gene expressions between experiments.
- (iii) The Knorm correlation is a weighted Pearson coefficient of gene expressions with weights proportional to the partial correlation of the experiments. This interpretation comes from that the (i,j) th element in the precision matrix $(\mathbf{R}^E)^{-1}$ is proportional to the partial correlation between experiments i and j conditional on all other experiments.

2.3 Practical Estimation Procedure For Real Dataset Applications

The high dimensionality of real datasets can result in large estimation errors and unstable covariance matrix estimates, especially when $\mathbf{\Sigma}^G$ and \mathbf{R}^E are not sparse. To mitigate this impact, we develop a practical estimation procedure consisting of a row (i.e. gene) sub-sampling and a covariance shrinkage technique that iteratively estimates the covariance matrices from equations (6)–(8).

This procedure consists of three main steps. The first step provides a sample of data matrices for parameter estimation as there is no actual observed expression matrix in practice; only replicates of each column (i.e. experiment) in \mathbf{X} are observed instead of matrix replicates. Following the model in equation (2), a parametric bootstrapping technique is used to construct the data matrices by bootstrapping the nuisance residuals $\boldsymbol{\epsilon}_b$ in the data.

The second step focuses on obtaining a reliable estimate of \mathbf{R}^E from the sample by reducing estimation errors in $\hat{\mathbf{R}}^E$. This is achieved by an iterative procedure of equations (6)–(8) with a row sub-sampling technique (to enable a comparable number of rows and columns in estimation) and a covariance shrinkage technique (to stabilize the estimated covariance matrices). The third step then uses the $\hat{\mathbf{R}}^E$ obtained from the previous step to estimate $\mathbf{\Sigma}^G$.

Practical estimation procedure

Step 1: Obtain a sample of data matrices $\mathbf{X}_1, \dots, \mathbf{X}_B$ by placing in the j th column of each \mathbf{X}_b a randomly selected replicate observed of the j th column in \mathbf{X} , $b = 1, \dots, B$.

Step 2: For each matrix \mathbf{X}_b , $b = 1, \dots, B$, obtain a sub-matrix \mathbf{X}_b^{sub} by sub-sampling the rows with the number of rows comparable to the number of columns in \mathbf{X}_b . Apply equations (6)–(8) iteratively to obtain $\hat{\mathbf{R}}_b^E$. In each iteration, apply a covariance shrinkage method to $\hat{\mathbf{\Sigma}}_b^{G,sub}$ to obtain $\hat{\mathbf{\Sigma}}_b^{G,sub,shrunken}$ which goes back into the next iteration as the “new” $\hat{\mathbf{\Sigma}}_b^{G,sub}$. $\hat{\mathbf{\Sigma}}_b^{G,sub}$ is the estimate of the row covariance matrix of \mathbf{X}_b^{sub} . This iterative procedure is initialized with a Pearson correlation matrix for \mathbf{R}^E , and terminates when the difference in

log likelihood of $\hat{\Lambda}$ in the last two iterations do not exceed a specified threshold. The final estimate of \mathbf{R}^E is given by the bagged estimate $\hat{\mathbf{R}}^E = \frac{1}{B} \sum_{b=1}^B \hat{\mathbf{R}}_b^E$.

Step 3: Using $\hat{\mathbf{R}}^E$ obtained in Step 2, for each \mathbf{X}_b , estimate Σ_b^G using equation (6). Apply a covariance shrinkage method to $\hat{\Sigma}_b^G$ to obtain $\hat{\Sigma}_b^{G,shrunken}$. The final estimate of Σ^G is given by the bagged estimate $\hat{\Sigma}^G = \frac{1}{B} \sum_{b=1}^B \hat{\Sigma}_b^{G,shrunken}$.

In the real dataset analyses in Section 4, the log likelihood difference threshold is set to 0.01 and B is 500. The covariance shrinkage method by Schäfer and Strimmer (2006) was used in the estimation procedure. This method obtains a shrunken covariance matrix with optimal mean squared error within the class of linear combinations of the unbiased covariance estimator and a pre-specified target matrix, does not require specifications on the underlying distribution, and it is computationally efficient. The diagonal matrix with unequal covariances was selected as the target matrix to represent the simplest parsimonious structure of the gene covariance matrix. We note that other shrinkage methods can be explored based on which set of assumptions one believe is appropriate for the dataset and its analysis.

3 Real Datasets And Data Processing

3.1 Real Datasets

We use two publicly available microarray datasets to evaluate our method in inferring functional gene relationships.

Yeast dataset. This dataset comes from a study by Sabet et. al. (2004) to investigate the influence of histone modifications on gene regulation. It consists of gene expressions from a wild type yeast and seven histone mutation experiments. There are two to three replicate arrays for each experiment. The descriptions of the experimental conditions can be found in the Appendix, and the dataset is accessible through the NCBI Gene Expression Omnibus Database by the accession number GDS772.

Human Th cell dataset. This dataset is generated by Lund et. al. (2005) to identify the immediate genes that are differentially regulated in response to activation and Th1- or Th2-inducing cytokine (IL-12 or IL-4, respectively) at 2 and 6 h after initiation of polarization. The dataset consists of 16 experiments conducted using 5 related treatments at three time points besides the untreated cells. There are two to four replicated arrays for each experiment, with a total of 34 microarrays. The descriptions of the experimental conditions can be found in the Appendix.

3.2 Data Processing

Raw data from each dataset are first normalized using the robust multi-array average (RMA) method developed by Irizarry et al. (2003). To enable a verification of inferred *context-specific* gene relationships, we use a set of Gene Ontology (GO) annotated genes with a strong specificity of response to the experiments in each dataset. Genes with the same GO category can be considered as being functionally related. As each dataset consists of

both wild type (i.e. control) and treatment experiments, genes with a strong specificity of response can be identified as genes that respond differently across the experiments. Similarly motivated by Li and Wong (2001), we identify these genes as follows: (i) for each experiment, rank the genes by their average expression over replicates, (ii) for each gene, obtain the difference between the maximum and minimum ranks across the experiments, (iii) a gene is identified as highly variably expressed if this rank difference exceeds a specified threshold. We chose the top 20% of such genes (~530 genes) for the yeast dataset, and the top 10% of such genes (~600 genes) for the human datasets.

4 Results

In this section, we report the performance of Knorm correlation in an illustrative simulation dataset, a yeast dataset, and a human dataset. We use the GO functional annotations to biologically evaluate the validity of inferred functional gene relationships. Since there is no gold standard measure for gene relationships, we will use the Pearson coefficient as a comparison benchmark because of its similar interpretations to Knorm correlation in terms of gene relationships and also its widespread use. The results demonstrate the success of our proposed method in inferring functional gene relationships.

4.1 Simulation Dataset

In this simulation study, we demonstrate the need to take into account the column-wise dependencies when estimating the row dependencies in \mathbf{X} in the case when the column correlation matrix is known. Using two *correlated* row vectors (i.e. genes), this study illustrates improved accuracies of Knorm correlation estimates over that of the Pearson coefficients even with increasing column (i.e. experiment) dependencies. At each $p\%$ dependency level (with $p=1,\dots,100$), we first generate 1000 *i.i.d.* column vectors of dimension two, each from a bivariate normal distribution with zero means, unit variances and a correlation of 0.17. We then assign the first $1000p\%$ vectors to be the same as the first vector, with the remaining $1000(1-p)\%$ independent vectors remain unchanged. Putting these 1000 column vectors of dimension 2 into a matrix, we now obtain two row vectors of dimension 1000 with a true row correlation of 0.17, and $p\%$ of the vector components being identical. We next compute both Pearson coefficient and Knorm correlation of the two row vectors, and plot the estimates in blue and red respectively in Figure 2. The Knorm correlation was computed using equation (9) with the column correlation matrix known by the construction procedure of the row vectors at the $p\%$ dependency level.

Figure 2 shows the improved estimation accuracy of Knorm correlation over that of Pearson coefficient. The Knorm correlation estimate is closer to the true correlation of 0.17 and has a much smaller variance until we reach about an 80% dependency. The Pearson coefficient, on the other hand, fails rapidly in accuracy after an approximate 5% dependency between the row vector components. We also have similar observations for simulation studies with different values of true correlations, both negative and positive (besides the value 0.17), and we only present the simulation study with a true correlation of 0.17 here as an illustrative example.

4.2 Real Datasets

We use the Gene Ontology (GO) to evaluate the *biological* accuracy of the inferred functional gene relationships. Two genes are said to be annotated as functionally related if they share the same GO category. Using the correlation measure, functional relationships between any two genes are predicted based on the sign and magnitude of their correlation estimates. The magnitude reflects the extent of a gene pair's synchronous response to the experiments, whereas a positive sign indicates a parallel response and a negative sign suggests an opposite response.

For each dataset we compare the percentages of GO annotated functionally related gene pairs among the top ranking genes ordered by the absolute Pearson coefficient and the absolute Knorm correlation. The Knorm correlation is computed using the estimation procedure described in Section 2.3 and the Pearson coefficient is computed on expressions that are averaged over the replicates within each experiment (a common approach in practice). We further provide examples of gene pairs to explicitly illustrate the improvement in inferring gene relationships by the Knorm correlation. However, the overall performance of each correlation measure in inferring gene relationships should be assessed by the number of gene pairs found to be functionally related by GO annotations. Note that a good correlation measure would put functionally related gene pairs high on the list and thus report higher percentages.

4.2.1 Yeast dataset

The Knorm correlation reports consistently higher percentages of GO annotated functionally related gene pairs than those obtained by the Pearson coefficient, see Table 1. In the top 10, 30, 50 and 100 gene pairs in estimated correlations, the Knorm correlation identified respectively 30.0%, 43.3%, 38.0% and 34.0% gene pairs that are known to be functionally related by GO annotations whereas the Pearson coefficient only identified respectively 10%, 20%, 26% and 21% gene pairs to be functionally related gene pairs. The distinction is especially strong for the gene pairs with highly ranked correlations. It is worthwhile to note that the percentages of functionally related gene pairs from both the proposed method and Pearson approach in Table 1 decrease generally and the percentage differences become stable as more top gene pairs are considered. This occurs since the presence of more gene pairs with weaker gene relationships would dilute and consequently stabilize the percentages of functionally related gene pairs found.

Gene pairs with biologically validated gene relationships. We have discovered gene pairs whose functional relationships are correctly predicted by the Knorm correlations. We provide some examples of such gene pairs as follows. These gene pairs have a high Knorm correlation estimate but a low Pearson coefficient. (**MCM1, SWI5**). This gene pair has a Knorm correlation of 0.52, but a Pearson coefficient of only -0.08 . The positive correlation of 0.52 is supported by experimental studies showing that MCM1 is a direct regulator of SWI5 (Kumar et al., 2000; Lee et al., 2002) and also that a reduced acetylation of histone amino termini is associated with reduced transcription levels of SWI5 (Deckert and Struhl, 2002; Shimizu et al., 2003). Therefore expressions of SWI5 and MCM1 are expected to show positive correlation in this dataset where histone amino termini have been deleted or modified, and the Knorm correlation confirms this expectation. (**CKA1, PMC1**). Both genes are known to be involved in maintaining cell ion homeostasis and yeast growth. Knorm correlation provides a positive estimate of 0.53 that reflects their related roles, while

the Pearson coefficient gives an estimate of only -0.02 . (**HSF1, CTK3**). These genes are biologically expected to be synchronously involved in the regulation of transcription from RNA polymerase II promoter. This is revealed by a positive Knorm correlation of 0.49 , but not by the Pearson coefficient (-0.03).

4.2.2 Human *Th* cell dataset

We see from Table 2 that the Knorm correlation reports favorably higher percentages of gene pairs found to be GO annotated as functionally related than those obtained by the Pearson coefficient, especially for the very highly ranked gene pairs. We note that the percentages in human dataset are generally lower than those in the yeast dataset, which can be attributed to the current incomplete annotations in human genome.

Gene pairs with biologically validated gene relationships. We have again discovered gene pairs, as follows, whose functional relationships are correctly predicted by the Knorm correlations. (**APEX1, MSH6**). The negative correlation of -0.44 by Knorm correlation is supported by a recent study reporting that the expression of APE protein leads to the suppression of DNA mismatch repair and that the MSH6 protein was markedly reduced in the APE-expressing cells (Chang et al., 2005). The Pearson coefficient, on the other hand, fails to capture this relationship with a value of -0.18 (34th percentile). (**RB1, CDKN1A**). The positive correlation of 0.40 by the Knorm is supported by a recent study reporting that the retinoblastoma protein RB1 is a cooperating factor for the transcription factor MITF to activate the expression of the cyclin-dependent kinase inhibitor gene CDKN1A, that contributes to cell cycle exit and activation of the differentiation program (Carreira et al., 2005). Contrary to this fact, the Pearson coefficient yields a value of -0.05 .

5 Empirical Model Justification

A key assumption in our probability model is the *i.i.d.* standard normal assumption on the elements in Λ in equation (5). In this section, we provide an empirical justification of this assumption using the yeast dataset as an illustrative example. We examine the qq-plot of the elements in $\hat{\Lambda}$, estimated from the yeast dataset, against a standard normal distribution, in addition to performing a Kolmogorov-Smirnov (K-S) test on the elements in $\hat{\Lambda}$. $\hat{\Lambda}$ is computed by equation (5) using the mean and covariance matrices estimated for the yeast dataset.

Figure 3 shows a randomly selected qq-plot among those obtained from 300 replicated expression matrices constructed through the bootstrapping procedure. This qq-plot is suggestive of a standard normal distribution for the elements in $\hat{\Lambda}$ with a p-value of 0.2 for the K-S test. Overall we observe good qq-plots with an average p-value of 0.68 for the K-S tests.

6 Discussion

We discuss some practical considerations when applying the approach to real datasets. First is the gene selection performed in the real data analyses. Genes that respond to the underlying biology of the experiments can be regarded as genes of higher specificity than those that do not respond; they carry larger biological variations than the nuisance variations. As such, our real dataset analyses focus on genes with high variable expressions across the experiments. We recognize that while there are genes that respond to the experiments but do not exhibit varied expression changes, it would be difficult to distinguish them from non-responsive genes due to the microarray technology limitation. Using genes highly specific to the biological process under study is crucial for a meaningful interpretation of inferred gene relationships. Next is the inference of causal relationships from the Knorm correlation. Microarray data is limited in its capacity to infer causal gene relationships because it only provides a snapshot of gene activities, and a gene expression is an overall measure of a gene's responses in multiple interactions with other genes. Measures like correlation only seek to provide a first step in inferring functional gene relationships and whether the genes are *associated* with one another in the biological process under study. After gene relationships have been established, and if of further interest, other technologies may be employed to specifically determine their directional relationships.

Although our method works well in practice, we bear in mind that it comes with two main assumptions: the normal distribution and the Kronecker product covariance matrix. The normal assumption is not unique to our work but is a commonly accepted assumption in numerous microarray studies. Besides the justification provided in Section 5, we also investigated the robustness of the Knorm correlation against the normality assumption using simulated poisson distributed "gene expressions". Results indicated the Knorm correlation is rather robust, yielding comparable mean square errors to those computed on the normally distributed "expressions" (more details and results in Appendix). We note that there are inherent difficulties in the data structure posed by study design (often by decisions beyond our control) that make developing direct model justifications less straightforward. The singularity of the high dimensional (estimated) gene covariance matrix poses difficulties in using statistical tests involving likelihoods that require a determinant of the singular high-dimensional covariance matrix. A complete statistical justification for the Kronecker product covariance matrix from such data turn out to be a very challenging problem.

We note that it is not the main aim of the article to suggest that the Knorm correlation is the *best* measure for inferring gene relationships. Rather, this work suggests that a measure adjusted for dependencies between the experiments is a *better* measure than one not adjusted for them, e.g. Knorm correlation *versus* Pearson coefficient. Therefore, a comparison between various measures of gene relationships would not be meaningful as each measure is defined to capture different aspects of a relationship. We have, however, provide a comparison between the Euclidean distance (one that ignores experiment dependencies) and the Mahalanobis distances (one that takes into account experiment dependencies) using the yeast dataset to further illustrate that the latter is a better measure than the former; the Mahalanobis distance is computed using the experiment covariance matrix estimated by the procedure in Section 2.3. Results are provided in the Appendix.

In conclusion, this work demonstrates that considering experimental dependencies is important in making more accurate inferences on functional gene relationships and its

practical usefulness in real datasets. Since assumptions in our method are largely motivated from the nature of the dataset and the purpose of analysis, this work could serve as an initial point to develop other models with appropriate modifications.

References

- Alizadeh, A. A et al. (2000), “Distinct types of diffuse large B-cell Lymphoma Identified by Gene Expression Profiling”, *Nature*, 403, 503–511.
- Carreira, S., Goodall, J., Aksan, I., La Rocca, S. A., Galibert, M. D., Denat, L., Larue, L., and Goding, C. R. (2005), “Mitf Cooperates with Rb1 and Activates p21Cip1 Expression to Regulate Cell Cycle Progression”, *Nature*, 433(7027), 764–769.
- Chang, I. Y., Kim, S. H., Cho, H. J., Lee, D. Y., Kim, M. H., Chung, M. H., and You, H. J. (2005), “Human AP Endonuclease Suppresses DNA Mismatch Repair Activity Leading to Microsatellite Instability”, *Nucleic Acids Res.*, 33(16), 5073–5081.
- Cromer, A., Carles, A., Millon, R., Ganguli, G., Chalmel, F., Lemaire, F., Young, J., Dembele, D., Thibault, C., Muller, D., Poch, O., Abecassis, J., and Wasylyk, B. (2004), “Identification of genes associated with tumorigenesis and metastatic potential of hypopharyngeal cancer by microarray analysis”, *Oncogene*, 23(14), 2484 – 2498.
- De Munck, J. C., Huizenga, H. M., Waldorp, L. J., and Heethaar, R. M. (2002), “Estimating Stationary Dipoles From MEG/EEG Data Contaminated With Spatially Temporally Correlated Background Noise”, *IEEE Transactions on Signal Processing*, 50, 7, 1565–1572.
- Deckert, J., and Struhl, K. (2002), “Targeted Recruitment of Rpd3 Histone Deacetylase Represses Transcription by Inhibiting Recruitment of Swi/Snf, SAGA, and TATA Binding Protein”, *Mol. Cell Biol.*, 22(18), 6458–6470.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Bostein, D. (1998), “Cluster analysis and display of genome-wide expression patterns”, *Proc. Nat. Acad. Sci. USA*, 95, 196 – 212.
- Efron, B. (1993), *Introduction to the bootstrap*, Chapman & Hall, New York.
- Fraley, C. and Raftery, A. E. (2000), “Model-Based Clustering, Discriminant Analysis, and Density Estimation”, University of Washington, Center for Statistics, and the Social Sciences, Working paper No. 11.
- Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. (2002), “Co-clustering of Biological Networks and Gene Expression Data”, *Bioinformatics*, 18, S145 – S154.
- Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003), “Summaries of Affymetrix GeneChip Probe Level Data”, *Nucleic Acids Res.*, 31(4) e15.
- Keppel, G. (1973), *Design and analysis: A researcher’s handbook*, Prentice-Hall Inc., NJ.
- Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N., and Davidson, G. S. (2001), “A Gene Expression Map for *C. elegans*”, *Science*, 293(5537), 2087 – 2092.
- Kumar, R., Reynolds, D. M., Shevchenko, A., Shevchenko, A., Goldstone, S. D., and Dalton, S. (2000), “Forkhead Transcription Factors, Fkh1p and Fkh2p, Collaborate with Mcm1p to Control Transcription Required for M-phase”, *Curr. Biol.*, 10(15), 896–906.
- Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L.,

- Fraenkel, E., Gifford, D. K., and Young, R. A. (2002), “Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*”, *Science*, 298(5594), 799–804.
- Li, K–C. (2002), “Genome-wide Coexpression Dynamics: Theory and Application”, *Proc. Natl. Acad. Sci. USA*, 99(26), 16875–16880.
- Lund, R., Ahlfors, H., T., Kainonen, E., Lahesmaa, A. M. Dixon, C., and Lahesmaa, R. (2005), “Identification of the Genes Involved in the Initiation of Type 1 and 2 T Helper Cell Commitment”, *Eur. J. Immunol.*, 35(11), 3307–3319.
- M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein (1998). Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc Natl Acad Sci U S A*, Vol. 95, 14863–14868.
- Penrose, R. (1995), A generalized inverse for matrices, *Proc. Cambridge Phil. Soc.*, 51, 406–413.
- Ramoni, M., Sebastiani, P., and Cohen , P. R. (2002), “Bayesian Clustering by Dynamics”, *Machine Learning*, 47(1), 91–121.
- Roy, A., and Khatree, R. (2005), “On Implementaion of a Test for Kronecker Product Covariance Structure for Multivariate Repeated Measures Data”, *Statistical Methodology*, 2, 297–306.
- Sabet, N., Volo, S., Yu, C., Madigan, J. P., and Morse, R. H. (2004), “Genome-wide Analysis of the Relationship between Transcriptional Regulation by Rpd3p and the Histone H3 and H4 Amino Termini in Budding Yeast Genome-wide Analysis of the Relationship between Transcriptional Regulation by Rpd3p and the Histone H3 and H4 Amino Termini in Budding Yeast”, *Mol. Cell. Biol.*, 24(20), 8823 –8833.
- Schäfer , J. and Strimmer, K. (2006), “A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics”, *Statistical Applications in Genetics and Molecular Biology*, 4(1), Article 32.
- Shimizu, M., Takahashi, K., Lamb, T. M., Shindo, H., and Mitchell, A. P. (2003), “Yeast Ume6p Repressor Permits Activator Binding but Restricts TBP Binding at the *HOP1* Promoter”, *Nucleic Acids Res.*, 31(12), 3033–3037.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Tyer, V. R., Anders, K., Eisen, M. B. m Brown, P. O., Bostein, D., and Futcher, B. (1998), “Comprehensive identification of cell cycle regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization”, *Molecular Biology of the Cell*, 9(12), 3273 – 3297.
- Svantesson, T., and Wallace, J. W. (2003), “Tests for Assessing Multivariate Normality and the Covariance Structure of MIMO Data”, *ICASSP*, IV, 656–659.
- Timm, N. H. (1980), “Multivariate Analysis of Variance of Repeated Measurements”, In: P.R. Krishnaiah (ed), *Handbook of Statistics*, 1, 41 – 87 (New York, North Holland).
- Wu, H., Chen, Y., Liang, J., Shi, B., Wu, G., Zhang, Y., Wang, D., Li, R., yi, X., Zhang, H., Sun, L., Shang, Y. (2005), “Hypomethylation-linked activation of PAX2 mediates tamoxifen-stimulated endometrial carcinogenesis”, *Nature*, 438, 981 – 987.
- Zhou, X. J., Kao, M. C., Huang, H., Wong, A., Nunez–Iglesias, J., Primig, M., Aparicio, O.M, Finch, C. E., Morgan, T. E., and Wong, W. H. (2005), “Functional Annotation and Network Reconstruction through Cross-platform Integration of Microarray Data”, *Nature Biotechnology*, 23(2), 238-243.

Table 1. Percentages of gene pairs found to be GO annotated as functionally related from among the top ranking gene pairs identified by Knorm correlation and the Pearson approach for the yeast dataset.

No. of top ranking gene pairs	Yeast microarray dataset	
	Knorm correlation	Pearson coefficient
Top 10	30.0	10.0
Top 30	43.3	20.0
Top 50	38.0	26.0
Top 100	34.0	21.0
Top 500	26.4	21.8

Table 2. Percentages of gene pairs found to be GO annotated as functionally related from among the top ranking gene pairs identified by Knorm correlation and the Pearson approach for the human *Th* cell microarray dataset.

No. of top ranking gene pairs	Human <i>Th</i> cell microarray dataset	
	Knorm correlation	Pearson coefficient
Top 10	10.0	10.0
Top 30	10.0	3.3
Top 50	10.0	4.0
Top 100	5.0	2.0
Top 500	4.0	3.4

Figure 1. Scatter plots of gene expressions of approximately 530 GO annotated yeast genes between four experiments in a yeast histone mutation dataset (Sabet et. al., 2004). Axes represent RMA normalized gene expression values.

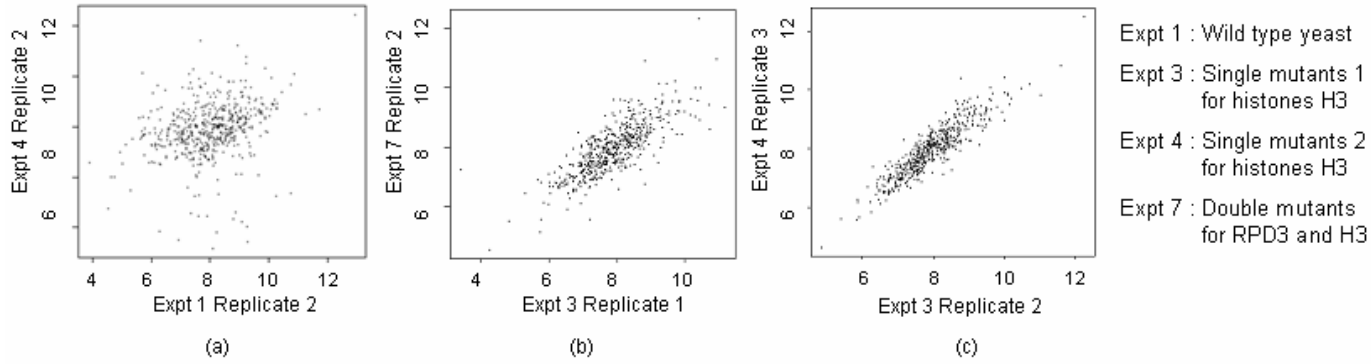


Figure 2. Correlation estimates of two simulated vectors by Knorm correlation (in red) and Pearson coefficients (in blue) in the presence of vector component dependencies at different levels. X-axis indicates the dependency level; Y-axis represents the estimated correlation. The true correlation value is 0.17.

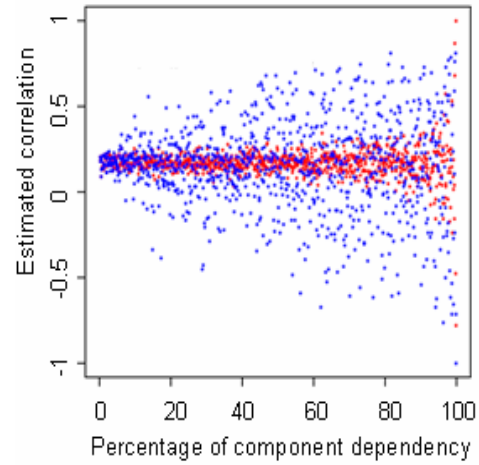
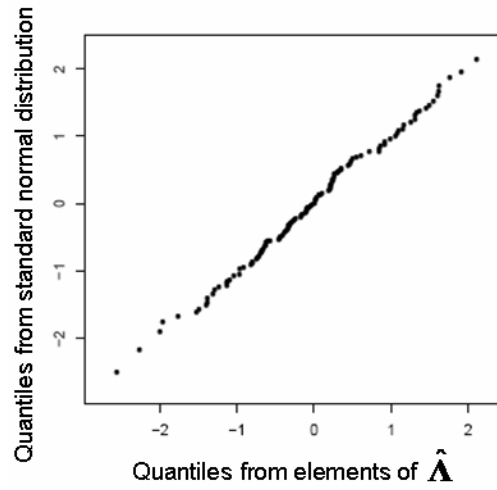


Figure 3. QQ-plot of elements in $\hat{\Lambda}$, estimated from a randomly selected expression matrix constructed through a bootstrapping procedure for the yeast dataset described in Section 3 against a standard normal distribution.



Appendix

1. IMPACT OF EXPERIMENTAL DEPENDENCIES ON PEARSON COEFFICIENTS

In addition to the results presented in Figure 2, we further investigate the adverse impact of experimental (column) dependencies on the Pearson coefficient of two *uncorrelated* genes across eight experiments (i.e. two row vectors of dimension 8). In this simulation study, we simulate gene expression matrices with rows corresponding to genes, and columns corresponding to experiments. Three different correlation matrices are used to describe the column dependencies: in Figure A1(a), the column correlation matrix is an identity matrix to simulate for row vectors with independent vector components; in Figure A1(b), the column correlation matrix consists of a mixture of zero and positive elements to simulate for row vectors with moderately positively correlated components; in Figure A1(c), the column correlation matrix consists of elements in a range of 0.8 to 1.0 to simulate for row vectors with highly positively correlated components. Each histogram in Figure A1 consists of 5000 Pearson coefficients, each computed from a pair of row vectors that are independently generated by a common multivariate normal distribution with zero means, unit variances and a specified correlation matrix as described above.

From Figures A.1(a)–A.1(c), we see a change in the distribution of the Pearson coefficients with increasing dependencies between the vector components. Figure A.1(a) shows a histogram representing the true distribution of Pearson coefficients between the two uncorrelated vectors. The distributions of the Pearson coefficients in Figures A.1(b) and A.1(c) become more skewed toward the larger absolute correlation values as dependencies between the components increase.

2. DERIVATION OF KRONECKER PRODUCT STRUCTURED COVARIANCE MATRIX OF GENE EXPRESSION MATRIX \mathbf{X} UNDER OUR MODEL

Theorem S1. Given a $p \times n$ matrix Λ of *i.i.d.* elements with mean 0 and unit variances, the covariance matrix of $\mathbf{X} = \mathbf{U} \mathbf{D}^{1/2} \Lambda \mathbf{P}^{1/2} \mathbf{V}^T$ is $\Sigma^G \otimes \Sigma^E$, where $\Sigma^G = \mathbf{U} \mathbf{D} \mathbf{U}^T$ and $\Sigma^E = \mathbf{V} \mathbf{P} \mathbf{V}^T$ are the singular value decompositions of Σ^G and Σ^E respectively.

Proof. Letting $\Omega = \Lambda \mathbf{P}^{1/2} \mathbf{V}^T$, we have $\mathbf{X} = \mathbf{U} \mathbf{D}^{1/2} \Omega$. Since Λ is a $p \times n$ matrix of *i.i.d.* elements with unit variances, the covariance matrix of Λ is \mathbf{I}_{pn} , or equivalently $\text{Cov}(\text{vec}(\Lambda)) = \mathbf{I}_{pn}$, where \mathbf{I}_{pn} is a $(pn) \times (pn)$ identity matrix. Now we consider the covariance matrix of Ω . Let Ω_{ij} be the (i,j) th element in Ω , \mathbf{e}_i be a p -dimensional column vector of zeroes except a value of 1 at the i th element and \mathbf{f}_j be a n -dimensional column vector of zeroes except a value of 1 at the j th element. Then we have

$$\begin{aligned}
\text{Cov}(\Omega_{ij}, \Omega_{i'j'}) &= \text{Cov}(\mathbf{e}_i^T \Lambda \mathbf{P}^{1/2} \mathbf{V}^T \mathbf{f}_j, \mathbf{e}_{i'}^T \Lambda \mathbf{P}^{1/2} \mathbf{V}^T \mathbf{f}_{j'}) \\
&= \mathbf{f}_j^T \mathbf{V} \mathbf{P}^{1/2} \text{Cov}(\mathbf{e}_i^T \Lambda, \mathbf{e}_{i'}^T \Lambda) \mathbf{P}^{1/2} \mathbf{V}^T \mathbf{f}_{j'} \\
&= \begin{cases} \mathbf{f}_j^T \mathbf{V} \mathbf{P}^{1/2} \mathbf{P}^{1/2} \mathbf{V}^T \mathbf{f}_{j'} & \text{when } i = i' \\ 0 & \text{when } i \neq i' \end{cases} \\
&= \begin{cases} \mathbf{f}_j^T \Sigma^E \mathbf{f}_{j'} & \text{when } i = i' \\ 0 & \text{when } i \neq i' \end{cases} \\
&= \begin{cases} (\Sigma^E)_{jj'} & \text{when } i = i' \\ 0 & \text{when } i \neq i' \end{cases}
\end{aligned} \tag{A1}$$

Therefore, the covariance matrix of Ω is $\mathbf{I}_p \otimes \Sigma^E$. Furthermore, letting X_{ij} to be the (i,j) th element in \mathbf{X} , we have

$$\begin{aligned}
\text{Cov}(X_{ij}, X_{i'j'}) &= \text{Cov}(\mathbf{e}_i^T \mathbf{U} \mathbf{D}^{1/2} \Omega \mathbf{f}_j, \mathbf{e}_{i'}^T \mathbf{U} \mathbf{D}^{1/2} \Omega \mathbf{f}_{j'}) \\
&= \mathbf{e}_i^T \mathbf{U} \mathbf{D}^{1/2} (\Sigma^E)_{jj'} \mathbf{I}_p \mathbf{D}^{1/2} \mathbf{U} \mathbf{e}_{i'} \\
&= (\mathbf{e}_i^T \Sigma^G \mathbf{e}_{i'}) (\Sigma^E)_{jj'} \\
&= (\Sigma^G)_{ii'} (\Sigma^E)_{jj'}
\end{aligned} \tag{A2}$$

Thus the covariance matrix of $\mathbf{X} = \mathbf{U} \mathbf{D}^{1/2} \Omega$ is $\Sigma^G \otimes \Sigma^E$. ■

3. DERIVATION OF MLEs IN EQUATIONS (6)–(8)

Theorem S2. Let the covariance matrices Σ^G and Σ^E be invertible. Given that $\text{vec}(\mathbf{X})$ follows a multivariate normal distribution with mean $\text{vec}(E(\mathbf{X})) = \text{vec}(\boldsymbol{\mu}\mathbf{1}^T)$ and covariance matrix $\Sigma^G \otimes \Sigma^E$, where $\mathbf{1}$ is a column vector of ones, the Maximum Likelihood Estimators (MLEs) of Σ^G , Σ^E and $\boldsymbol{\mu}$, conditional on remaining parameters, are given in equations (6)–(6) in Section 2 respectively.

Proof. By the assumed multivariate normal model, the log-likelihood function of an observed \mathbf{X} is

$$l(\mathbf{X}; \boldsymbol{\mu}, \Sigma^E, \Sigma^G) = -\frac{p}{2} \log |\Sigma^E| - \frac{n}{2} \log |\Sigma^G| - \frac{1}{2} \text{tr} \left((\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T) (\Sigma^E)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T)^T (\Sigma^G)^{-1} \right).$$

Then the first partial derivatives of $l(\mathbf{X}; \boldsymbol{\mu}, \Sigma^E, \Sigma^G)$ with respect to Σ^G , Σ^E and $\boldsymbol{\mu}$ are

$$\begin{aligned} \frac{\partial l}{\partial \Sigma^G} &= -\frac{n}{2} \left(\frac{\partial}{\partial \Sigma^G} \log |\Sigma^G| \right) - \frac{1}{2} \left(\frac{\partial}{\partial \Sigma^G} \text{tr} \left((\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T) (\Sigma^E)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T)^T (\Sigma^G)^{-1} \right) \right) \\ &= -\frac{n}{2} (\Sigma^G)^{-1} + \frac{1}{2} \left((\Sigma^G)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T) (\Sigma^E)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T)^T (\Sigma^G)^{-1} \right), \end{aligned}$$

$$\begin{aligned} \frac{\partial l}{\partial \Sigma^E} &= -\frac{p}{2} \left(\frac{\partial}{\partial \Sigma^E} \log |\Sigma^E| \right) - \frac{1}{2} \left(\frac{\partial}{\partial \Sigma^E} \text{tr} \left((\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T) (\Sigma^E)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T)^T (\Sigma^G)^{-1} \right) \right) \\ &= -\frac{p}{2} (\Sigma^E)^{-1} + \frac{1}{2} \left((\Sigma^E)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T)^T (\Sigma^G)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T) (\Sigma^E)^{-1} \right), \end{aligned}$$

$$\frac{\partial l}{\partial \boldsymbol{\mu}} = (\Sigma^G)^{-1} (\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T) (\Sigma^E)^{-1} \mathbf{1}.$$

As the normal distribution belongs to the exponential family and its log-density function is concave, the MLEs can be obtained by equating the above derivatives to zero and solving for Σ^G , Σ^E and $\boldsymbol{\mu}$, by which, we then obtain the MLEs of Σ^G , Σ^E and $\boldsymbol{\mu}$ conditional on remaining parameters, respectively as given in equations (6)–(8) in Section 2. Note that the Σ^E here is equivalent to the \mathbf{R}^E in equation (7) as Σ^E is assumed to have unit variances in the main article. ■

4. DESCRIPTION OF EXPERIMENTS

Yeast dataset (Sabet et. al. 2004)

- Experiment 1: Wild type yeast
- Experiment 2: Single mutants for RPD3P
- Experiment 3: Single mutants 1 for histones H3
- Experiment 4: Single mutants 2 for histones H3
- Experiment 5: Single mutants 1 for histones H4
- Experiment 6: Single mutants 2 for histones H4
- Experiment 7: Double mutants for RPD3 and histones H3
- Experiment 8: Double mutants for RPD3 and histones H4

Human *Th* cell dataset (Lund et. al. 2005)

- Experiment 1: Untreated cells
- Experiment 2: AntiCD3 + AntiCD28 (2h)
- Experiment 3: AntiCD3 + AntiCD28 (6h)
- Experiment 4: AntiCD3 + AntiCD28 (48h)
- Experiment 5: AntiCD3 + AntiCD28 + IL-12 (2h)
- Experiment 6: AntiCD3 + AntiCD28 + IL-12 (6h)
- Experiment 7: AntiCD3 + AntiCD28 + IL-12 (48h)
- Experiment 8: AntiCD3 + AntiCD28 + IL-12 +TGFbeta (26h)
- Experiment 9: AntiCD3 + AntiCD28 + IL-12 +TGFbeta (6h)
- Experiment 10: AntiCD3 + AntiCD28 + IL-12 +TGFbeta (48h)
- Experiment 11: AntiCD3 + AntiCD28 + IL-4 (2h)
- Experiment 12: AntiCD3 + AntiCD28 + IL-4 (6h)
- Experiment 13: AntiCD3 + AntiCD28 + IL-4 (48h)
- Experiment 14: AntiCD3 + AntiCD28 + IL-4 + TGFbeta (2h)
- Experiment 15: AntiCD3 + AntiCD28 + IL-4 + TGFbeta (6h)
- Experiment 16: AntiCD3 + AntiCD28 + IL-4 + TGFbeta (48h)

5. ROBUSTNESS OF KNORM CORRELATION

We additionally performed the following simulation study to examine the robustness of the Knorm correlation when the multivariate normality assumption of the gene expression matrix \mathbf{X} does not hold.

The data were simulated and analyzed as follows:

- We generated two datasets from the Normal and Poisson distributions respectively. The dependencies between the vector components are introduced by having the appropriate number of components to be identical. For example, a 20% component dependency indicates that the first 20 components in the vector of dimension 100 are the same.

In the *Normally* distributed dataset, at each $p\%$ dependency level (with $p=1, \dots, 100$), we first generate 100 *i.i.d.* column vectors of dimension 2, each from a bivariate normal distribution with zero means, unit variances and a correlation of 0.17, and then assign the first $100p\%$ vectors to be the same as the first vector (while remaining the last $100(1-p)\%$ independent vectors unchanged). Putting these 100 column vectors of dimension 2 into a matrix, we now obtain two row vectors of dimension 100 with a true row correlation of 0.17, and $p\%$ of the vector components being identical (the correlation between components within the vector is either 0 or 1).

In the *Poisson* distributed dataset, we first sampled pairs of values from two dependent Poisson processes (the correlation between the Poisson processes is 0.17). Similar to what we did in the normally distributed dataset, we then constructed the pairs of vectors and introduced the dependencies between the vector components.

Each dataset consists of 30 independent pairs of vectors at each component dependency level.

- By our construction procedure, the component covariance matrix for each vector is known. We then computed the Knorm correlation for each vector pair using equation (6) with the known component covariance matrix, and estimated the mean squared errors of the Knorm correlation with the true correlation 0.17.

Figure A.2. shows the mean squared errors of the Knorm correlation. The blue points represent the Normal dataset and red points represent the Poisson dataset. We see that the two sets of mean squared errors are close to each other, suggesting that the Knorm correlation is more or less robust against the normality assumption. Note that the normal dataset here is the same one used in Figure 2 in the revised manuscript.

6. RESULTS USING EUCLIDEAN DISTANCE VERSUS THE MAHALANOBIS DISTANCE (YEAST DATASET)

We perform an additional study to further demonstrate the advantage of adjusting for experiment dependencies using another distance metric. Here, we use the Mahalanobis distance and Euclidean distance on standardized expressions to infer gene relationships. The experiment correlation matrix used in computing the Mahalanobis distance is from the Knorm iterative estimation procedure. Table A.1 presents the results on the yeast dataset. It further reinforces the advantage of taking into account experiment dependencies in the distance calculations. We observe that the Knorm correlations yield higher percentages of GO functionally related gene pairs than that by the Mahalanobis distance except for the top 10 genes.

Figure A.1. Adverse impact of increasing component dependencies on the distribution of the Pearson coefficients for a pair of uncorrelated vectors. Each histogram consists of Pearson coefficients estimated from 5000 random pairs of uncorrelated vectors.

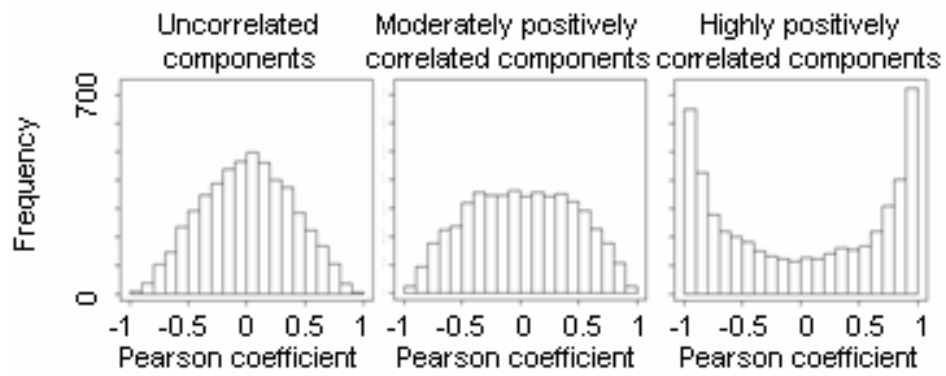


Figure A.2. A plot comparing the mean square errors of Knorm correlation from the normally distributed vectors (shown in blue) against that from the poisson distributed vectors (shown in red) across various component dependencies.

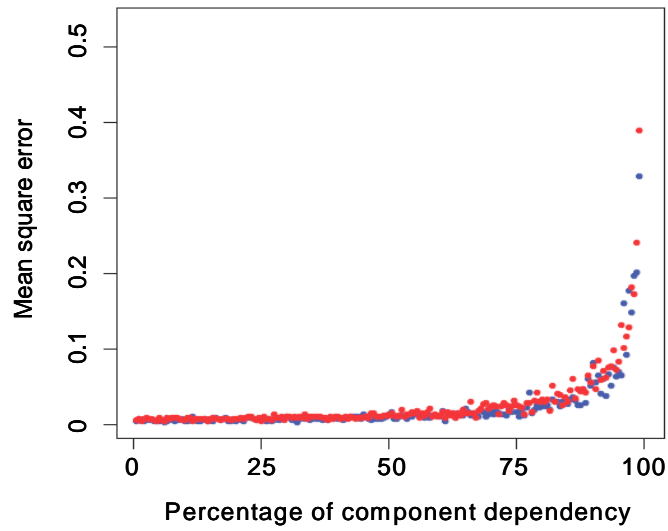


Table A.1. Comparison of percentages of GO functionally related gene pairs identified by the Euclidean and Mahalanobis distances for the **yeast dataset**.

No. of top ranking gene pairs	Mahalanobis distance	Euclidean distance	Knorm correlation	Pearson coefficient
Top 10	50.0%	20.0%	30.0%	10.0%
Top 30	36.7%	16.7%	43.3%	20.0%
Top 50	32.0%	26.0%	38.0%	26.0%
Top 100	26.0%	24.0%	34.0%	21.0%
Top 500	26.4%	21.2%	26.4%	21.8%