SUPPLEMENTAL MATERIALS

http://www.amstat.org/publications/jasa/supplemental_materials

S1. 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 A.1(a), the column correlation matrix is an identity matrix to simulate for row vectors with independent vector components; in Figure A.1(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 A.1(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 A.1 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.

S2. DERIVATION OF MLEs IN EQUATIONS (6)–(8)

Let the covariance matrices $\mathbf{\Sigma^G}$ and $\mathbf{\Sigma^E}$ be invertible. Given that $\text{vec}(\mathbf{X}^T)$ follows a multivariate normal distribution with mean $\text{vec}(E(\mathbf{X}^T)) = \text{vec}(\mathbf{1\mu}^T)$ and covariance matrix $\mathbf{\Sigma^G} \otimes \mathbf{\Sigma^E}$, where $\mathbf{1}$ is a column vector of ones, the Maximum Likelihood Estimators (MLEs) of $\mathbf{\Sigma^G}$, $\mathbf{\Sigma^E}$ and $\boldsymbol{\mu}$, conditional on remaining parameters, are given in equations (6)–(8) in Section 2 of the main article respectively.

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

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

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

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

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

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

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 $\mathbf{\Sigma^G}$, $\mathbf{\Sigma^E}$ and $\boldsymbol{\mu}$, by which, we then obtain the MLEs of $\mathbf{\Sigma^G}$, $\mathbf{\Sigma^E}$ and $\boldsymbol{\mu}$, conditional on remaining parameters, respectively as given in equations (6)–(8) in Section 2. Note that the $\mathbf{\Sigma^E}$ here is equivalent to the $\mathbf{R^E}$ in equation (7) as $\mathbf{\Sigma^E}$ is assumed to have unit variances in the main article. ■

S3. APPLICATION OF KNORM CORRELATION TO ADDITIONAL TWO HUMAN CANCER DATASETS

The first dataset[1] (Cromer et al., 2004) consists of gene expressions from normal patients and hypopharyngeal cancer patients at four different disease progression stages. The second dataset[2] (Wu et al., 2005) consists of gene expressions from normal cells and endometrioid carcinoma cells treated with oestrogen, tamoxifen or no treatment. Each experiment has two replicates. The experiments are described in Table A.1.

By the biological descriptions of the experiments in Table A.1, there should exist varied levels of dependencies between the experiments. For example in the first dataset, gene expressions are expected to be more similar between the normal and disease early stage (i.e. disease-related mutations may not happen yet in most genes) compared to between the normal and disease late stage. In this context, genes with variable expressions are potential genes triggering the cancer development and tumor differentiations. The results of applications of our method to the two datasets are summarized in Table A.2. Again, we see the advantage of our method over Pearson correlation by taking into account these experiment dependencies.

[1] Accessible by accession number GDS1070 in the Gene Expression Omnibus.
[2] Accessible by accession number GSE3013 in the Gene Expression Omnibus.

S4. SIMULATION STUDY DEMONSTRATING SUPERIORITY OF KNORM APPROACH FOR SPARSE AND NON-SPARSE GENE CORRELATION MATRICES EVEN WHEN NUMBER OF AVAILABLE REPLICATES IS SMALL

Given any experiment covariance matrix $\mathbf{\Sigma^E}$ and gene correlation matrix $\mathbf{\Sigma^G}$, we first generate a gene expression matrix from a matrix normal distribution with zero-means and a Kronecker structured covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma^G} \otimes \mathbf{\Sigma^E}$. For every column, the column replicates were constructed by adding independent zero-mean normal nuisance noises of a small standard deviation (0.01 in this simulation study) to each column entry. This simulation was performed following the practical data structure: the replicates of entire expression matrix are *not* available in practice. What we observe are actually the replicates of each column (i.e. experiment) in the matrix.

We simulated 10 datasets under different settings: we used two different $\mathbf{\Sigma^G}$ to represent scenarios in which a gene correlation matrix can be either sparse or not sparse (its dimension is 100×100); the number of replicates varies between 1, 2, 5, 10 and 50; the $\mathbf{\Sigma^E}$ is fixed (we used the experiment covariance matrix (8×8) estimated from the yeast dataset). When more than 1 replicates of experiments are available, we constructed 50 bootstrapped data matrices by sampling the replicates from each experiment (see Section 2.3 in the main text).

We evaluated the results by examining the Frobenius norm of the error matrix (difference between estimated and true correlation matrices). We applied the following three approaches to estimate the gene correlation matrix:

Method I. Knorm correlation computed using the *true* experiment covariance matrix (see equation (7)) in the revised manuscript) and averaged over the bootstrapped data matrices.
Method II. Knorm correlation computed using the iterative estimation procedure described in Section 2.3 in the revised manuscript. This method *estimates* both the gene and experiment correlation matrix, instead of using the true correlation matrix in the above approach.
Method III. Pearson coefficient computed and averaged over the bootstrapped data matrices.

The results are summaried in Table A.3. The simulation results overall suggest that:

(i) Using the true experiment correlation matrix, the mean Frobenius norm of the error matrix is (as expected) the smallest (the second and fifth column). The estimates from the Pearson approach are the worst (the fourth and seventh column). This is not surprising as Pearson approach does not adjust for experiment dependencies or data redundancies.
(ii) The iterative estimation procedure works quite nicely when the gene correlation matrix is sparse (the third column), while its advantages over Pearson correlation is moderate when the gene correlation matrix is not sparse (the sixth column). This is because higher dependencies in a non-sparse gene correlation matrix introduce more redundancies in the data and reduce the amount of information. These results (the sixth column), on the other hand, suggest that our estimation procedure could be further improved to yield estimates closer to that obtained in column five.

(iii) In general, when the number of replicates is ≥ 2, increasing the number of replicates does not dramatically improve the estimate accuracy. This is because having a larger number of replicates only increases the amount of information on the nuisance variations and not on the biological variations that we wish to estimate (i.e., in practice, the replicates of entire expression matrix are *not* available; what we observe are actually the replicates of each column/experiment in the matrix). We note that our model is identifiable even with only one replicate due to the Kronecker product structured covariance matrix and some other constraint conditions discussed in the main text.

S5. RESULTS USING GENES RANKED WITHOUT TAKING ABSOLUTE CORRELATION VALUES (YEAST AND HUMAN *TH* CELL DATASETS)

See Tables A.4 and A.5.

S6. ROBUSTNESS OF KNORM CORRELATION AGAINST NORMALITY

The Knorm correlation was derived based on two assumptions - the multivariate normality of data and the Kronecker product structured covariance matrix. When either assumption is violated ***seriously***, the method can be invalid.

In the case that the assumption of the Kronecker product structured covariance matrix holds and the normality assumption does not, the effectiveness of our method would be associated with the performance of Pearson coefficient as a correlation measure. In more specific details, let us consider the random matrix $\boldsymbol{Z}$ with two row vectors $X = (X_1, X_2,\ldots, X_n)$ and $Y = (Y_1, Y_2,\ldots, Y_n)$:

$$\boldsymbol{Z} = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X_1 & X_2 & \ldots & X_n \\ Y_1 & Y_2 & \ldots & Y_n \end{pmatrix}.$$

Without loss of generality, we assume that X_i's and Y_i's follow the same distribution with zero mean and unit variance, and that the covariance matrix of $\boldsymbol{Z}$ is in the form of

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \otimes \mathbf{R}_n,$$

where $\rho = corr(X_i, Y_i)$ and $\mathbf{R}_n$ is the $n \times n$ nonsingular column-wise correlation matrix.

We remark the following facts related to the performance of Knorm correlation:

F1. The covariance matrix of $\boldsymbol{Z}\mathbf{R}_n^{-1/2}$ is $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \otimes \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Let $\boldsymbol{Z}\mathbf{R}_n^{-1/2} = \begin{pmatrix} X \\ Y \end{pmatrix}\mathbf{R}_n^{-1/2} = \begin{pmatrix} X_1^* & X_2^* & \ldots & X_n^* \\ Y_1^* & Y_2^* & \ldots & Y_n^* \end{pmatrix}$. Then we have $\text{var}(X_i^*) = \text{var}(Y_i^*) = 1$, $corr(X_i^*, Y_i^*) = \rho$ and $\begin{pmatrix} X_i^* \\ Y_i^* \end{pmatrix}$ is independent of $\begin{pmatrix} X_j^* \\ Y_j^* \end{pmatrix}$ for all $i, j = 1,\ldots,n$ and $i \neq j$.

F2. The ***Knorm*** correlation (as an estimate of ρ) can be understood as the Pearson coefficient between $(x_1^*,\ldots, x_n^*)$ and $(y_1^*,\ldots, y_n^*)$, the observed row vectors of $\begin{pmatrix} X_1^* & X_2^* & \ldots & X_n^* \\ Y_1^* & Y_2^* & \ldots & Y_n^* \end{pmatrix}$.

F3. $X_1^*,\ldots, X_n^*$ are independent of each other but may not follow the same distribution when $X_1,\ldots, X_n$ are not normally distributed. The same applies to Y_i^*'s. The differences among X_i^*'s (Y_i^*'s) in distribution will be the key factor affecting the performance of the Pearson coefficient between $(x_1^*,\ldots, x_n^*)$ and $(y_1^*,\ldots, y_n^*)$ as an estimate of ρ.

Consequently, the performance of Knorm correlation would rely on the distributional differences among X_i^*'s (Y_i^*'s). When $X_1,...,X_n$ are normally distributed, X_i^*'s will be *i.i.d*. When $X_1,...,X_n$ are closer to normals, X_i^*'s would be closer to be *i.i.d* and then the Knorm correlation could be more efficient.

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,…,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 100*p*% 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 read 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.

S7. 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.6 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.

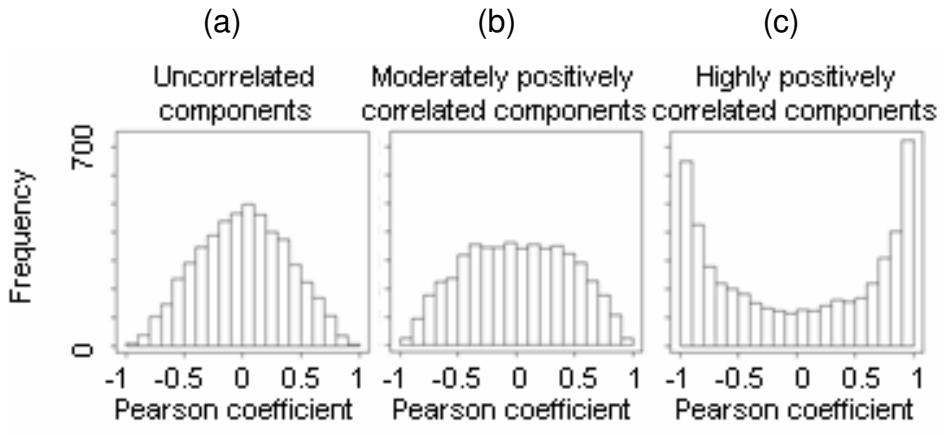

Figure A.2. A plot comparing the mean square errors of Knorm correlation from the normally distributed vectors (shown in blue) against that from the poisson distributed vectors (shown in red) across various component dependencies.

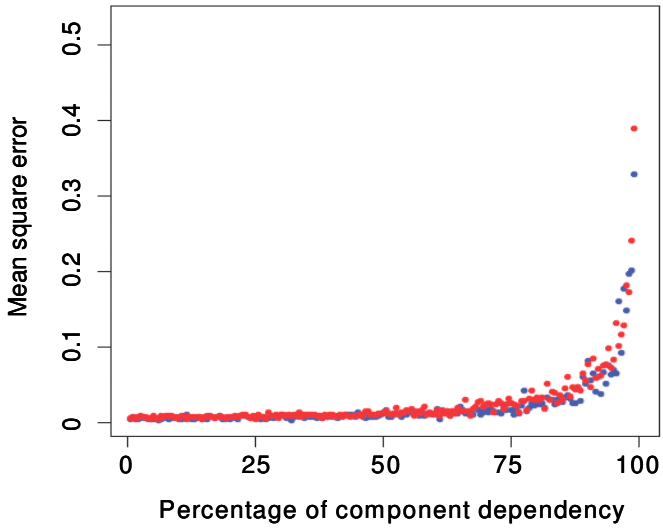

Table A.1. Descriptions of experiments used in the cancer datasets by Cromer et al. (2004) and Wu et al. (2005). Shown in parentheses () are the number of replicates available for each experiment.

H. Sapiens (human) dataset (Cromer et al., 2004)	Normal (no disease) (2) Disease early stage (2) Disease late stage with local recurrence (2) Disease late state with no metastasis (2) Disease late stage with metastasis (2)
H. Sapiens (human) dataset (Wu et al., 2005)	Stage I endometrioid carcinomas control (2) Stage I endometrioid carcinomas oestrogen (2) Stage I endometrioid carcinomas tamoxifen (2) Stage II endometrioid carcinomas control (2) Stage II endometrioid carcinomas oestrogen (2) Stage II endometrioid carcinomas tamoxifen (2) Normal endometrial epithelium control (2) Normal endometrial epithelium oestrogen (2) Normal endometrial epithelium tamoxifen (2)

Table A.2. Comparison of percentages of GO functionally related gene pairs among the top gene pairs identified by different methods based on absolute correlation values.

No. of top ranking gene pairs	Cromer et al. (2004) dataset		Wu et al. (2005) dataset	
	Knorm correlation	Pearson coefficient	Knorm correlation	Pearson coefficient
Top 10	10.0%	0.0%	10.0%	0.0%
Top 30	10.0%	0.0%	6.7%	6.7%
Top 50	6.0%	0.0%	6.0%	6.0%
Top 100	4.0%	3.0%	6.0%	6.0%
Top 500	3.4%	2.6%	4.2%	3.4%

Table A.3. Comparison of Different Methods by Mean Frobenius Norms						
	True gene correlation matrix is *sparse*			True gene corr. matrix is *not* sparse		
Number of replicates per each experiment	Method I (column 2)	Method II (column 3)	Method III (column 4)	Method I (column 5)	Method II (column 6)	Method III (column 7)
1	34.98	44.60*	44.10	32.75	41.90*	39.87
2	34.98	36.23	44.08	32.74	38.47	39.86
5	34.97	36.22	44.08	32.74	38.45	39.87
10	34.97	36.22	44.07	32.72	38.41	39.87
50	34.96	36.22	44.07	32.72	38.42	39.87

* indicates that these values are estimated using only one bootstrapped matrix in the proposed estimation procedure (since there is only one replicate available for each experiment).

Table A.4. Percentages of gene pairs found to be GO functionally related from among the top ranking gene pairs identified by Knorm correlation and the Pearson approach for the **yeast** dataset.

Yeast dataset						
No. of top ranking gene pairs	Gene pairs ranked by their …					
	absolute correlation values		*positive* correlation values		*negative* correlation values	
	Knorm correlation	Pearson coefficient	Knorm correlation	Pearson coefficient	Knorm correlation	Pearson coefficient
Top 10	30.0	10.0	30.0	20.0	0.0	10.0
Top 30	43.3	20.0	46.7	26.7	10.0	10.0
Top 50	38.0	26.0	40.0	30.0	8.0	6.0
Top 100	34.0	21.0	37.0	29.0	7.0	5.0
Top 500	26.4	21.8	29.2	29.4	5.4	4.4

Table A.5. Percentages of gene pairs found to be GO functionally related from among the top ranking gene pairs identified by Knorm correlation and the Pearson approach for the **human** *Th* cell dataset.

Human *Th* cell dataset						
No. of top ranking gene pairs	Gene pairs ranked by their …					
	absolute correlation values		*positive* correlation values		*negative* correlation values	
	Knorm correlation	Pearson coefficient	Knorm correlation	Pearson coefficient	Knorm correlation	Pearson coefficient
Top 10	10.0	10.0	10.0	10.0	10.0	0
Top 30	10.0	3.3	6.7	3.3	10.0	0
Top 50	10.0	4.0	8.0	4.0	8.0	0
Top 100	5.0	2.0	4.0	2.0	6.0	2.0
Top 500	4.0	3.4	3.6	3.6	3.6	1.4

Table A.6. 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%