





## A Statistical Framework to Infer Functional Gene Relationships From Biologically Interrelated Microarray Experiments

Siew Leng Teng & Haiyan Huang


**To cite this article:** Siew Leng Teng & Haiyan Huang (2009) A Statistical Framework to Infer Functional Gene Relationships From Biologically Interrelated Microarray Experiments, Journal of the American Statistical Association, 104:486, 465-473, DOI: [10.1198/jasa.2009.0037](https://doi.org/10.1198/jasa.2009.0037)

**To link to this article:** <https://doi.org/10.1198/jasa.2009.0037>

 View supplementary material [↗](#)

 Published online: 01 Jan 2012.

 Submit your article to this journal [↗](#)

 Article views: 302

 View related articles [↗](#)

 Citing articles: 1 View citing articles [↗](#)

# A Statistical Framework to Infer Functional Gene Relationships From Biologically Interrelated Microarray Experiments

Siew Leng TENG and Haiyan HUANG

A major task in understanding biological processes is to elucidate the 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 presence of expression dependencies attributable to the biologically interrelated experiments, however, has been widely ignored. When unaccounted for, these (experiment) dependencies can result in inaccurate inferences of functional gene relationships, and hence incorrect biological conclusions. This article contributes a framework consisting of a model and an estimation procedure to infer 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 a Kronecker product covariance matrix to model the gene-experiment interactions. The resulting novel gene coexpression measure, named Knorm correlation, can be understood as a natural extension of the widely used Pearson coefficient when the experiment correlation matrix is known. Compared with the Pearson coefficient, the Knorm correlation has a smaller estimation variance. The Knorm is also asymptotically consistent with the Pearson coefficient. When the experiment correlation matrix is unknown, the Knorm correlation is computed based on the estimated experiment correlation matrix by an iterative estimation procedure. We demonstrate the advantages of the Knorm correlation in both simulation studies and real datasets. The Knorm correlation estimation procedure is implemented in an R package (Knorm) that is freely available from the Bioconductor website.

KEY WORDS: Experiment dependency; Knorm correlation; Kronecker product; Two-way dependencies.

## 1. INTRODUCTION

A major task in understanding biological processes is to elucidate the relationships between genes involved in the underlying biological pathways. Of particular interest are the functional gene relationships that arise as genes respond in different ways 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, Volo, Yu, Madigan, and Morse 2004; Cromer et al. 2004; Lund et al. 2005; Wu et al. 2005). As such, the functional gene relationships are context specific.

In this article, we address an understudied but critical issue in inferring the functional gene relationships: the presence of experiment dependencies in the 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 *coexist* 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 those between experiment 1 and the remaining experiments. This is due to the similar cellular changes induced by histone H3 mutations in experiments 3, 4, and 7 but *not* in experiment 1. Figure 1 also illustrates that the extent of dependencies in the

gene expressions between the experiments can vary according to the extent and type of histone mutations being introduced. The 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 the experiment dependencies is that they introduce redundancies that can overwhelm the important signals and lead to inaccurate estimates of gene relationships. In particular, the widely used Pearson coefficient (e.g., Eisen, Spellman, Brown, and Bostein 1998; Kim et al. 2001; Hanisch, Zien, Zimmer, and Lengauer 2002; Li, 2002; Zhou et al. 2005) suffers from an increased estimation variance and an almost random sign when experiment dependencies are unaccounted for (see Fig. 2 in Sec. 4). 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 in Sec. 4). 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 the experiment dependencies are naturally strong. A simulation study that further investigates the adverse impact of experiment dependencies on the Pearson coefficient is presented in S1 of the supplementary material.

Another fundamental issue faced in inferring gene relationships from real datasets is the complex data structure. The observed data structure in a typical dataset does *not* consist of replicates of expression matrices. Instead, it consists of replicates of gene expressions from each experiment. The number of replicates is often small (e.g., 2–3) and can be different for each experiment. The data structure is made more complicated by the copresence of biological and nuisance variations in each gene expression experiment.

---

Siew Leng Teng is a Ph.D. student from the Division of Biostatistics, University of California, Berkeley, CA 94720 (E-mail: [slteng@stat.berkeley.edu](mailto:slteng@stat.berkeley.edu)). Haiyan Huang is Assistant Professor, Department of Statistics, University of California, Berkeley, CA 94720 (Email: [hhuang@stat.berkeley.edu](mailto:hhuang@stat.berkeley.edu)). The work of Teng and Huang are supported by NIH R01GM075312. The authors thank X. Jasmine Zhou in University of Southern California for her great help on providing the biological evaluations of the results, in addition to many inspiring discussions. The authors thank Riikka Lund for providing access to their data, and Yu Huang in University of Southern California for providing the Gene Ontology nodes. The authors also thank the associate editor, the referees, Wing Wong, Terry Speed, and Peter Bickel for their helpful comments and suggestions.

---

© 2009 American Statistical Association  
Journal of the American Statistical Association  
June 2009, Vol. 104, No. 486, Applications and Case Studies  
DOI 10.1198/jasa.2009.0037

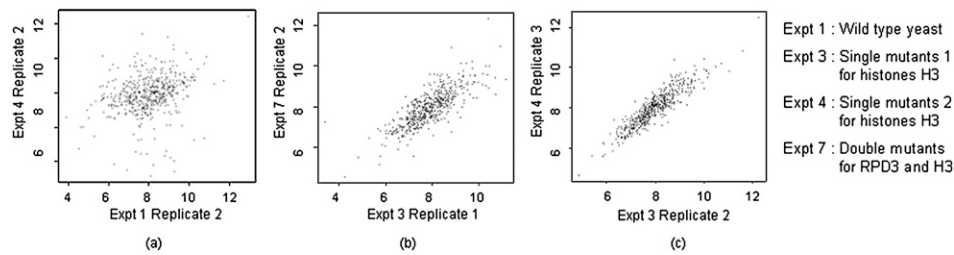


Figure 1. Scatter plots of gene expressions of approximately 530 GO annotated yeast genes between four experiments in a yeast histone mutation dataset (Sabet et al. 2004). Axes represent RMA normalized gene expression values.

Motivated by the previous issues, this article attempts to address the following questions. “How do we model and estimate the *experiment* dependencies from the complex data structure, and how would the *gene correlation measure* be adjusted for experiment dependencies?” To answer these questions, we present a framework consisting of a statistical model and a practical estimation procedure. The model is a linear additive model with random gene, experiment, and gene-experiment interaction effects. A Kronecker product covariance matrix is used to represent the dependencies among the interaction terms, which allows modeling and estimating the *experiment* and *gene* covariance matrices simultaneously. This model also delineates the biological and nuisance variations in the gene expression data. The expression matrices are reconstructed by bootstrapping the replicates observed in each experiment.

We note that although two-way dependencies in matrix data had been studied (e.g., Timm 1980; Fang and Zhang 1990; de Munck, Huizenga, Waldorp, and Heethaar 2002; Svantesson and Wallace 2003; Roy and Khattree 2005), the modeling of coexisting gene and experiment dependencies is relatively unexplored in conjunction with the complex data structure and the “large  $p$ , small  $n$ ” problem. The “large  $p$ , small  $n$ ” problem in gene expression data is well known and it makes impossible to directly estimate the experiment and gene covariance matrices by the maximum likelihood estimators (MLEs) obtained through the previous linear additive model. In other words, the estimation algorithms employed in the existing literature [e.g., the one outlined in the work of de Munck et al. (2002) where the data matrix is  $150 \times 100$  – a matrix with more balanced dimensions in row and column] cannot be applied to the microarray data. There are also existing works that model specific spatial dependencies between experiments [e.g., the Fourier series approach by Spellman et al. (1998) and the autoregressive models by Ramoni, Sebastiani, and Cohen (2002)]. These approaches, however, often require specific assumptions that may not be generally satisfied by typical datasets used in pathway or gene function network studies (e.g., as illustrated in Fig. 1). Our work, on the other hand, is motivated to model the generally monotonic and varied experiment dependencies observed in many real datasets.

By combining a gene subsampling strategy and a covariance shrinkage technique, we develop a practical estimation procedure to mitigate issues due to the “large  $p$ , small  $n$ ” problem. The shrinkage technique helps in getting robust estimates of the covariance matrix and its inverse, whereas the subsampling (of genes) strategy helps to diminish the dimension difference

between genes and experiments and thereafter improves the precision of the estimates. The newly derived gene correlation measure, named Knorm correlation then adjusts for experiment dependencies by weighting the gene expression proportionally to the partial correlation between the experiments. The Knorm correlation has smaller estimation variability than the Pearson coefficient, especially when there are only a few replicates available for each experiment. When the experiments are uncorrelated, the Knorm correlation simplifies to the Pearson coefficient.

Regarding the model validation, we note that existing methods for testing the Kronecker structure of the covariance matrix are not applicable to the data in our context. These methods mainly use the likelihood ratio test to test a probability model with the Kronecker structured covariance matrix against one with the full unrestricted covariance matrix (e.g., Svantesson and Wallace 2003; Roy and Khattree 2005). This general approach uses the following assumptions: (1) independent replicates of a random matrix are available, (2) there is sufficient data to estimate the full unrestricted covariance matrix, and (3) likelihood computation is possible. The data structure in our study, however, does not meet these assumptions. In particular, the singularity of the high dimensional (estimated) gene covariance matrix poses a difficult problem to evaluating the likelihood where a determinant of the covariance matrix is required. A complete statistical justification of our model based on the high dimensional and complex structured microarray data turns out to be a very challenging problem. Empirical evidence is employed to support our model in this article.

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. Overall, the Knorm correlation reports higher percentages of functionally related Gene Ontology (GO) annotated gene pairs in real datasets. Using the yeast dataset as an illustrative example, Section 5 provides empirical justifications of our model. Finally, Section 6 discusses several practical and technical issues encountered in practice.

## 2. STATISTICAL FRAMEWORK

### 2.1 Statistical Model

Let  $X_{ijk}$  represent 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 the gene effect is a *random* effect that consists of three components:

- (1) A *fixed* component  $\mathbf{G}$  that measures the average gene expression level. This component depends only on the gene and is independent of the experiments.
- (2) A *random* component that accounts for the **nuisance variation** arising from nuisance effects, such as a measurement error. This component explains changes in gene expression that are independent of the experiments.
- (3) A *random* component that accounts for the **biological variation** in a gene expression. This variation is triggered as a gene responds to the different experiments. This component thus also represents the gene-experiment interaction effect that would dominate the random component in (2) when the gene expresses itself differently in the experiments.

These three assumptions also apply to the experiment effect. Putting together the previous components, a linear additive model of the 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 the fixed components representing the individual gene and experiment effects, respectively,  $\gamma_{ij}^{GE}$  denotes the random gene-experiment interaction effect, and  $\varepsilon_{ijk}$  is a random term representing the nuisance effects in  $X_{ijk}$ . The nuisance terms  $\varepsilon_{ijk}$  are iid with zero means and are independent of the interaction terms  $\gamma_{ij}^{GE}$ . For an appropriately reconstructed  $p \times n$  gene expression matrix  $\mathbf{X}$  (to be elaborated in Section 2.3), we can rewrite the 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),  $\mathbf{E} = \mathbf{1} \cdot (v_1 \dots v_n)$ ,  $\mathbf{\Gamma}^{GE} = (\gamma_{ij}^{GE})_{i=1, \dots, p, j=1, \dots, n}$  is a zero-mean random matrix with elements representing the gene-experiment effects,  $\boldsymbol{\varepsilon}$  is a zero-mean random matrix with elements representing iid 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 respectively denote the gene and experiment dependencies, 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 nuisance effects,  $\text{vec}(\mathbf{X}^T)$  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  $\boldsymbol{\Sigma}^G \otimes \mathbf{I}_n$ , where  $\mathbf{I}_n$  denotes the  $n \times n$  identity matrix. The covariance matrix  $\boldsymbol{\Sigma}^G \otimes \mathbf{I}_n$  is the dependency structure used in multivariate normal models that result in the Pearson coefficient. Second,  $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^E$  can be interpreted as a dependency structure between the interaction terms caused by a dual projection of a matrix of iid random variables with unit variances onto the eigenspaces determined by  $\boldsymbol{\Sigma}^G$  and  $\boldsymbol{\Sigma}^E$ . This can be seen from the fact that  $\mathbf{\Gamma}^{GE}$  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 iid 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 components in the singular value decompositions of  $\boldsymbol{\Sigma}^E$  and  $\boldsymbol{\Sigma}^G$ . Consequently, the covariance matrix of  $\mathbf{\Gamma}^{GE}$  is  $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^E$ . Following Equation (3),

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

When nuisance effects are negligible,

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

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 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 iid  $N(0,1)$  random variables and the nuisances epsilon are negligible, then from (4) it is straightforward to show that  $\text{vec}(\mathbf{X}^T)$  is multivariate normal with mean  $\mathbf{G} + \mathbf{E}$  and covariance matrix  $\boldsymbol{\Sigma}^G \otimes \boldsymbol{\Sigma}^E$  (Fang and Zhang 1990).

## 2.2 Parameter Estimation and Knorm Correlation

For the model in (2) to be identifiable, we assume that  $(\mathbf{E})_{ij} = E_j = 0$  and that  $\boldsymbol{\Sigma}^E$  has unit diagonal elements (i.e., the experiment effect is random with zero mean and unit variance). These assumptions are reasonable in the context of gene expression datasets as normalized gene expressions will have the same mean and variance in each experiment [e.g., Robust Multi-array Analysis (RMA) approach 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 thus should be based on the nature of the data and the purpose of analysis. For the remainder of this 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 MLEs of  $\boldsymbol{\Sigma}^G$ ,  $\mathbf{R}^E$ , and  $\boldsymbol{\mu}$ , conditional upon the remaining parameters, are

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

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

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

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



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

where  $\mathbf{W}$  is a diagonal matrix with the same diagonal elements in  $\hat{\Sigma}^G$ . The MLE of  $\mu$  in Equation (8), conditional on  $\Sigma^G$  and  $\mathbf{R}^E$ , is unbiased and consistent. Similarly, the MLEs of  $\Sigma^G$  and  $\mathbf{R}^E$ , conditional upon the remaining two parameters, are also consistent estimators. The reader is referred to Fang and Zhang (1990) and S2 of the supplemental material for a detailed derivation of Equations (6)–(8).

**Knorm correlation.** The Knorm correlation is defined by Equation (9), and has several appealing interpretations.

- (1) When experiments are uncorrelated (i.e.,  $\mathbf{R}^E = \mathbf{I}_n$ ), the Knorm correlation reduces to the matrix of Pearson correlations. This follows from the previous argument that the covariance matrix  $\Sigma^G \otimes \Sigma^E$  is a natural extension of  $\Sigma^G \otimes \mathbf{I}_n$ . A model with the latter covariance matrix assumes an absence of experiment dependencies that result in the Pearson coefficient.
- (2) The Knorm correlation is the correlation between transformed gene expression profiles. The transformation is achieved through a projection onto the eigenspace of  $\mathbf{R}^E$  that removes the experiment dependencies in the gene expressions.
- (3) The Knorm correlation is a weighted Pearson coefficient of gene expressions with weights proportional to the partial correlation between the experiments. This interpretation comes from the fact 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 the remaining experiments.

### 2.3 Practical Estimation Procedure for Real Dataset Applications

The “large  $p$ , small  $n$ ” problem of the high dimensional real datasets can result in large estimation errors and unstable covariance matrix estimates, especially when  $\Sigma^G$  and  $\mathbf{R}^E$  are not sparse. To mitigate this impact, we develop a practical estimation procedure consisting of a row (i.e., gene) subsampling 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. Because only replicates of each column (i.e., experiment) in  $\mathbf{X}$  are observed instead of matrix replicates, a parametric bootstrapping technique (Efron 1993) is used to reconstruct the data matrices. Based on the model in (2), this essentially bootstraps the nuisance residuals  $\varepsilon$ . We also note that the interaction effects are expected to dominate the nuisance effects in our study as only the differentially expressed genes are of interest. 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 use of Equations (6)–(8) with a row subsampling technique (to enable a comparable number of rows and columns in estimation) and a covariance shrinkage technique (to stabilize the estimated covariance matrices). Finally, the third step uses the  $\hat{\mathbf{R}}^E$  obtained from the previous step to estimate the  $\Sigma^G$ .

We describe the full implementation of the procedure, followed by more details on the row subsampling and covariance shrinkage technique.

**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 of the  $j$ th experiment,  $b = 1, \dots, B$ .

**Step 2.** For each matrix  $\mathbf{X}_b$ ,  $b = 1, \dots, B$ , obtain a submatrix  $\mathbf{X}_b^{\text{sub}}$  by the row subsampling technique (see details later).

Apply Equations (6)–(8) iteratively to obtain  $\hat{\mathbf{R}}_b^E$  and  $\hat{\Sigma}_b^{G, \text{sub}}$  from  $\mathbf{X}_b^{\text{sub}}$ . Note that  $\hat{\Sigma}_b^{G, \text{sub}}$  is the estimate of the row covariance matrix of  $\mathbf{X}_b^{\text{sub}}$ . In each iteration, apply a covariance shrinkage method (see details later) to  $\hat{\Sigma}_b^{G, \text{sub}}$  to obtain  $\hat{\Sigma}_b^{G, \text{sub}*}$ , which goes back into the next iteration as the “new”  $\hat{\Sigma}_b^{G, \text{sub}}$ . This iterative procedure is initialized with a Pearson correlation matrix for  $\mathbf{R}^E$  and terminates when the difference in the log-likelihood in the last two iterations do not exceed a specified threshold. The final estimate of  $\mathbf{R}^E$  is given by the bagged estimate  $\hat{\mathbf{R}}^E = 1/B \sum_{b=1}^B \hat{\mathbf{R}}_b^E$ . In the real dataset analyses in Section 4, the threshold of the log-likelihood difference is set to 0.01 and  $B$  is 500.

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

**Row subsampling.** This technique simply randomly selects a subset of rows to create submatrices  $\mathbf{X}^{\text{sub}}$  with comparable row and column dimensions. From our extensive simulation studies, this technique was observed to greatly reduce estimation errors in the experiment correlation matrix estimate.

**Covariance shrinkage.** A covariance shrinkage method is generally used to obtain a stable estimate of the sample covariance matrix and its inverse. Without using shrinkage, applying Equations (6)–(8) directly could result in large estimation errors due to more vanishing eigenvalues of the covariance matrix as its dimension,  $p$ , becomes much larger than  $n$  (the well-known “large  $p$  small  $n$ ” problem). The reader can refer to Schäfer and Strimmer (2005) for a simulation study characterizing this phenomenon. In particular, we note that the application of a covariance shrinkage method is critical in some extreme situations (e.g., when many genes are highly dependent on one another), which results in an almost singular and nonsparse gene covariance matrix even for a small subset of genes. More discussions on this situation can be found in S4 of the supplemental material. Briefly, for our method to be generally applicable, it is necessary to combine the shrinkage and the row subsampling techniques.

The particular covariance shrinkage method used in our practical estimation procedure is the one developed by Schäfer and Strimmer (2005). In brief, this approach obtains a shrunken covariance estimate in the form of a linear combination of the unstructured MLE covariance matrix and a parsimonious structured matrix by optimizing its mean squared error with some bias introduced (Ledoit and Wolf 2004; Schäfer and Strimmer 2005). We used the diagonal matrix with unequal covariances as the target matrix to represent the simplest

parsimonious structure of the gene covariance matrix. The advantages of this approach are that it does not require specification of the underlying distribution and is computationally efficient. However, other shrinkage methods can also be explored depending on which set of assumptions one believes is appropriate for the dataset and its analysis (e.g., Daniels and Kass 2001; Ledoit and Wolf 2004; Schäfer and Strimmer 2005; Bickel and Levina 2006).

### 3. REAL DATASETS AND DATA PROCESSING

#### 3.1 Real Datasets

We present the applications to two publicly available microarray datasets to illustrate 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 and the dataset can be found in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus Database by the accession number GDS772.

*Human *Th* cell dataset.* This dataset was generated by Lund et al. (2005) to identify the 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 hours after initiation of polarization. The dataset consists of 16 experiments conducted using five 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.

Besides these two datasets, we also applied our method to two additional human cancer datasets to see a broader range of applications of the approach. Results of these applications can be found in S3 of the supplemental material.

#### 3.2 Data Processing

The raw data from each dataset are first normalized using the RMA method developed by Irizarry et al. (2003). To verify the inferred context-specific gene relationships, we use a set of GO annotated genes with strong specificity of response to the experiments in each dataset. We consider genes with the same GO category at level 6 or more 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: (1) for each experiment, rank the genes by their average expression over replicates; (2) for each gene, obtain the difference between the maximum and minimum ranks across the experiments; (3) 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 *Th* cell datasets.

### 4. RESULTS

In this section, we report the performance of the Knorm correlation in an illustrative simulation dataset, the yeast dataset, and the human *Th* cell dataset. We use the GO functional annotations to biologically evaluate the validity of the inferred functional gene relationships. Because there is no gold standard measure for gene relationships, we will use the Pearson coefficient as a comparison benchmark because of its similar interpretation to the Knorm correlation and its widespread use. Overall, the Knorm correlation achieves an improved accuracy in correlation estimates in simulation studies, and reports higher (if not comparable) percentages of functionally related GO annotated gene pairs in the real datasets.

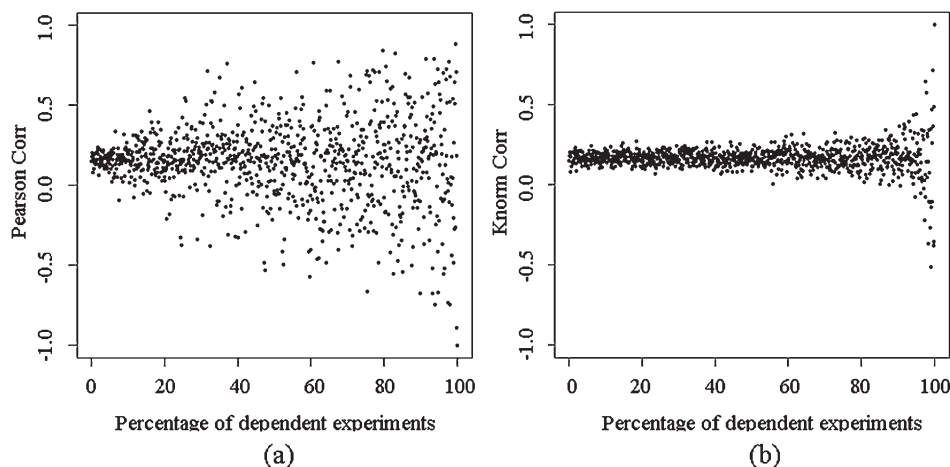


Figure 2. Correlation estimates of two simulated vectors by (a) Pearson coefficients and (b) Knorm correlation in the presence of vector component dependencies at different levels. X-axis indicates the dependency level; Y-axis represents the estimated correlation. The true correlation value is 0.17.

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

#### 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 the data matrix  $\mathbf{X}$  in the case when the column correlation matrix is known. Using two *correlated* row vectors (i.e., genes), this study illustrates an improved accuracy in the correlation estimate by the Knorm correlation over that of the Pearson coefficients despite an increasing number of perfectly correlated columns (i.e., increasing experiment dependencies). For each value of  $p\%$ , where  $p\%$  represents the percentage of columns being identical to one another,  $p = 1, \dots, 100$ , we first generate 1,000 iid 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  $1,000p\%$  vectors to be the same as the first vector, with the remaining  $1,000(1 - p\%)$  independent vectors unchanged. Putting these 1,000 column vectors of dimension 2 into a matrix, we now obtain two row vectors of dimension 1,000 with a true row correlation of 0.17. We next compute both the Pearson coefficient and Knorm correlation of the two row vectors, and plot these estimates respectively in Figure 2(a) and Figure 2(b). 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 the Pearson coefficient. The Knorm correlation estimate is closer to the true correlation of 0.17 and has a much smaller variance until we reach to about an 80% dependency. The Pearson coefficient, on the other hand, fails rapidly in accuracy after an approximate 5% dependency

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

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.

An additional simulation study was carried out on two general types of gene correlation matrices encountered in practice—sparse and nonsparse matrices—to further demonstrate the superiority of the Knorm estimation procedure over the Pearson approach. The Knorm approach consistently achieves smaller Frobenius norms between the estimated and true gene correlation matrices over the Pearson approach even when the number of available replicates is small. Results of this additional study can be found in S4 of the supplemental material.

#### 4.2 Real Datasets

We use the 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 an opposite response.

For each dataset we compare the percentages of gene pairs found to be functionally related by GO annotations 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 gene expressions 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 gene pairs found to be functionally related by GO annotations than those obtained by the Pearson coefficient. From Table 1, among the top 10, 30, 50, and 100 gene pairs, the Knorm correlation identified respectively 30.0%, 43.3%, 38.0%, and 34.0% of gene pairs that are functionally related by GO annotations, whereas the Pearson coefficient only identified respectively 10%, 20%, 26%, and 21% of such 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 because the presence of more gene pairs with weaker gene relationships would dilute and consequently stabilize the percentages of functionally related gene pairs found.

We have discovered gene pairs whose functional relationships are correctly predicted by the Knorm correlations. These gene pairs have a high Knorm correlation estimate but a low Pearson coefficient. Examples of such gene pairs are as follows.

**(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). It is also known that a reduced acetylation of histone amino termini is associated with reduced transcription levels of SWI5 (Deckert and Struhl 2002; Shimizu, Takahashi, Lamb, Shindo, and Mitchell 2003). Therefore, expressions of SWI5 and MCM1 are expected to show a positive correlation in this dataset where the 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. The Knorm correlation provides a positive estimate of 0.53 that reflects their related roles, whereas 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 higher percentages of gene pairs found to be functionally related by GO annotations than those obtained by the Pearson coefficient, especially for the very highly ranked gene pairs. The percentages in the human dataset are observed to be lower than those in the yeast dataset, which can be likely due to the current relatively poor annotations of the human genome.

We have again discovered gene pairs, whose functional relationships are correctly predicted by the Knorm correlation. Examples of such genes include:

**(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  $-0.18$ .

**(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  $-0.05$ .

We note that we also performed similar percentage comparisons based on genes ranked without using the absolute values of correlations. The improved performance and accuracy in inferring gene relationships by the Knorm correlation

remain. Results are presented in S5 of the supplemental material.

## 5. EMPIRICAL MODEL JUSTIFICATION

Our probability model assumes that the elements of  $\Lambda$  in Equation (5) are iid standard normal. In this section, we provide an empirical justification of this assumption using the yeast dataset as an illustration. We examine the quantile-quantile (Q-Q) plot of the elements in  $\hat{\Lambda}$ , estimated from the yeast dataset, against a standard normal distribution; we also perform a Kolmogorov-Smirnov (K-S) test on the elements in  $\hat{\Lambda}$ .  $\hat{\Lambda}$  is computed by Equation (5) using the estimated mean and covariance matrices.

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

To some extent, the previous study reflects positively on the validity of our method. To further support our model, we attempt a test of the Kronecker product structured covariance matrix assumption as follows. As we have discussed in Section 1, existing methods for testing the Kronecker structure are not applicable to the gene expression data. Those methods mainly use the likelihood ratio test to test a probability model with the Kronecker structured covariance matrix against one with the full unrestricted covariance matrix (e.g., Svantesson and Wallace, 2003; Roy and Khattree, 2005). For gene expression data, it is almost impossible to estimate the full unrestricted covariance matrix because of the unavailability of independent matrix replicates. Furthermore, the singularity of the high

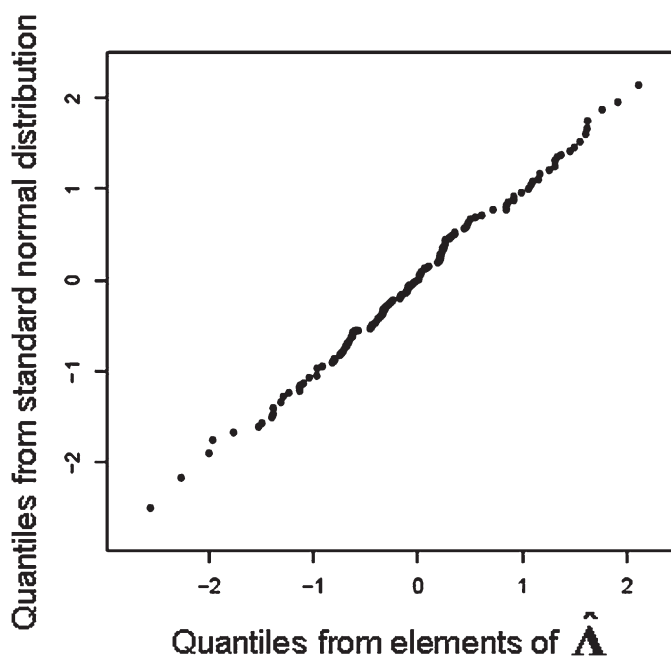


Figure 3. Q-Q plot of elements in  $\hat{\Lambda}$ , estimated from a randomly selected expression matrix constructed through a bootstrapping procedure for the yeast dataset described in Section 3 against a standard normal distribution.



dimensional (estimated) gene covariance matrix prevents the computation of the likelihood. It turns out that a complete statistical justification of our model based on the high dimensional data structure is a very challenging problem.

We seek empirical evidence to support the need to take into account the experiment dependencies. Specifically, we examine the *pseudo* log-likelihood ratio for the null hypothesis of the  $\Sigma^G \otimes \mathbf{I}_n$  covariance structure against the alternative hypothesis of the  $\Sigma^G \otimes \mathbf{R}^E$  structure, under the assumption that the expression data matrix is normally distributed. Here,  $\mathbf{I}_n$  denotes the identity matrix and  $\mathbf{R}^E$  is the unrestricted experiment correlation matrix. The  $\Sigma^G \otimes \mathbf{I}_n$  model ignores the experiment dependencies and results in the Pearson coefficient. The  $\Sigma^G \otimes \mathbf{R}^E$  model takes into account the experiment dependencies and results in the Knorm correlation.

The joint likelihoods of the data under both hypotheses are either not computable or unreliable because the estimated  $\Sigma^G$  is high dimensional and singular (or nonsingular but with a very small determinant). So we instead use an *approximate pseudo joint log-likelihood ratio* to compare the two models. For a vector  $\mathbf{y} = (y_1, \dots, y_q)^T$ , its pseudo joint log-likelihood ratio is given by

$$\log LR = \log f_{H_0} - \log f_{H_1},$$

where  $f_{H_T}$  is the pseudo joint likelihood from the multivariate matrix normal distribution with the corresponding covariance matrix under the null ( $T = 0$ ) and alternative ( $T = 1$ ) hypotheses, respectively, and  $f_{H_T}(y_1, \dots, y_q) = \prod_{i=1}^q f_{H_T}(y_i | y_j \text{ for all } \{Y_i, Y_j\} \in E)$  with  $E$  being the set of dependencies between random variables in  $\mathbf{Y}$ . For computation purposes, we approximate the pseudo log-likelihoods by considering a subset of  $E$  (e.g., for each  $Y_i$ , we used the two  $Y_j$ 's with the strongest dependencies with  $Y_i$  in the following reported results). We used the Bayesian information criterion (BIC) for the model selection. Following the above and letting  $E$  be the set of dependent genes, the human *Th* cell dataset with 30 bootstrapped data matrices yields a BIC adjusted approximate pseudo log-likelihood ratio of  $-23827.3$ , which suggests that the  $\Sigma^G \otimes \mathbf{R}^E$  model would likely be a more appropriate fit for the human *Th* cell data than the  $\Sigma^G \otimes \mathbf{I}_n$  model.

## 6. DISCUSSION

There are several practical considerations when applying the approach to real datasets. One issue is the gene selection performed in the real data analyses. The genes with highly variable expressions are regarded as genes that respond to the experiments and so they carry larger biological variation than the nuisance effect variation. 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 the nonresponsive genes due to microarray technology limitations. Being able to use genes that are highly specific to the biological process under study is crucial for a meaningful interpretation of the inferred gene relationships. Another issue is of making inference on causal relationships from the Knorm correlation. The microarray data are limited in its capacity to infer causal gene relationships because it only provides a snapshot of gene activities and because a gene expression is

only an overall measure of a gene's responses in multiple interactions with other genes. Measures such as correlation only provide a first step in inferring functional gene relationships by establishing if the genes are *associated* with one another in the biological process under study. Other technologies may be further employed to specifically determine the directional relationships if such associations exist.

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. We have investigated the robustness of the Knorm correlation against the normality assumption using simulated Poisson distributed "gene expressions." These results are presented in S6 of the supplemental material. For the Kronecker covariance matrix assumption, there are inherent difficulties in the data structure posed by the study design (often by decisions beyond our control) that make developing direct model justifications less straightforward, see Section 5.

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 gene relationship. We have, however, provided 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 demonstrate 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 S7 of the supplemental material.

In conclusion, this work demonstrates that considering experiment dependencies is important in making accurate inferences on functional gene relationships and has practical use in real datasets.

[Received July 2006. Revised March 2008.]

## REFERENCES

- Bickel, P., and Levina, E. (2006), "Regularized Estimation of Large Covariance Matrices," Technical report, Department of Statistics, University of California, Berkeley.
- 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 Research*, 33, 5073–5081.
- Cromer, A., Carles, A., Millon, R., Ganguli, G., Chalmel, F., Lemaire, F., Young, J., Demb  l  , 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, 2484–2498.
- Daniels, M. J., and Kass, R. E. (2001), "Shrinkage Estimators for Covariance Matrices," *Biometrics*, 57, 1173–1184.

- 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, 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," *Molecular and Cellular Biology*, 22, 6458–6470.
- Efron, B. (1993), *Introduction to the Bootstrap*, New York: Chapman & Hall.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Bostein, D. (1998), "Cluster Analysis and Display of Genome-wide Expression Patterns," *Proceedings of the National Academy of Sciences of the United States of America*, 95, 196–212.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-wide Expression Patterns," *Proceedings of the National Academy of Sciences of the United States of America*, 95, 14863–14868.
- Fang, K., and Zhang, Y. (1990), *Generalized Multivariate Analysis*, New York: Springer.
- Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. (2002), "Co-clustering of Biological Networks and Gene Expression Data," *Bioinformatics (Oxford, England)*, 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 Research*, 31, e15.
- 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," *Current Biology*, 10, 896–906.
- Ledoit, O., and Wolf, M. (2004), "A Well Conditioned Estimator for Large-Dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 88, 365–411.
- 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, 799–804.
- Li, C., and Wong, W. H. (2001), "Model-based Analysis of Oligonucleotide Arrays: Model Validation, Design Issues and Standard Error Application," *Genome Biology*, 2(8): research0032.1-0032.11.
- Li, K.-C. (2002), "Genome-wide Coexpression Dynamics: Theory and Application," *Proceedings of the National Academy of Sciences of the United States of America*, 99, 16875–16880.
- Lund, R., Ahlfors, H., 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," *European Journal of Immunology*, 35, 3307–3319.
- Penrose, R. (1995), "A Generalized Inverse for Matrices," *Proceedings of the Cambridge Philosophical Society*, 51, 406–413.
- Ramoni, M., Sebastiani, P., and Cohen, P. R. (2002), "Bayesian Clustering by Dynamics," *Machine Learning*, 47, 91–121.
- Roy, A., and Khatree, R. (2005), "On Implementation 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," *Molecular and Cellular Biology*, 24, 8823–8833.
- Schäfer, J., and Strimmer, K. (2005), "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics," *Statistical Applications in Genetics and Molecular Biology*, 4, Article 32, Issue 1.
- 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 Research*, 31, 3033–3037.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., 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, 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 *Handbook of Statistics* (Vol. 1), ed. P. R. Krishnaiah, New York: North-Holland, pp. 41–87.
- Wu, H., Chen, Y., Liang, J., Shi, B., Wu, G., Zhang, Y., Wang, D., Li, R., Yi, X., Zhang, H., Sun, L., and 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, 238–243.