Shrinkage-Based Similarity Metric for Cluster Analysis of Microarray Data *
(Full Technical Report)

TR2003-845

Vera Cherepinsky[1]†, Jiawu Feng[1], Marc Rejali[1], *and* Bud Mishra[1,2]

[1] Courant Institute, New York University, 251 Mercer Street, New York, NY 10012; and
[2] Cold Spring Harbor Lab, 1 Bungtown Road, Cold Spring Harbor, NY 11724.

ABSTRACT

The current standard correlation coefficient used in the analysis of microarray data, including gene expression arrays, was introduced in [1]. Its formulation is rather arbitrary. We give a mathematically rigorous derivation of the correlation coefficient of two gene expression vectors based on James-Stein Shrinkage estimators. We use the background assumptions described in [1], also taking into account the fact that the data can be treated as transformed into normal distributions. While [1] uses zero as an estimator for the expression vector mean μ, we start with the assumption that for each gene, μ is itself a zero-mean normal random variable (with *a priori* distribution $\mathcal{N}(0, \tau^2)$), and use Bayesian analysis to update that belief, to obtain *a posteriori* distribution of μ in terms of the data. The estimator for μ, obtained after shrinkage towards zero, differs from the mean of the data vectors and ultimately leads to a statistically robust estimator for correlation coefficients.

To evaluate the effectiveness of shrinkage, we conducted *in silico* experiments and also compared similarity metrics on a biological example using the data set from [1]. For the latter, we classified genes involved in the regulation of yeast cell-cycle functions by computing clusters based on various definitions of correlation coefficients, including the one using shrinkage, and contrasting them against clusters based on the activators known in the literature. In addition, we conducted an extensive computational analysis of the data from [1], empirically testing the performance of different values of the shrinkage factor γ and comparing them to the values of γ corresponding to the three metrics adressed here, namely, $\gamma = 0$ for the Eisen metric, $\gamma = 1$ for the Pearson correlation coefficient, and γ computed from the data for the Shrinkage metric.

The estimated "false-positives" and "false-negatives" from this study indicate the relative merits of clustering algorithms based on different statistical correlation coefficients as well as the sensitivity of the clustering algorithm to small perturbations in the correlation coefficients. These results indicate that using the shrinkage metric improves the accuracy of the analysis.

All derivation steps are described in detail; all mathematical assertions used in the derivation are proven in the appendix.

[1] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998), *PNAS USA* 95, 14863–14868.

1 Background

Traditionally, biology has proceeded as an observational science. Robert Hooke, whose work "Micrographia" of 1665 included the first identification of biological cells through his microscopical investigations, had said, "The truth is, the science of Nature has already been too long made only a work of

*This research was conducted under the Department of Energy Grant DoE-25-74100-F1799, the National Cancer Institute Grant NCI 5 RO1 CA79063-03, the NSF Career Grant IRI-9702071, the NSF Graduate Research Fellowship, and the NYU McCracken Fellowship.

†To whom correspondence should be addressed. E-mail: vera@cat.nyu.edu

the brain and the fancy. It is now high time that it should return to the plainness and soundness of observations on material and obvious things." Recently, we have seen an unprecedented progress in our observational and experimental abilities, allowing us to understand the structure of a largely unobservable transparent cell. The most prominent step in this direction has been through microarray-based gene expression analysis, providing us with the ability to quantify the transcriptional states of cells.

The most interesting insight can be obtained from transcriptome abundance data within a single cell under different experimental conditions. In the absence of technology to provide one with such a detailed picture, we have to make do with mRNA collected from a small population of cells, even when individual cells within the population may not be completely synchronized. Furthermore, these mRNAs will only give a partial picture, supported only by those genes that we are already familiar with and possibly missing many crucial undiscovered genes. Of course, without the proteomic data, transcriptomes tell less than half the story. Nonetheless, it goes without saying that microarrays have already revolutionized our understanding of biology even though they only provide occasional, noisy, unreliable, partial, and occluded snapshots of the transcriptional states of cells.

If one hypothesizes that the number of potential genes involved in cellular processes is relatively large compared to the regulatory elements and their effective combinations responsible for controlling these genes, then the transcriptional state-space should be rather low-dimensional compared to its apparent dimension. As a result, understanding this structure accurately from transcriptome data has many non-trivial implications to functional understanding of the cell. Partitioning genes into closely related groups has thus become the key mathematical first step in practically all statistical analyses of microarray data.

Traditionally, algorithms for cluster analysis of genome-wide expression data from DNA microarray hybridization are based upon statistical properties of gene expressions and result in organizing genes according to similarity in pattern of gene expression. These algorithms display the output graphically, often in a binary tree form, conveying the clustering and the underlying expression data simultaneously. If two genes belong to a cluster (or, equivalently, if they belong to a subtree of small depth) then one may infer a common regulatory mechanism for the two genes or interpret this information as an indication of the status of cellular processes. Furthermore, coexpression of genes of known function with novel genes may lead to a discovery process for characterizing unknown or poorly characterized genes. In general, since false-negatives (where two coexpressed genes are assigned to distinct clusters) may cause the discovery process to ignore useful information for certain novel genes, and false-positives (where two independent genes are assigned to the same cluster) may result in noise in the information provided to the subsequent algorithms used in analyzing regulatory patterns, it is important that the statistical algorithms for clustering be reasonably robust. Unfortunately, as the microarray experiments that can be carried out in an academic laboratory for a reasonable cost are small in number and suffer from experimental noise, often a statistician must resort to unconventional algorithms to deal with small-sample data.

A popular and one of the earliest clustering algorithms reported in the literature was introduced in [1]. In this paper, the gene-expression data were collected on spotted DNA microarrays [2] and were based upon gene expression in the budding yeast *Saccharomyces cerevisiae* during the diauxic shift [3], the mitotic cell division cycle [4], sporulation [5], and temperature and reducing shocks. In all experiments, RNA from experimental samples (taken at selected times during the process) was labeled during reverse transcription with the red-fluorescent dye Cy5 and was mixed with a reference sample labeled in parallel with the green-fluorescent dye Cy3. After hybridization and appropriate washing steps, separate images were acquired for each fluorophore, and fluorescence intensity ratios were obtained for all target elements. The experimental data were given in an $M \times N$ matrix structure, in which the M rows represented all genes for which data had been collected, the N columns represented individual array experiments (e.g., single time points or conditions), and each entry represented the measured Cy5/Cy3 fluorescence ratio at the corresponding target element on the appropriate array. All ratio values were log transformed to treat inductions and repressions of identical magnitude as numerically equal but opposite in sign. It was assumed that the raw ratio values followed log-normal distributions and hence, the log-transformed data followed normal distributions. While our mathematical derivations will rely on this assumption for the sake of simplicity, we note that our approach can be generalized in a straightforward manner to deal with other situations where this assumption is violated.

The gene similarity metric employed was a form of correlation coefficient. Let G_i be the (log-transformed) primary data for gene G in condition i. For any two genes X and Y observed over a series of N conditions, the classical similarity score based upon Pearson correlation coefficient is:

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right),$$

where

$$\Phi_G{}^2 = \sum_{i=1}^{N} \frac{(G_i - G_{offset})^2}{N}$$

and G_{offset} is the estimated mean of the observations, i.e.,

$$G_{offset} = \bar{G} = \frac{1}{N}\sum_{i=1}^{N} G_i.$$

Note that Φ_G is simply the (rescaled) estimated standard deviation of the observations. In the analysis presented in [1], "values of G_{offset} which are not the average over observations on G *were* used when there was an assumed unchanged or reference state represented by the value of G_{offset}, against which changes were to be analyzed; in all of the examples presented there, G_{offset} was set to 0, corresponding to a fluorescence ratio of 1.0." To distinguish this modified correlation coefficient from the classical Pearson correlation coefficient, we shall refer to it as Eisen correlation coefficient. Our main innovation is in suggesting a different value for G_{offset}, namely $G_{offset} = \gamma\bar{G}$, where γ is allowed to take a value between 0.0 and 1.0. Note that when $\gamma = 1.0$, we have the classical Pearson correlation coefficient and when $\gamma = 0.0$, we have replaced it by Eisen correlation coefficient. For a non-unit value of γ, the estimator for $G_{offset} = \gamma\bar{G}$ can be thought of as the unbiased estimator $\bar{G}$ being shrunk towards the believed value for $G_{offset} = 0.0$. We address the following questions: What is the optimal value for the shrinkage parameter γ from a Bayesian point of view? How do the gene expression data cluster as the correlation coefficient is modified with this optimal shrinkage parameter?

In order to achieve a consistent comparison, we leave the rest of the algorithms undisturbed. Namely, once the similarity measure has been assumed, we cluster the genes using the same hierarchical clustering algorithm as the one used by Eisen *et al.* Their hierarchical clustering algorithm is based on the centroid-linkage method (referred to as "average-linkage method" of Sokal and Michener [6] in [1]) and computes a binary tree (dendrogram) that assembles all the genes at the leaves of the tree, with each internal node representing possible clusters at different levels. For any set of M genes, an upper-triangular similarity matrix is computed by using a similarity metric of the type described above, which contains similarity scores for all pairs of genes. A node is created joining the most similar pair of genes, and a gene expression profile is computed for the node by averaging observations for the joined genes. The similarity matrix is updated with this new node replacing the two joined elements, and the process is repeated $(M-1)$ times until only a single element remains. The modified algorithm has been implemented by the authors within the "NYUMAD" microarray database system and can be freely downloaded from: http://bioinformatics.cat.nyu.edu/nyumad/clustering/. As each internal node can be labeled by a value representing the similarity between its two children nodes (i.e., the two elements that were combined to create the internal node), one can create a set of clusters by simply breaking the tree into subtrees by eliminating all the internal nodes with labels below a certain predetermined threshold value. The clusters created in this manner were used to compare the effects of choosing differing similarity measures.

2 Model

Recall that a family of correlation coefficients parametrized by $0 \le \gamma \le 1$ may be defined as follows:

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right), \quad (1)$$

where

$$\begin{aligned} \Phi_G &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(G_i - G_{offset})^2} \quad \text{and} \quad (2)\\ G_{offset} &= \gamma\bar{G} \quad \text{for} \quad G \in \{X, Y\} \end{aligned}$$

- *Pearson Correlation Coefficient* uses

$$G_{offset} = \bar{G} = \frac{1}{N}\sum_{j=1}^{N} G_i \quad \text{for every gene } G, \text{ or } \quad \gamma = 1.$$

- *Eisen et al.* (in [1]) use

$$G_{offset} = 0 \quad \text{for every gene } G, \text{ or } \quad \gamma = 0.$$

- We propose using the general form of equation (1) to derive a similarity metric which is dictated by the data and reduces the occurrence of false-positives (relative to the Eisen metric) and false-negatives (relative to the Pearson correlation coefficient).

2.1 Motivation and Setup

As mentioned above, the metric used by Eisen *et al.* in [1] had the form of equation (1) with G_{offset} set to 0 for every gene G (as a reference state against which to measure the data). Here, we rigorously examine the mathematical validity of setting G_{offset} to 0 arbitrarily. Even if it is initially

assumed that each gene G has zero mean, that assumption must be updated when data becomes available. To this end, we derive a correlation coefficient formula which is dictated by the data, and can be justified by a Bayesian argument.

The microarray data is given in the form of the levels of M genes expressed under N experimental conditions. The data can be viewed as

$$\{\{X_{ij}\}_{i=1}^{N}\}_{j=1}^{M}$$

where $M \gg N$ and $\{X_{ij}\}_{i=1}^{N}$ is the data vector for gene j.

2.2 Derivation

We begin by rewriting S in our notation:

$$\begin{aligned} & S(X_j, X_k) \qquad (3) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij}-(X_j)_{offset}}{\Phi_j}\right)\left(\frac{X_{ik}-(X_k)_{offset}}{\Phi_k}\right), \\ & {\Phi_j}^2 = \frac{1}{N}\sum_{i}\left(X_{ij}-(X_j)_{offset}\right)^2 \end{aligned}$$

In the most general setting, we can make the following assumptions on the data distribution: let all values X_{ij} for gene j have a Normal distribution with mean θ_j and standard deviation β_j (variance ${\beta_j}^2$); i.e.,

$$X_{ij} \sim \mathcal{N}(\theta_j, {\beta_j}^2) \quad \text{for} \quad i = 1, \ldots, N$$

with j fixed ($1 \le j \le M$), where θ_j is an unknown parameter (taking different values for different j). To estimate θ_j, it is convenient to assume that θ_j is itself a random variable taking values close to zero:

$$\theta_j \sim \mathcal{N}(0, \tau^2).$$

The assumed distribution aids us in obtaining the estimate of θ_j given in (14).

For convenience, let us also assume that the data are range-normalized, so that ${\beta_j}^2 = \beta^2$ for every j. If this assumption does not hold on the given data set, it is easily corrected by scaling each gene vector appropriately. Following common practice, we adjusted the range to scale to an interval of unit length, i.e., its maximum and minimum values differ by 1. Thus,

$$X_{ij} \sim \mathcal{N}(\theta_j, \beta^2) \quad \text{and} \quad \theta_j \sim \mathcal{N}(0, \tau^2).$$

Replacing $(X_j)_{offset}$ in (3) by the exact value of the mean θ_j yields a *Clairvoyant* correlation coefficient of X_j and X_k. In reality, since θ_j is itself a random variable, it must be estimated from the data. Therefore, to get an explicit formula for $S(X_j, X_k)$, we must derive estimators $\widehat{\theta_j}$ for all j.

In Pearson correlation coefficient, θ_j is estimated by the vector mean $\overline{X}_{\cdot j}$; Eisen correlation coefficient corresponds to replacing θ_j by 0 for every j, which is equivalent to assuming $\theta_j \sim \mathcal{N}(0,0)$ (i.e., $\tau^2 = 0$). We propose to find an estimate of θ_j (call it $\widehat{\theta_j}$) that takes into account both the prior assumption and the data.

2.3 Estimation of θ_j

First, let us obtain the posterior distribution of θ_j from the prior $\mathcal{N}(0, \tau^2)$ and the data. This derivation can be done either from the Bayesian considerations, or via the James-Stein Shrinkage estimators (see [7], or [8] for a recent review). Here, we discuss the former method.

2.3.1 $N = 1$

Assume initially that $N = 1$, i.e., we have one data point for each gene, and denote the variance by σ^2 for the moment:

$$\begin{aligned} X_j &\sim \mathcal{N}(\theta_j, \sigma^2) \qquad (4) \\ \theta_j &\sim \mathcal{N}(0, \tau^2) \qquad (5) \end{aligned}$$

For clarity, we denote the probability density function (pdf) of θ_j by $\pi(\cdot)$ and the pdf of X_j by $f(\cdot)$. It is immediate from (4) and (5) that

$$\begin{aligned} \pi(\theta_j) &= \frac{1}{\sqrt{2\pi}\tau}\,\exp\left(-{\theta_j}^2/2\tau^2\right), \\ f(X_j|\theta_j) &= \frac{1}{\sqrt{2\pi}\sigma}\,\exp\left(-(X_j-\theta_j)^2/2\sigma^2\right). \end{aligned}$$

By Bayes' Rule, the joint pdf of X_j and θ_j is given by

$$\begin{aligned} & f(X_j, \theta_j) = f(X_j|\theta_j)\ \pi(\theta_j) \qquad (6) \\ &= \frac{1}{2\pi\sigma\tau}\,\exp\left(-\left[\frac{{\theta_j}^2}{2\tau^2}+\frac{(X_j-\theta_j)^2}{2\sigma^2}\right]\right) \end{aligned}$$

Then $f(X_j)$, the marginal pdf of X_j alone is

$$\begin{aligned} f(X_j) &= \mathbf{E}_{\theta_j} f(X_j|\theta_j) = \int_{\theta=-\infty}^{\infty} f(X_j|\theta)\pi(\theta)d\theta \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\,\exp\left(-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}\right), \qquad (7) \end{aligned}$$

where the equality in equation (7) is written out in Appendix A.2. It follows that the posterior distribution of θ_j,

again by Bayes' Theorem, is given by

$$\begin{aligned}\pi(\theta_j|X_j) &= \frac{f(X_j,\theta_j)}{f(X_j)} \\ &= \frac{f(X_j|\theta_j)\ \pi(\theta_j)}{f(X_j)} \quad \text{by (6)} \\ &= \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}\ \exp\left[-\frac{\left(\theta_j-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2}{2\left(\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)}\right]. \qquad (8)\end{aligned}$$

(See Appendix A.3 for derivation of (8).)

Since this has Normal form, we can read off the mean and variance

$$\begin{aligned}\mathbf{E}(\theta_j|X_j) &= \frac{\tau^2}{\sigma^2+\tau^2}X_j \\ &= \left(1-\frac{\sigma^2}{\sigma^2+\tau^2}\right)X_j, \qquad (9) \\ Var(\theta_j|X_j) &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.\end{aligned}$$

We can estimate θ_j by its mean.

2.3.2 N ARBITRARY

Now, if $N>1$ is arbitrary, X_j becomes a vector $X_{\cdot j}$. It can be easily shown by using likelihood functions that the vector of values $\{X_{ij}\}_{i=1}^N$, with $X_{ij}\sim\mathcal{N}(\theta_j,\beta^2)$, can be treated as a single data point $Y_j=\overline{X}_{\cdot j}=\sum_{i=1}^N X_{ij}/N$ from the distribution $\mathcal{N}(\theta_j,\beta^2/N)$ (see Appendix A.4).

Thus, following the above derivation with $\sigma^2=\beta^2/N$, we have a Bayesian estimator for θ_j given by $\mathbf{E}(\theta_j|X_{\cdot j})$:

$$\widehat{\theta_j}=\left(1-\frac{\beta^2/N}{\beta^2/N+\tau^2}\right)Y_j. \qquad (10)$$

Unfortunately, (10) cannot be used in (3) directly, because τ^2 and β^2 are unknown, so must be estimated from the data.

2.3.3 ESTIMATION OF $1/\left(\beta^2/N+\tau^2\right)$

Let

$$W=\frac{M-2}{\sum_{j=1}^M {Y_j}^2}. \qquad (11)$$

The form of W comes from James-Stein estimation ([7]), but its derivation will not be discussed here; instead we treat it as an educated guess and verify that it is indeed an appropriate estimator for $1/\left(\beta^2/N+\tau^2\right)$.

$$\begin{aligned}Y_j &\sim \theta_j+\frac{\beta^2}{N}\mathcal{N}(0,1) \\ &\sim \tau^2\mathcal{N}(0,1)+\frac{\beta^2}{N}\mathcal{N}(0,1) \\ &\sim \left(\frac{\beta^2}{N}+\tau^2\right)\mathcal{N}(0,1)\sim\mathcal{N}\left(0,\frac{\beta^2}{N}+\tau^2\right) \qquad (12)\end{aligned}$$

The transition in (12) is justified in Appendix A.5. Let $\alpha^2=\beta^2/N+\tau^2$. Then from (12) it follows that

$$\frac{Y_j}{\sqrt{\alpha^2}}=\frac{Y_j}{\alpha}\sim\mathcal{N}(0,1),$$

and hence

$$\sum_{j=1}^M {Y_j}^2=\alpha^2\sum_{j=1}^M\left(\frac{Y_j}{\alpha}\right)^2=\alpha^2\chi^2_M,$$

where χ^2_M is a Chi-square random variable with M degrees of freedom. By properties of the Chi-square distribution and the linearity of expectation,

$$\begin{aligned}\mathbf{E}\left(\frac{\alpha^2}{\sum {Y_j}^2}\right) &= \frac{1}{M-2} \quad \text{(see Appendix A.6)} \\ \mathbf{E}(W) &= \mathbf{E}\left(\frac{M-2}{\sum {Y_j}^2}\right)=\frac{1}{\alpha^2}=\frac{1}{\frac{\beta^2}{N}+\tau^2}\end{aligned}$$

Thus, W is an unbiased estimator of $1/\left(\beta^2/N+\tau^2\right)$, and can be used to replace $1/\left(\beta^2/N+\tau^2\right)$ in (10).

2.3.4 ESTIMATION OF β^2

It can be shown (see Appendix A.7) that

$${S_j}^2=\frac{1}{N-1}\sum_{i=1}^N(X_{ij}-Y_j)^2$$

is an unbiased estimator for β^2 based solely on data from gene j, and that $\frac{N-1}{\beta^2}{S_j}^2$ has Chi-square distribution with $(N-1)$ degrees of freedom. Since this holds for every j, we can get a more accurate estimate for β^2 by pooling all available data, i.e., by averaging the estimates for each j:

$$\begin{aligned}\widehat{\beta^2} &= \frac{1}{M}\sum_{j=1}^M {S_j}^2=\frac{1}{M}\sum_{j=1}^M\left(\frac{1}{N-1}\sum_{i=1}^N(X_{ij}-Y_j)^2\right) \\ &= \frac{1}{M(N-1)}\sum_{j=1}^M\sum_{i=1}^N(X_{ij}-Y_j)^2. \qquad (13)\end{aligned}$$

$\widehat{\beta^2}$ is an unbiased estimator for β^2, since

$$\begin{aligned}\mathbf{E}(\widehat{\beta^2}) &= \mathbf{E}\left(\frac{1}{M}\sum_{j=1}^{M} {S_j}^2\right) \\ &= \frac{1}{M}\sum_{j=1}^{M}\mathbf{E}({S_j}^2) = \frac{1}{M}\sum_{j=1}^{M}\beta^2 = \beta^2.\end{aligned}$$

Substituting the estimates (11) and (13) into (10), we obtain the explicit estimate for θ_j:

$$\begin{aligned}&\widehat{\theta_j} \\ &= \left(1 - \frac{\widehat{1}}{\frac{\beta^2}{N}+\tau^2}\frac{\widehat{\beta^2}}{N}\right)Y_j \\ &= \left(1 - W\cdot\frac{\widehat{\beta^2}}{N}\right)Y_j \\ &= \left(1 - \left(\frac{M-2}{\sum_{k=1}^{M}{Y_k}^2}\right)\cdot\frac{1}{N}\cdot\frac{1}{M(N-1)}\sum_{k=1}^{M}\sum_{i=1}^{N}(X_{ik}-Y_k)^2\right)Y_j \\ &= \underbrace{\left(1 - \frac{M-2}{MN(N-1)}\cdot\frac{\sum_{k=1}^{M}\sum_{i=1}^{N}(X_{ik}-Y_k)^2}{\sum_{k=1}^{M}{Y_k}^2}\right)}_{\gamma}Y_j \qquad (14) \\ &= \gamma\overline{X}_{\cdot j}\end{aligned}$$

Finally, we can substitute $\widehat{\theta_j}$ from equation (14) into the correlation coefficient in (3) wherever $(X_j)_{offset}$ appears to obtain an explicit formula for $S(X_{\cdot j}, X_{\cdot k})$.

3 Algorithm & Implementation

The implementation of hierarchical clustering proceeds in a greedy manner, always choosing the most similar pair of elements (starting with genes at the bottom-most level) and combining them to create a new element. The "expression vector" for the new element is simply the weighted average of the expression vectors of the two most similar elements that were combined. This structure of repeated pair-wise combinations is conveniently represented in a binary tree, whose leaves are the set of genes and internal nodes are the elements constructed from the two children nodes. The algorithm is described below in pseudocode.

3.1 Hierarchical clustering pseudocode

Given $\{\{X_{ij}\}_{i=1}^{N}\}_{j=1}^{M}$:

`Switch:`

`Pearson:` $\gamma = 1$;

`Eisen:` $\gamma = 0$;

`Shrinkage:` {

`Compute` $W = (M-2)\Big/\sum_{j=1}^{M}{\overline{X}_{\cdot j}}^2$

`Compute` $\widehat{\beta^2} = \sum_{j=1}^{M}\sum_{i=1}^{N}\left(X_{ij}-\overline{X}_{\cdot j}\right)^2\Big/\left(M(N-1)\right)$

$\gamma = 1 - W\cdot\widehat{\beta^2}/N$

}

`While (# clusters > 1) do`

- `Compute similarity table:`
  $S(G_j, G_k) = \frac{\sum_i\left(G_{ij}-(G_j)_{offset}\right)\left(G_{ik}-(G_k)_{offset}\right)}{\sqrt{\sum_i\left(G_{ij}-(G_j)_{offset}\right)^2\cdot\sum_i\left(G_{ik}-(G_k)_{offset}\right)^2}}$,
  `where` $(G_\ell)_{offset} = \gamma\overline{G_\ell}$.
- `Find` (j^*, k^*) :
  $S(G_{j^*}, G_{k^*}) \geq S(G_j, G_k)\ \forall$ `clusters` j, k
- `Create new cluster` $N_{j^*k^*}$
  $=$ `weighted average of` G_{j^*} `and` G_{k^*}.
- `Take out clusters` j^* `and` k^*.

The implementation of generalized hierarchical clustering with options to choose different similarity measures has been incorporated into NYUMAD (NYU MicroArray Database), an integrated system to maintain and analyze biological abundance data along with associated experimental conditions and protocols. While the initial goal was to provide a system to manage microarray data, the system has been designed to store any type of abundance data, including protein levels. This system uses a relational database management system for the storage of data and has a flexible database schema that stores abundance data along with general research data such as experimental conditions and protocols. The database schema is defined using standard SQL (Structured Query Language) and is therefore portable to any SQL database platform. To enable widespread utility, NYUMAD supports the MAGE-ML standard ([9]) for the exchange of gene expression data, defined by the Microarray Gene Expression Data Group (MGED) — web site at http://www.mged.org/.

There are several ways to access the system: using the NYUMAD Java application, through web pages, or through custom applications (for details, see http://bioinformatics.cat.nyu.edu/nyumad/). Data transfer is affected using the world wide web (WWW) with the HTTP protocol. The use of the WWW for communication ensures accessibility from any location.

The graphical user interface (GUI) provided by the Java application facilitates easy data submission, retrieval, and

analysis. The Java application presents data in a logical manner and allows easy navigation through the data. The GUI also allows straightforward updating of existing data and insertion of new data.

NYUMAD supports collaborative research efforts by allowing groups to submit data from any location (via HTTP) and to view, retrieve, or analyze each other's data immediately. Groups can share protocols and divide a large project covering a wide range of experimental conditions into subprojects performed by individual groups.

NYUMAD is a secure repository for both public and private data. Users can control the visibility of their data so that initially the data might be private but after the publication of the results, the data can be marked public and made visible to the larger research community. Public users can log in with a general login ID without the need for a password and view and retrieve any of the public data.

The system provides a wide range of data analysis and interpretation tools and algorithms that help in identifying patterns and relationships. A general feature of NYUMAD is the flexibility for users to build their own queries and utilize their own parameters, data transformations, and filters where appropriate. Users can retrieve queried data for input to their own tools or use other tools within NYUMAD — for example, perform a clustering of their microarray data or determine the statistical significance of differential expression values for a specific set of genes. Data analysis tools are supplemented with visualization tools.

4 Results

4.1 Mathematical Simulation

To compare the performance of these algorithms, we started with a relatively simple *in silico* experiment. In such an experiment, one can create two genes X and Y and simulate N (about 100) experiments as follows:

$$\begin{aligned} X_i &= \theta_X + \sigma_X(\alpha_i(X,Y) + \mathcal{N}(0,1)), \text{ and} \\ Y_i &= \theta_Y + \sigma_Y(\alpha_i(X,Y) + \mathcal{N}(0,1)), \end{aligned}$$

where α_i, chosen from a uniform distribution over a range $[L,H]$ ($\mathcal{U}(L,H)$), is a "bias term" introducing a correlation (or none if all α's are zero) between X and Y. $\theta_X \sim \mathcal{N}(0,\tau^2)$ and $\theta_Y \sim \mathcal{N}(0,\tau^2)$ are the means of X and Y, respectively. Similarly, σ_X and σ_Y are the standard deviations for X and Y, respectively.

Note that, with this model

$$\begin{aligned} S(X,Y) &= \frac{1}{N}\sum_{i=1}^{N}\frac{(X_i-\theta_X)}{\sigma_X}\frac{(Y_i-\theta_Y)}{\sigma_Y} \\ &\sim \frac{1}{N}\sum_{i=1}^{N}(\alpha_i+\mathcal{N}(0,1))(\alpha_i+\mathcal{N}(0,1)) \\ &\sim \frac{1}{N}\left[\left(\sum_{i=1}^{N}\alpha_i^2\right)+\chi^2_N+2\mathcal{N}(0,1)\sum_{i=1}^{N}\alpha_i\right] \end{aligned}$$

if the exact values of the mean and variance are used. We denote the distribution of S by $\mathcal{F}(\mu,\delta)$, where μ is the mean and δ is the standard deviation.

The model was implemented in Mathematica [10]; the following parameters were used in the simulation: $N = 100$, $\tau \in \{0.1, 10.0\}$ (representing very low or high variability among the genes), $\sigma_X = \sigma_Y = 10.0$, and $\alpha = 0$ representing no correlation between the genes or $\alpha \sim \mathcal{U}(0,1)$ representing some correlation between the genes. Once the parameters were fixed for a particular *in silico* experiment, the gene-expression vectors for X and Y were generated many thousand times, and for each pair of vectors $S_c(X,Y)$, $S_p(X,Y)$, $S_e(X,Y)$, and $S_s(X,Y)$ were estimated by four different algorithms and further examined to see how the estimators of S varied over these trials. These four different algorithms estimated S according to equations (1), (2) as follows: *Clairvoyant* estimated S_c using the true values of θ_X, θ_Y, σ_X, and σ_Y; *Pearson* estimated S_p using the unbiased estimators $\bar{X}$ and $\bar{Y}$ of θ_X and θ_Y (for X_{offset} and Y_{offset}), respectively; *Eisen* estimated S_e using the value 0.0 as the estimator of both θ_X and θ_Y; and *Shrinkage* estimated S_s using the shrunk biased estimators $\widehat{\theta}_X$ and $\widehat{\theta}_Y$ of θ_X and θ_Y, respectively. In the latter three, the standard deviation was estimated as in (2). The histograms corresponding to these *in silico* experiments can be found in Figure 1. Our observations can be summarized as follows:

- When X and Y are not correlated and the noise in the input is low ($N = 100$, $\tau = 0.1$, and $\alpha = 0$), Pearson does just as well as Eisen, Shrinkage, or Clairvoyant ($S_c \sim \mathcal{F}(-0.000297, 0.0996)$, $S_p \sim \mathcal{F}(-0.000269, 0.0999)$, $S_e \sim \mathcal{F}(-0.000254, 0.0994)$, and $S_s \sim \mathcal{F}(-0.000254, 0.0994)$).

- When X and Y are not correlated but the noise in the input is high ($N = 100$, $\tau = 10.0$, and $\alpha = 0$), Pearson does just as well as Shrinkage or Clairvoyant, but Eisen introduces far too many false-positives ($S_c \sim \mathcal{F}(-0.000971, 0.0994)$, $S_p \sim \mathcal{F}(-0.000939, 0.100)$, $S_e \sim \mathcal{F}(-0.00119, 0.354)$, and $S_s \sim \mathcal{F}(-0.000939, 0.100)$).

- When X and Y are correlated and the noise in the input is low ($N = 100$, $\tau = 0.1$, and $\alpha \sim \mathcal{U}(0,1)$), Pearson does much more poorly compared to Eisen, Shrinkage, or Clairvoyant — these three doing equally well; Pearson introduces too many false-negatives ($S_c \sim \mathcal{F}(0.331, 0.132)$, $S_p \sim \mathcal{F}(0.0755, 0.0992)$, $S_e \sim \mathcal{F}(0.248, 0.0915)$, and $S_s \sim \mathcal{F}(0.245, 0.0915)$).

- Finally, when X and Y are correlated and the noise in the input is high, the signal-to-noise ratio becomes extremely poor and all the algorithms fail, i.e., introduce errors ($S_c \sim \mathcal{F}(0.333, 0.133)$, $S_p \sim \mathcal{F}(0.0762, 0.100)$, $S_e \sim \mathcal{F}(0.117, 0.368)$, and $S_s \sim \mathcal{F}(0.0762, 0.0999)$).

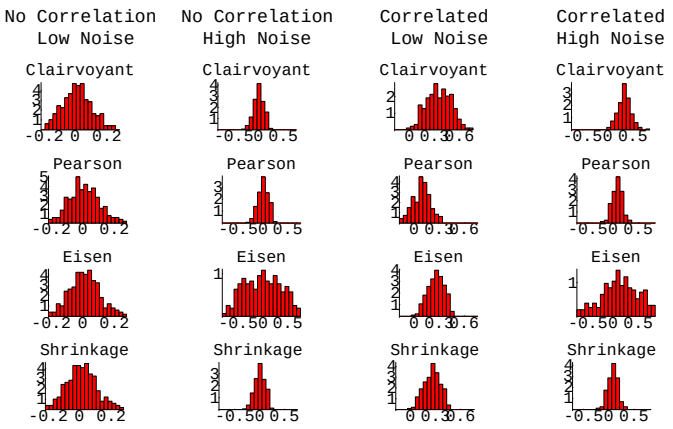

Figure 1: Histograms.

In summary, one can conclude that for the same clustering algorithm, Pearson tends to introduce more false-negatives and Eisen tends to introduce more false-positives than Shrinkage. Shrinkage, on the other hand, reduces these errors by combining the good properties of both algorithms.

4.2 Biological Example

We then proceeded to test the algorithms on a biological example. We chose a biologically well-characterized system, and analyzed the clusters of genes involved in the yeast cell cycle. These clusters were computed using the hierarchical clustering algorithm with the underlying similarity measure chosen from the following three: Pearson, Eisen, or Shrinkage. As a reference, the computed clusters were compared to the ones implied by the common cell-cycle functions and regulatory systems inferred from the roles of various transcriptional activators (see Figure 2).

Note that our experimental analysis is based on the assumption that the groupings suggested by the ChIP (Chromatin ImmunoPrecipitation) analysis are, in fact, correct and thus, provide a direct approach to compare various correlation coefficients. It is quite likely that the ChIP-based groupings themselves contain many false relations (both positive and negative) and corrupt our inference in some unknown manner. Nonetheless, we observe that the trends of reduced false positives and negatives in shrinkage analysis with these biological data are consistent with the analysis based on mathematical simulation and hence, reassuring.

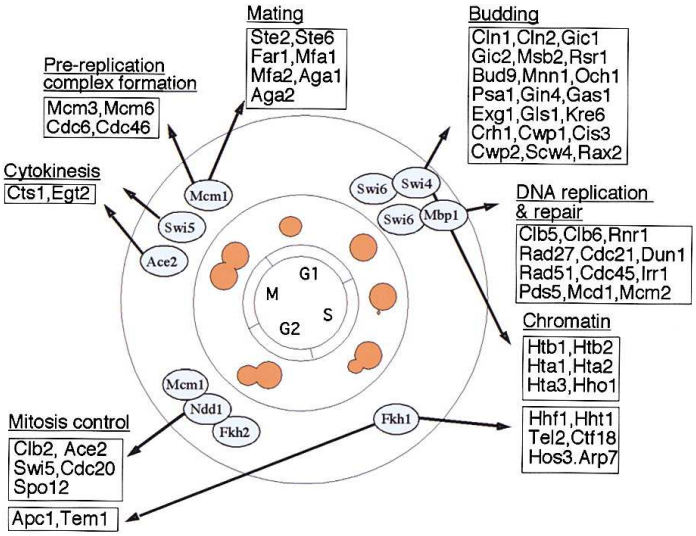

Figure 2: Regulation of cell-cycle functions by the activators. [Reproduced with permission from [11] (Copyright 2001, Elsevier)].

In the work of Simon *et al.* ([11]), genome-wide location analysis was used to determine how the yeast cell cycle gene expression program is regulated by each of the nine known cell cycle transcriptional activators: Ace2, Fkh1, Fkh2, Mbp1, Mcm1, Ndd1, Swi4, Swi5, and Swi6. It was also found that cell cycle transcriptional activators which function during one stage of the cell cycle regulate transcriptional activators that function during the next stage. This serial regulation of transcriptional activators together with various functional properties suggests a simple way of partitioning some selected cell cycle genes into nine clusters, each one characterized by a group of transcriptional activators working together and their functions (see Table 1): for instance, Group 1 is characterized by the activators Swi4 and Swi6 and the function of budding; Group 2 is characterized by the activators Swi6 and Mbp1 and the function involving DNA replication and repair at the juncture of G1

and S phases, etc.

Table 1: Genes in our data set, grouped by transcriptional activators and cell-cycle functions.

	Activators	Genes	Functions
1	Swi4, Swi6	Cln1, Cln2, Gic1, Gic2, Msb2, Rsr1, Bud9, Mnn1, Och1, Exg1, Kre6, Cwp1	Budding
2	Swi6, Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45, Mcm2	DNA replication and repair
3	Swi4, Swi6	Htb1, Htb2, Hta1, Hta2, Hta3, Hho1	Chromatin
4	Fkh1	Hhf1, Hht1, Tel2, Arp7	Chromatin
5	Fkh1	Tem1	Mitosis Control
6	Ndd1, Fkh2, Mcm1	Clb2, Ace2, Swi5, Cdc20	Mitosis Control
7	Ace2, Swi5	Cts1, Egt2	Cytokinesis
8	Mcm1	Mcm3, Mcm6, Cdc6, Cdc46	Pre-replication complex formation
9	Mcm1	Ste2, Far1	Mating

Our initial hypothesis can be summarized as follows: *Genes expressed during the same cell cycle stage, and regulated by the same transcriptional activators should be in the same cluster.* Below we list some of the deviations from the hypothesis observed in the raw data.

Possible False-Positives:

- Bud9 (Group 1: Budding) and {Cts1, Egt2} (Group 7: Cytokinesis) are placed in the same cluster by all three metrics: P49 = S82 $\simeq$ E47; however, the Eisen metric also places Exg1 (Group 1) and Cdc6 (Group 8: Pre-replication complex formation) in the same cluster.
- Mcm2 (Group 2: DNA replication and repair) and Mcm3 (Group 8) are placed in the same cluster by all three metrics: P10 = S20 $\simeq$ E73; however, the Eisen metric places several more genes from different groups in the same cluster: {Rnr1, Rad27, Cdc21, Dun1, Cdc45} (Group 2), Hta3 (Group 3: Chromatin), and Mcm6 (Group 8) are also placed in cluster E73.

Possible False-Negatives:

- Group 1: Budding (Table 1) is split into four clusters by the Eisen metric:
  {Cln1, Cln2, Gic2, Rsr1, Mnn1} $\in$ Cluster *a* (E39), Gic2 $\in$ Cluster *b* (E62), {Bud9, Exg1} $\in$ Cluster *c* (E47), and {Kre6, Cwp1} $\in$ Cluster *d* (E66);
  and into six clusters by both the Shrinkage and Pearson metrics:
  {Cln1, Cln2, Gic2, Rsr1, Mnn1} $\in$ Cluster *a* (S3=P66), {Gic1, Kre6} $\in$ Cluster *b* (S39=P17), Msb2 $\in$ Cluster *c* (S24=P71), Bud9 $\in$ Cluster *d* (S82=P49), Exg1 $\in$ Cluster *e* (S48=P78), and Cwp1 $\in$ Cluster *f* (S8=P4).

Table 1 contains those genes from Figure 2 that were present in our data set. The following tables contain these genes grouped into clusters by a hierarchical clustering algorithm using the three metrics (Eisen in Table 2, Pearson in Table 3, and Shrinkage in Table 4) thresholded at a correlation coefficient value of 0.60. The choice of the threshold parameter is discussed further in section 5. Genes that have not been grouped with any others at a similarity of 0.60 or higher are absent from the tables; in the subsequent analysis they are treated as *singleton* clusters.

Table 2: Eisen Clusters

E39	Swi4/Swi6	Cln1, Cln2, Gic2, Rsr1, Mnn1
E62	Swi4/Swi6	Gic1
E47	Swi4/Swi6 Ace2/Swi5 Mcm1	Bud9, Exg1 Cts1, Egt2 Cdc6
E66	Swi4/Swi6	Kre6, Cwp1
E71	Swi6/Mbp1 Fkh1 Ndd1/Fkh2/Mcm1 Mcm1	Clb5, Clb6, Rad51 Tel2 Cdc20 Cdc46
E73	Swi6/Mbp1 Swi4/Swi6 Mcm1	Rnr1, Rad27, Cdc21, Dun1, Cdc45, Mcm2 Hta3 Mcm3, Mcm6
E63	Swi4/Swi6 Fkh1	Htb1, Htb2, Hta1, Hta2, Hho1 Hhf1, Hht1
E32	Fkh1	Arp7
E38	Fkh1 Ndd1/Fkh2/Mcm1	Tem1 Clb2, Ace2, Swi5
E51	Mcm1	Ste2, Far1

The value $\gamma \simeq 0.89$ estimated from the raw yeast data was surprisingly high, contrary to the suggestion in [1] that the value $\gamma = 0$ performed better than $\gamma = 1$. It also did not yield as great an improvement in the yeast data clusters as the simulations indicated. This suggested that the true value of γ is closer to 0. Upon closer examination of the data, we observed that the data in its raw "pre-normalized" form is inconsistent with the assumptions used in deriving γ:

Table 3: Pearson Clusters

P66	Swi4/Swi6	Cln1, Cln2, Gic2, Rsr1, Mnn1
P17	Swi4/Swi6	Gic1, Kre6
P71	Swi4/Swi6	Msb2
P49	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
P78	Swi4/Swi6	Exg1
P4	Swi4/Swi6	Cwp1
P12	Swi6/Mbp1	Clb5, Clb6, Rnr1, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
P10	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
P54	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
P37	Fkh1	Arp7
P16	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
P50	Mcm1	Ste2, Far1

Table 4: Shrinkage Clusters

S3	Swi4/Swi6	Cln1, Cln2, Gic2, Rsr1, Mnn1
S39	Swi4/Swi6	Gic1, Kre6
S24	Swi4/Swi6	Msb2
S82	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
S48	Swi4/Swi6	Exg1
S8	Swi4/Swi6	Cwp1
S14	Swi6/Mbp1	Clb5, Clb6, Rnr1, Cdc21, Dun1, Rad51, Cdc45
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
S20	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
S4	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
S13	Swi4/Swi6	Hta3
S63	Fkh1	Arp7
S22	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
S83	Mcm1	Ste2, Far1

1. The gene vectors are not range-normalized, so ${\beta_j}^2 \neq \beta^2$ for every j, and
2. The N experiments are not necessarily independent.

4.3 Corrections

We attempted to remedy the first shortcoming by normalizing all gene vectors with respect to range (dividing each entry in gene X by $(X_{\max} - X_{\min})$), recomputing the estimated γ value, and repeating the clustering process. As normalized gene expression data yielded the estimate $\gamma \simeq 0.91$, still too high a value, we conducted an extensive computational experiment to determine the best empirical γ value by also clustering with the shrinkage factors of 0.2, 0.4, 0.6, and 0.8. The clusters taken at the correlation factor cut-off of 0.60, as above, are presented in Tables 5–11.

Table 5: RN Data, $\gamma = 0.0$ (Eisen Clusters)

E8	Swi4/Swi6	Cln1, Msb2, Mnn1
E71	Swi4/Swi6	Cln2, Rsr1
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
E14	Swi4/Swi6	Gic1
E17	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
	Mcm1	Ste2, Far1
E16	Swi4/Swi6	Exg1
E59	Swi4/Swi6	Kre6
E18	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
E86	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
E10	Fkh1	Arp7
E19	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
E11	Mcm1	Cdc6

To compare the resulting sets of clusters, we introduced the following notation. Write each cluster set as follows:

$$\left\{x \to \{\{y_1, z_1\}, \{y_2, z_2\}, \ldots, \{y_{n_x}, z_{n_x}\}\}\right\}_{x=1}^{\#\text{ of groups}}$$

where x denotes the group number (as described in Table 1), n_x is the number of clusters group x appears in, and for each cluster $j \in \{1, \ldots, n_x\}$ there are y_j genes from group

Table 6: Range-normalized data, $\gamma = 0.2$

$S_{0.2}59$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.2}26$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
$S_{0.2}23$	Swi4/Swi6	Gic1
$S_{0.2}58$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.2}57$	Swi4/Swi6	Exg1
	Fkh1	Arp7
$S_{0.2}61$	Swi4/Swi6	Kre6
$S_{0.2}18$	Swi6/Mbp1	Clb5
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.2}28$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.2}25$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.2}29$	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.2}4$	Mcm1	Ste2
$S_{0.2}55$	Mcm1	Far1

Table 7: Range-normalized data, $\gamma = 0.4$

$S_{0.4}64$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.4}13$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.4}44$	Swi4/Swi6	Gic1, Kre6
$S_{0.4}27$	Swi4/Swi6	Msb2
$S_{0.4}46$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.4}73$	Swi4/Swi6	Exg1
$S_{0.4}2$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.4}48$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.4}26$	Fkh1	Arp7
$S_{0.4}25$	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.4}16$	Mcm1	Cdc6
$S_{0.4}47$	Mcm1	Ste2
$S_{0.4}58$	Mcm1	Far1

x and z_j genes from other groups in Table 1. A value of "$*$" for z_j denotes that cluster j contains additional genes, although none of them are cell cycle genes; in subsequent computations, this value is treated as 0.

This notation naturally lends itself to a scoring function for measuring the number of false-positives, number of false-negatives, and total error score, which aids in the comparison of cluster sets.

$$\mathrm{FP}(\gamma) = \frac{1}{2}\sum_{x}\sum_{j=1}^{n_x} y_j \cdot z_j \tag{15}$$

$$\mathrm{FN}(\gamma) = \sum_{x}\sum_{1\le j<k\le n_x} y_j \cdot y_k \tag{16}$$

$$\mathrm{Error\_score}(\gamma) = \mathrm{FP}(\gamma) + \mathrm{FN}(\gamma) \tag{17}$$

In this notation, the cluster sets with their error scores can be listed as follows:

$$\begin{aligned}
&\gamma = 0.0 (E) \Longrightarrow \\
\{1 &\rightarrow \{\{3,*\},\{2,13\},\{1,*\},\{1,*\},\\
&\quad\ \{1,*\},\{1,4\},\{1,0\},\{1,0\},\{1,0\}\},\\
2 &\rightarrow \{\{8,7\},\{1,1\}\},\\
3 &\rightarrow \{\{5,2\},\{1,14\}\},\\
4 &\rightarrow \{\{2,5\},\{1,14\},\{1,*\}\},\\
5 &\rightarrow \{\{1,3\}\},\\
6 &\rightarrow \{\{3,1\},\{1,14\}\},\\
7 &\rightarrow \{\{2,3\}\},\\
8 &\rightarrow \{\{2,13\},\{1,1\},\{1,0\}\},\\
9 &\rightarrow \{\{2,3\}\}\\
\}&\\
&\mathrm{Error_score}(0.0) = 97 + 88 = 185
\end{aligned}$$

Table 8: Range-normalized data, $\gamma = 0.6$

$S_{0.6}34$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.6}77$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.6}35$	Swi4/Swi6	Gic1, Kre6
$S_{0.6}47$	Swi4/Swi6	Msb2
$S_{0.6}62$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.6}20$	Swi4/Swi6	Exg1
$S_{0.6}73$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.6}91$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.6}48$	Fkh1	Arp7
$S_{0.6}37$	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.6}64$	Mcm1	Ste2
$S_{0.6}63$	Mcm1	Far1

Table 9: Range-normalized data, $\gamma = 0.8$

$S_{0.8}51$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.8}7$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.8}64$	Swi4/Swi6	Gic1, Kre6
$S_{0.8}90$	Swi4/Swi6	Msb2
$S_{0.8}31$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.8}43$	Swi4/Swi6	Exg1
$S_{0.8}65$	Swi4/Swi6	Cwp1
$S_{0.8}13$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.8}17$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.8}76$	Fkh1	Arp7
$S_{0.8}74$	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.8}33$	Mcm1	Ste2
$S_{0.8}32$	Mcm1	Far1

$\gamma = 0.2 \Longrightarrow$

$$
\begin{aligned}
\{1 &\rightarrow \{\{4,*\},\{1,7\},\{1,*\},\{1,*\},\\
&\quad \{1,1\},\{1,2\},\{1,0\},\{1,0\},\{1,0\}\},\\
2 &\rightarrow \{\{7,1\},\{1,5\},\{1,1\}\},\\
3 &\rightarrow \{\{5,2\},\{1,5\}\},\\
4 &\rightarrow \{\{2,5\},\{1,5\},\{1,1\}\},\\
5 &\rightarrow \{\{1,3\}\},\\
6 &\rightarrow \{\{3,1\},\{1,5\}\},\\
7 &\rightarrow \{\{2,1\}\},\\
8 &\rightarrow \{\{2,4\},\{1,1\},\{1,0\}\},\\
9 &\rightarrow \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}
$$

$\text{Error\_score}(0.2) = 38 + 94 = 132$

$\gamma = 0.4 \Longrightarrow$

$$
\begin{aligned}
\{1 &\rightarrow \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},\\
&\quad \{2,*\},\{1,2\},\{1,0\},\{1,0\}\},\\
2 &\rightarrow \{\{8,6\},\{1,1\}\},\\
3 &\rightarrow \{\{5,2\},\{1,13\}\},\\
4 &\rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},\\
5 &\rightarrow \{\{1,3\}\},\\
6 &\rightarrow \{\{3,1\},\{1,13\}\},\\
7 &\rightarrow \{\{2,1\}\},\\
8 &\rightarrow \{\{2,12\},\{1,*\},\{1,1\}\},\\
9 &\rightarrow \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}
$$

$\text{Error\_score}(0.4) = 78 + 86 = 164$

Table 10: RN Data, $\gamma = 0.91$ (Shrinkage Clusters)

S49	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
S73	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
S45	Swi4/Swi6	Gic1, Kre6
S15	Swi4/Swi6	Msb2
S90	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
S56	Swi4/Swi6	Exg1
S46	Swi4/Swi6	Cwp1
S71	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
S61	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
S37	Fkh1	Arp7
S7	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
S91	Mcm1	Ste2
S92	Mcm1	Far1

Table 11: RN Data, $\gamma = 1.0$ (Pearson Clusters)

P10	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
P68	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
P1	Swi4/Swi6	Gic1, Kre6
P39	Swi4/Swi6	Msb2
P66	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
P20	Swi4/Swi6	Exg1
P2	Swi4/Swi6	Cwp1
P72	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
P53	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
P12	Fkh1	Arp7
P46	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
P64	Mcm1	Ste2
P65	Mcm1	Far1

$\gamma = 0.6 \Longrightarrow$

$$
\begin{aligned}
\{1 &\rightarrow \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},\\
&\quad\ \{2,*\},\{1,2\},\{1,0\},\{1,0\}\},\\
2 &\rightarrow \{\{8,6\},\{1,1\}\},\\
3 &\rightarrow \{\{5,2\},\{1,13\}\},\\
4 &\rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},\\
5 &\rightarrow \{\{1,0\}\},\\
6 &\rightarrow \{\{3,*\},\{1,13\}\},\\
7 &\rightarrow \{\{2,1\}\},\\
8 &\rightarrow \{\{2,12\},\{1,1\},\{1,0\}\},\\
9 &\rightarrow \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}
$$

Error_score(0.6) $= 75 + 86 = 161$

$\gamma = 0.8 \Longrightarrow$

$$
\begin{aligned}
\{1 &\rightarrow \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},\\
&\quad\ \{1,*\},\{2,*\},\{1,2\},\{1,0\}\},\\
2 &\rightarrow \{\{8,6\},\{1,1\}\},\\
3 &\rightarrow \{\{5,2\},\{1,13\}\},\\
4 &\rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},\\
5 &\rightarrow \{\{1,0\}\},\\
6 &\rightarrow \{\{3,*\},\{1,13\}\},\\
7 &\rightarrow \{\{2,1\}\},\\
8 &\rightarrow \{\{2,12\},\{1,1\},\{1,0\}\},\\
9 &\rightarrow \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}
$$

Error_score(0.8) $= 75 + 86 = 161$

$\gamma = 0.91(S) \Longrightarrow$

$$\begin{aligned}\{1 &\rightarrow \{\{4,*\},\{1,13\}\{1,*\},\{1,*\},\\ &\quad \{1,*\},\{2,*\},\{1,2\},\{1,0\}\},\\ 2 &\rightarrow \{\{8,6\},\{1,1\}\},\\ 3 &\rightarrow \{\{5,2\},\{1,13\}\},\\ 4 &\rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},\\ 5 &\rightarrow \{\{1,0\}\},\\ 6 &\rightarrow \{\{3,*\},\{1,13\}\},\\ 7 &\rightarrow \{\{2,1\}\},\\ 8 &\rightarrow \{\{2,12\},\{1,1\},\{1,0\}\},\\ 9 &\rightarrow \{\{1,*\},\{1,*\}\}\\ \}&\end{aligned}$$

$\text{Error\_score}(0.91) = 75 + 86 = 161$

$\gamma = 1.0(P) \Longrightarrow$

$$\begin{aligned}\{1 &\rightarrow \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},\\ &\quad \{1,*\},\{2,*\},\{1,2\},\{1,0\}\},\\ 2 &\rightarrow \{\{8,6\},\{1,1\}\},\\ 3 &\rightarrow \{\{5,2\},\{1,13\}\},\\ 4 &\rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},\\ 5 &\rightarrow \{\{1,0\}\},\\ 6 &\rightarrow \{\{3,*\},\{1,13\}\},\\ 7 &\rightarrow \{\{2,1\}\},\\ 8 &\rightarrow \{\{2,12\},\{1,1\},\{1,0\}\},\\ 9 &\rightarrow \{\{1,*\},\{1,*\}\}\\ \}&\end{aligned}$$

$\text{Error\_score}(1.0) = 75 + 86 = 161$

Clearly, in this notation, γ values of 0.8, 0.91, and 1.0 give identical cluster groupings, and the best error score is attained at $\gamma = 0.2$.

To improve the estimated value of γ, we proceeded to correct the second shortcoming due to the statistical dependence among the experiments. We sought to reduce the effective number of experiments by subsampling from the set of all (possibly correlated) experiments — the candidates were chosen via clustering all the experiments, i.e., columns of the data matrix, and then selecting one representative experiment from each cluster of experiments. We then clustered the subsampled data, once again using the cut-off correlation value of 0.60. The resulting cluster sets under the Eisen, Shrinkage, and Pearson metrics are given in Tables 12, 13, and 14, respectively.

The subsampled data yielded the lower estimated value $\gamma \simeq 0.66$. In our set notation, the resulting clusters with the corresponding error scores can be written as follows:

Table 12: RN Subsampled Data, $\gamma = 0.0$ (Eisen)

E58	Swi4/Swi6	Cln1, Och1
E68	Swi4/Swi6	Cln2, Msb2, Rsr1, Bud9, Mnn1, Exg1
	Swi6/Mbp1	Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45, Mcm2
	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1, Arp7
	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
	Ace2/Swi5	Egt2
	Mcm1	Mcm3, Mcm6, Cdc6
E29	Swi4/Swi6	Gic1
E64	Swi4/Swi6	Gic2
E33	Swi4/Swi6	Kre6, Cwp1
	Swi6/Mbp1	Clb5, Clb6
	Swi4/Swi6	Hta3
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
E73	Fkh1	Tel2
E23	Ace2/Swi5	Cts1
E43	Mcm1	Ste2
E66	Mcm1	Far1

$\gamma = 0.0(E) \Longrightarrow$

$$\begin{aligned}\{1 &\rightarrow \{\{6,23\},\{2,*\},\{2,5\},\{1,*\},\{1,*\}\},\\ 2 &\rightarrow \{\{7,22\},\{2,5\}\},\\ 3 &\rightarrow \{\{5,24\},\{1,6\}\},\\ 4 &\rightarrow \{\{3,26\},\{1,*\}\},\\ 5 &\rightarrow \{\{1,28\}\},\\ 6 &\rightarrow \{\{3,26\},\{1,6\}\},\\ 7 &\rightarrow \{\{1,*\},\{1,28\}\},\\ 8 &\rightarrow \{\{3,26\},\{1,6\}\},\\ 9 &\rightarrow \{\{1,*\},\{1,*\}\}\\ \}&\end{aligned}$$

$\text{Error\_score}(0.0) = 370 + 79 = 449$

Table 13: RN Subsampled Data, $\gamma = 0.66$ (Shrinkage)

S49	Swi4/Swi6	Cln1, Bud9, Och1
	Ace2/Swi5	Egt2
	Mcm1	Cdc6
S6	Swi4/Swi6	Cln2, Gic2, Msb2, Rsr1, Mnn1, Exg1
	Swi6/Mbp1	Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
S32	Swi4/Swi6	Gic1
S65	Swi4/Swi6	Kre6, Cwp1
	Swi6/Mbp1	Clb5, Clb6
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
S15	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
S11	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
S60	Swi4/Swi6	Hta3
S30	Fkh1	Arp7
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
S62	Fkh1	Tem1
S53	Ace2/Swi5	Cts1
S14	Mcm1	Mcm6
S35	Mcm1	Ste2
S36	Mcm1	Far1

Table 14: RN Subsampled Data, $\gamma = 1.0$ (Pearson)

P1	Swi4/Swi6	Cln1, Och1
P15	Swi4/Swi6	Cln2, Rsr1, Mnn1
	Swi6/Mbp1	Cdc21, Dun1, Rad51, Cdc45, Mcm2
	Mcm1	Mcm3
P29	Swi4/Swi6	Gic1
P2	Swi4/Swi6	Gic2
P3	Swi4/Swi6	Msb2, Exg1
	Swi6/Mbp1	Rnr1
P51	Swi4/Swi6	Bud9
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
	Ace2/Swi5	Egt2
	Mcm1	Cdc6
P11	Swi4/Swi6	Kre6
P62	Swi4/Swi6	Cwp1
	Swi6/Mbp1	Clb5, Clb6
	Swi4/Swi6	Hta3
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
P49	Swi6/Mbp1	Rad27
	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
P10	Fkh1	Tel2
	Mcm1	Mcm6
P23	Fkh1	Arp7
P50	Fkh1	Tem1
P69	Ace2/Swi5	Cts1
P42	Mcm1	Ste2
P13	Mcm1	Far1

$$\gamma = 0.66(S) \Longrightarrow$$
$$\begin{aligned}\{1 &\rightarrow \{\{6,6\},\{3,2\},\{2,5\},\{1,*\}\},\\ 2 &\rightarrow \{\{6,6\},\{2,5\},\{1,1\}\},\\ 3 &\rightarrow \{\{5,2\},\{1,*\}\},\\ 4 &\rightarrow \{\{2,5\},\{1,3\},\{1,6\}\},\\ 5 &\rightarrow \{\{1,*\}\},\\ 6 &\rightarrow \{\{3,1\},\{1,6\}\},\\ 7 &\rightarrow \{\{1,*\},\{1,4\}\},\\ 8 &\rightarrow \{\{1,*\},\{1,1\},\{1,4\},\{1,6\}\},\\ 9 &\rightarrow \{\{1,*\},\{1,*\}\}\\ \}&\end{aligned}$$
$$\text{Error\_score}(0.66) = 76 + 88 = 164$$

$$\gamma = 1.0(P) \Longrightarrow$$
$$\begin{aligned}\{1 &\rightarrow \{\{3,6\},\{2,*\},\{2,1\},\{1,*\},\\ &\quad\ \{1,*\},\{1,*\},\{1,5\},\{1,5\}\},\\ 2 &\rightarrow \{\{5,4\},\{2,4\},\{1,2\},\{1,7\}\},\\ 3 &\rightarrow \{\{5,3\},\{1,5\}\},\\ 4 &\rightarrow \{\{2,6\},\{1,*\},\{1,1\}\},\\ 5 &\rightarrow \{\{1,*\}\},\\ 6 &\rightarrow \{\{3,3\},\{1,5\}\},\\ 7 &\rightarrow \{\{1,*\},\{1,5\}\},\\ 8 &\rightarrow \{\{1,1\},\{1,5\},\{1,5\},\{1,8\}\},\\ 9 &\rightarrow \{\{1,*\},\{1,*\}\}\\ \}&\end{aligned}$$
$$\text{Error\_score}(1.0) = 69 + 107 = 176$$

From the tables for the range-normalized, subsampled

yeast data, as well as by comparing the error scores, one can conclude that for the same clustering algorithm and threshold value, Pearson tends to introduce more false-negatives and Eisen tends to introduce more false-positives than Shrinkage, as Shrinkage reduces these errors by combining the good properties of both algorithms. This observation is consistent with our mathematical analysis and the simulation presented in section 4.1.

5 Discussion

Microarray-based genomic analysis and other similar high-throughput methods have begun to occupy an increasingly important role in biology, as they have helped to create a visual image of the state-space trajectories at the core of the cellular processes. This analysis will address directly to the observational nature of the "new" biology. As a result, we need to develop our ability to "see," accurately and reproducibly, the information in the massive amount of quantitative measurements produced by these approaches — or be able to ascertain when what we "see" is unreliable and forms a poor basis for proposing novel hypotheses. Our investigation demonstrates the fragility of many of these analysis algorithms when used in the context of a small number of experiments. In particular, we see that a small perturbation of, or a small error in the estimation of, a parameter (the shrinkage parameter) has a significant effect on the overall conclusion. The errors in the estimators manifest themselves by missing certain biological relations between two genes (false-negatives) or by proposing phantom relations between two otherwise unrelated genes (false-positives).

A global picture of these interactions can be seen in Figure 3, the Receiver Operator Characteristic (ROC) figure, with each curve parametrized by the cut-off threshold in the range of $[-1, 1]$. An ROC curve ([12]) for a given metric plots sensitivity against (1−specificity), where

Sensitivity = fraction of positives detected by a metric

$$= \frac{\mathrm{TP}(\gamma)}{\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma)},$$

Specificity = fraction of negatives detected by a metric

$$= \frac{\mathrm{TN}(\gamma)}{\mathrm{TN}(\gamma) + \mathrm{FP}(\gamma)},$$

and $\mathrm{TP}(\gamma)$, $\mathrm{FN}(\gamma)$, $\mathrm{FP}(\gamma)$, and $\mathrm{TN}(\gamma)$ denote the number of True Positives, False Negatives, False Positives, and True Negatives, respectively, arising from a metric associated with a given γ. (Recall that γ is 0.0 for Eisen, 1.0 for Pearson, and is computed according to (14) for Shrinkage, which yields 0.66 on this data set.) For each pair of genes, $\{j, k\}$, we define these events using our hypothesis (see section 4.2) as a measure of truth:

TP: $\{j, k\}$ are in the same group (see Table 1) and $\{j, k\}$ are placed in the same cluster;

FP: $\{j, k\}$ are in different groups, but $\{j, k\}$ are placed in the same cluster;

TN: $\{j, k\}$ are in different groups and $\{j, k\}$ are placed in different clusters; and

FN: $\{j, k\}$ are in the same group, but $\{j, k\}$ are placed in different clusters.

$\mathrm{FP}(\gamma)$ and $\mathrm{FN}(\gamma)$ were already defined in equations (15) and (16), respectively, and we define

$$\mathrm{TP}(\gamma) = \sum_{x} \sum_{j=1}^{n_x} \binom{y_j}{2} \tag{18}$$

and

$$\mathrm{TN}(\gamma) = \mathrm{Total} - (\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma) + \mathrm{FP}(\gamma)) \tag{19}$$

where Total= $\binom{44}{2} = 946$ is the total # of gene pairs $\{j, k\}$ in Table 1.

The ROC figure suggests the best threshold to use for each metric, and can also be used to select the best metric to use for a particular sensitivity.

The dependence of the error scores on the threshold can be more clearly seen from Figure 4. It shows that the conclusions we draw in section 4.3 hold for a wide range of threshold values, and hence a threshold value of 0.60 is a reasonable representative value.

As a result, in order to study the clustering algorithms and their effectiveness, one may ask the following questions. If one must err, is it better to err on the side of more false-positives or more false-negatives? What are the relative costs of these two kinds of errors? In general, since false-negatives may cause the inference process to ignore useful information for certain novel genes, and since false-positives may result in noise in the information provided to the algorithms used in analyzing regulatory patterns, intelligent answers to our questions depend crucially on how the cluster information is used in the subsequent discovery processes.

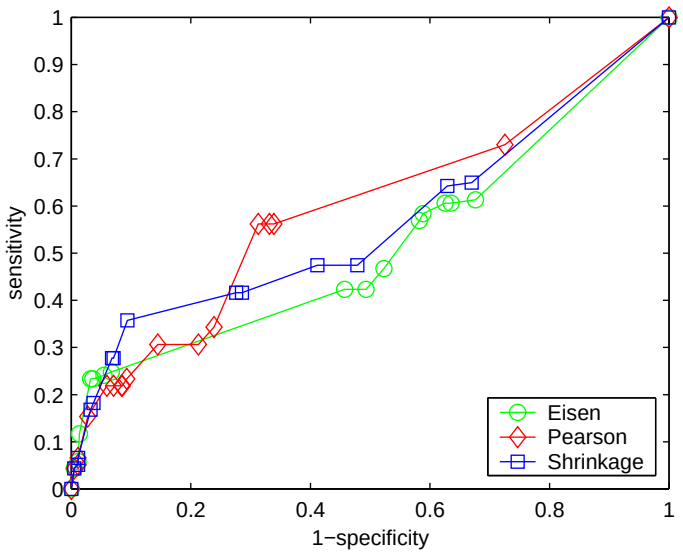

Figure 3: Receiver operator characteristic curves. Each curve is parametrized by the cut-off value $\theta \in \{1.0, 0.95, \ldots, -1.0\}$

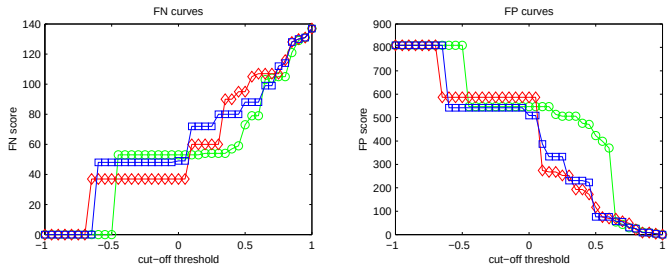

Figure 4: FN and FP curves, plotted as functions of θ.

We thank Mike Wigler, Misha Gromov, Ale Carbone, Thomas Anantharaman, Gloria Coruzzi, Fabio Piano, and Kris Gunslaus for offering many exciting ideas, useful contributions, and ways to think about transcriptomes, microarrays, statistics, and algorithms. Many of our close colleagues from NYU Bioinformatics group, Cold Spring Harbor Laboratory, and Mt. Sinai School of Medicine have directly and indirectly contributed to this effort: Toto Paxia, Raoul Daruwala, Joey Zhou, Archi Rudra, Naomi Silver, Frank Park, Ilya Nemenman, Will Casey, Marco Antoniotti, and Joe McQuown. To all of them, we are grateful. The work reported in this paper was supported by grants from NSF's Qubic program, DARPA, HHMI biomedical support research grant, the US Department of Energy, the US Air Force, National Institutes of Health, and New York State Office of Science, Technology & Academic Research.

References

[1] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. Cluster Analysis and Display of Genome-wide Expression Patterns. *Proceedings of the National Academy of Sciences, USA* **95**: 14863–14868, 1998.

[2] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W. Parallel Human Genome Analysis: Microarray-based Expression Monitoring of 1000 Genes. *Proceedings of the National Academy of Sciences, USA* **93**: 10614–10619, 1996.

[3] DeRisi, J.L., Iyer, V.R., and Brown, P.O. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* **278**: 680–686, 1997.

[4] Spellman, P.T., Sherlock, G., Iyer, V.R., Zhang, M., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9**: 3273–3297, 1998.

[5] Chu, S., DeRisi, J.L., Eisen, M.B., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. The Transcriptional Program of Budding Yeast. *Science* **282**: 699–705, 1998.

[6] Sokal, R.R. and Michener, C.D. A Statistical Method for Evaluating Systematic Relationships. *The University of Kansas Scientific Bulletin* **38**: 1409–1438, 1958.

[7] James, W. and Stein, C. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium Mathematical Statistics and Probability*, (ed. Neyman, J.), Vol. **1**: 361–379. University of California Press, 1961.

[8] Hoffman, K. Stein Estimation - A Review. *Statistical Papers*, **41(2)**: 127–158, 2000.

[9] Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W.L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B.J., Robinson, A., Bassett, D., Stoeckert, C.J. Jr., and Brazma, A. Design and Implementation of Microarray Gene Expression Markup Language (MAGE-ML). *Genome Biology* **3(9)**: research0046.1–0046.9, 2002.

[10] Wolfram, S. *The Mathematica Book*. Cambridge University Press, 4th edition, 1999.

[11] Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle. *Cell* **106**: 697–708, 2001.

[12] Egan, J.P. *Signal Detection Theory and ROC analysis*. Academic Press, New York, 1975.

A Appendix

A.1 Receiver Operator Characteristic Curves (More Details)

A.1.1 Definitions

As a measure of truth, we take our working hypothesis, namely, the transcriptional activator table (Table 1). Thus, if two genes are in the same group, they "belong in the same cluster", and if they are in different groups, they "belong in different clusters". We will generate an ROC curve for each metric used (i.e., one for Eisen, one for Pearson, and one for Shrinkage).

Event: grouping of (cell cycle) genes into clusters;

Threshold: cut-off similarity value at which the hierarchy tree is cut into clusters.

Our cell-cycle gene table consists of 44 genes, which gives us $C(44,2) = 946$ gene pairs. For each (unordered) gene pair $\{j,k\}$, we define the following events:

TP: $\{j,k\}$ are in the same group and $\{j,k\}$ are placed in the same cluster;

FP: $\{j,k\}$ are in different groups, but $\{j,k\}$ are placed in the same cluster;

TN: $\{j,k\}$ are in different groups and $\{j,k\}$ are placed in different clusters; and

FN: $\{j,k\}$ are in the same group, but $\{j,k\}$ are placed in different clusters.

Thus,

$$\begin{aligned}
\mathrm{TP}(\gamma) &= \sum_{\{j,k\}} \mathrm{TP}(\{j,k\}) \\
\mathrm{FP}(\gamma) &= \sum_{\{j,k\}} \mathrm{FP}(\{j,k\}) \\
\mathrm{TN}(\gamma) &= \sum_{\{j,k\}} \mathrm{TN}(\{j,k\}) \\
\mathrm{FN}(\gamma) &= \sum_{\{j,k\}} \mathrm{FN}(\{j,k\})
\end{aligned}$$

where the sums are taken over all 946 unordered pairs of genes.

Two other quantities involved in ROC curve generation are

Sensitivity = fraction of positives detected by a metric

$$= \frac{\mathrm{TP}(\gamma)}{\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma)}. \tag{20}$$

Specificity = fraction of negatives detected by a metric

$$= \frac{\mathrm{TN}(\gamma)}{\mathrm{TN}(\gamma) + \mathrm{FP}(\gamma)}. \tag{21}$$

An ROC curve plots sensitivity, on the y-axis, as a function of (1− specificity), on the x-axis, with each point on the plot corresponding to a different cut-off value. We create a different curve for each of the three metrics.

The following sections describe how the quantities $\mathrm{TP}(\gamma)$, $\mathrm{FN}(\gamma)$, $\mathrm{FP}(\gamma)$, and $\mathrm{TN}(\gamma)$ can be computed using our set notation for clusters. Recall from section 4.3:

$$\Big\{x \to \{\{y_1,z_1\},\{y_2,z_2\},\ldots,\{y_{n_x},z_{n_x}\}\}\Big\}_{x=1}^{\#\text{ of groups}}$$

A.1.2 Computation

TP

$$\mathrm{TP}(\gamma) = \sum_{\{j,k\}} \mathrm{TP}(\{j,k\}) =$$

gene pairs that were placed in the same cluster and belong in the same group.

For each group x given in set notation as

$$x \to \{\{y_1,z_1\},\ldots,\{y_{n_x},z_{n_x}\}\},$$

we count pairs from each y_j, i.e.,

$$\mathrm{TP}(x) = \binom{y_1}{2} + \cdots + \binom{y_{n_x}}{2} = \sum_{j=1}^{n_x} \binom{y_j}{2}$$

Totaling over all groups yields

$$\mathrm{TP}(\gamma) = \sum_{x=1}^{\#\text{ groups}} \mathrm{TP}(x) = \sum_x \sum_{j=1}^{n_x} \binom{y_j}{2}$$

FN

$$\mathrm{FN}(\gamma) = \sum_{\{j,k\}} \mathrm{FN}(\{j,k\}) =$$

gene pairs that belong in the same group but were placed into different clusters.

We must count every pair that got separated.

$$\mathrm{FN}(x) = \begin{cases} \displaystyle\sum_{j=1}^{n_x} \sum_{k=j+1}^{n_x} y_j \cdot y_k & \text{if } n_x \geq 2, \text{ or} \\ 0, & \text{if } n_x = 1. \end{cases}$$

However, when $n_x = 1$, there is no pair $\{j,k\}$ that satisfies the triple inequality $1 \leq j < k \leq n_x$, and hence, we do not have to treat it as a special case.

$$\therefore \mathrm{FN}(\gamma) = \sum_{x=1}^{\#\text{ groups}} \mathrm{FN}(x) = \sum_x \sum_{1 \leq j < k \leq n_x} y_j \cdot y_k$$

FP

$$\mathrm{FP}(\gamma) = \sum_{\{j,k\}} \mathrm{FP}(\{j,k\}) =$$

gene pairs that belong in different groups but got placed in the same cluster.

The expression

$$\sum_x \sum_{j=1}^{n_x} y_j \cdot z_j$$

counts every false-positive pair $\{j,k\}$ twice: first, when looking at j's group, and again, when looking at k's group.

$$\therefore \mathrm{FP}(\gamma) = \frac{1}{2} \sum_x \sum_{j=1}^{n_x} y_j \cdot z_j$$

TN

$$\mathrm{TN}(\gamma) = \sum_{\{j,k\}} \mathrm{TN}(\{j,k\}) =$$

gene pairs that belong in different groups and got placed in different clusters.

Instead of counting true-negatives from our notation, we use the fact that we know the other three scores and the total they all add up to.

Complementarity Given a gene pair $\{j,k\}$, exactly one of the events $\{\mathrm{TP}(\{j,k\}), \mathrm{FN}(\{j,k\}), \mathrm{FP}(\{j,k\}), \mathrm{TN}(\{j,k\})\}$ is true, i.e., exactly one of them $= 1$, while the rest $= 0$. This implies

$$\begin{aligned}
&\sum_{\{j,k\}} \mathrm{TP}(\{j,k\}) + \sum_{\{j,k\}} \mathrm{FN}(\{j,k\}) + \\
&\quad + \sum_{\{j,k\}} \mathrm{FP}(\{j,k\}) + \sum_{\{j,k\}} \mathrm{TN}(\{j,k\}) = \\
&= \mathrm{TP}(\gamma) + \mathrm{FN}(\gamma) + \mathrm{FP}(\gamma) + \mathrm{TN}(\gamma) = \\
&= \binom{44}{2} = \frac{44 \cdot 43}{2} = 946 = \mathrm{Total}
\end{aligned}$$

$$\therefore \mathrm{TN}(\gamma) = \mathrm{Total} - (\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma) + \mathrm{FP}(\gamma))$$

A.1.3 Plotting ROC curves

For each cut-off value θ, we can compute $\mathrm{TP}(\gamma)$, $\mathrm{FN}(\gamma)$, $\mathrm{FP}(\gamma)$, and $\mathrm{TN}(\gamma)$ as described in the previous section, with $\gamma \in \{0.0, 0.66, 1.0\}$ corresponding to Eisen, Shrinkage, and Pearson, respectively. Then, the sensitivity and specificity are computed from equations (20) and (21), and we can plot sensitivity vs (1− specificity), as shown in Figure 3.

We can also examine the effect of the cut-off threshold θ on the FN and FP scores individually, as shown in Figure 4.

A 3-dimensional plot of (1− specificity) on the $x-$axis, sensitivity on the $y-$axis, and threshold on the $z-$axis offers an interesting view, as shown in Figure 5.

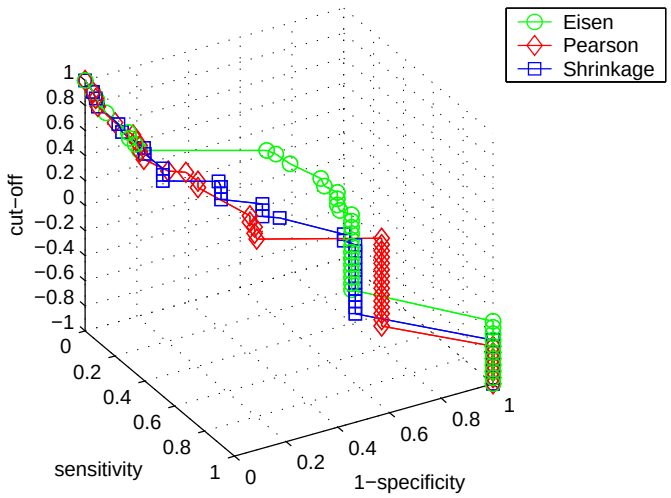

Figure 5: ROC curves, with threshold plotted on the $z-$axis.

A.2 Computing the marginal pdf for X_j

$$\begin{aligned}
f(X_j) &= \mathbf{E}_{\theta_j} f(X_j|\theta_j) = \int_{-\infty}^{\infty} f(X_j|\theta)\pi(\theta)d\theta \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(X_j-\theta)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\tau}\, e^{-\frac{\theta^2}{2\tau^2}} d\theta \\
&= \frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{(X_j-\theta)^2}{\sigma^2} + \frac{\theta^2}{\tau^2}\right)} d\theta \qquad (22)
\end{aligned}$$

First, rewrite the exponent as a complete square:

$$\begin{aligned}
\frac{(X_j-\theta)^2}{\sigma^2} + \frac{\theta^2}{\tau^2} &= \frac{1}{\sigma^2\tau^2}\left[\tau^2(X_j-\theta)^2 + \sigma^2\theta^2\right] \\
&= \frac{1}{\sigma^2\tau^2}\left[\tau^2 {X_j}^2 - 2\tau^2 X_j\theta + \tau^2\theta^2 + \sigma^2\theta^2\right] \\
&= \frac{1}{\sigma^2\tau^2}\left[(\sigma^2+\tau^2)\theta^2 - 2\tau^2 X_j\theta + \tau^2 {X_j}^2\right] \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left[\theta^2 - 2\frac{\tau^2}{\sigma^2+\tau^2}X_j\theta + \frac{\tau^2}{\sigma^2+\tau^2}{X_j}^2\right] \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left[\left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2 \right. \\
&\quad \left. \underbrace{-\left(\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2 + \frac{\tau^2}{\sigma^2+\tau^2}{X_j}^2}\right] \qquad (23)
\end{aligned}$$

- $$\begin{aligned}
&\frac{\tau^2}{\sigma^2+\tau^2}{X_j}^2 - \left(\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2 \\
&= {X_j}^2\left(\frac{\tau^2}{\sigma^2+\tau^2}\right)\left(1 - \frac{\tau^2}{\sigma^2+\tau^2}\right) \\
&= {X_j}^2\left(\frac{\tau^2}{\sigma^2+\tau^2}\right)\left(\frac{\sigma^2}{\sigma^2+\tau^2}\right) \\
&= {X_j}^2\frac{\sigma^2\tau^2}{(\sigma^2+\tau^2)^2} \qquad (24)
\end{aligned}$$

Substituting (24) into (23) yields

$$\begin{aligned}
\frac{(X_j-\theta)^2}{\sigma^2} + \frac{\theta^2}{\tau^2} &= \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2 + \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}{X_j}^2\frac{\sigma^2\tau^2}{(\sigma^2+\tau^2)^2} \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2 + \frac{{X_j}^2}{\sigma^2+\tau^2} \qquad (25)
\end{aligned}$$

Now use the completed square in (25) to continue the computation in (22).

$$\begin{aligned}
&f(X_j) \\
&= \frac{1}{2\pi\sigma\tau}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2} e^{-\frac{1}{2}\frac{{X_j}^2}{\sigma^2+\tau^2}} d\theta \\
&= \frac{e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}}}{2\pi\sigma\tau}\int_{-\infty}^{\infty} \exp\left[-\left(\frac{\theta - \frac{\tau^2}{\sigma^2+\tau^2}X_j}{\sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}}\right)^2\right] d\theta
\end{aligned}$$

Make the substitution

$$\varphi = \left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X_j\right) \Big/ \sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}$$

Then

$$\begin{aligned} d\varphi &= d\theta \Big/ \sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}} \\ d\theta &= \sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}\, d\varphi \\ \theta = \pm\infty &\implies \varphi = \pm\infty \end{aligned}$$

and

$$\begin{aligned} f(X_j) &= \frac{e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}}}{2\pi\sigma\tau} \int_{-\infty}^{\infty} e^{-\varphi^2} \sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}\, d\varphi \\ &= \frac{e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}}}{\pi\sqrt{2(\sigma^2+\tau^2)}} \underbrace{\int_{-\infty}^{\infty} e^{-\varphi^2} d\varphi}_{\sqrt{\pi}} \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\, e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}} \end{aligned}$$

Therefore

$$f(X_j) = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\, e^{-\frac{{X_j}^2}{2(\sigma^2+\tau^2)}} \tag{26}$$

A.3 Calculation of the posterior distribution of θ_j

Since the subscript j remains constant throughout the calculation, it will be dropped in this appendix. Herein, θ_j will be replaced by θ, and X_j by X.

$$\begin{aligned} \pi(\theta|X) &= \frac{f(X|\theta)\,\pi(\theta)}{f(X)} = \frac{f(X,\theta)}{f(X)} \\ &= \frac{(1/2\pi\sigma\tau)\, \exp\left[-\left(\frac{\theta^2}{2\tau^2} + \frac{(X-\theta)^2}{2\sigma^2}\right)\right]}{\left(1\Big/\sqrt{2\pi(\sigma^2+\tau^2)}\right)\, \exp\left[-\frac{X^2}{2(\sigma^2+\tau^2)}\right]} \\ &= \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}\, \exp\left[-\frac{1}{2}\underbrace{\left(\frac{\theta^2}{\tau^2} + \frac{(X-\theta)^2}{\sigma^2} - \frac{X^2}{\sigma^2+\tau^2}\right)}\right] \end{aligned}$$

- $\frac{\theta^2}{\tau^2} + \frac{(X-\theta)^2}{\sigma^2} - \frac{X^2}{\sigma^2+\tau^2} =$

$$\begin{aligned} &= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\Big[\sigma^2(\sigma^2+\tau^2)\theta^2 \\ &\qquad + \tau^2(\sigma^2+\tau^2)\overbrace{(X-\theta)^2}^{X^2-2X\theta+\theta^2} - \sigma^2\tau^2X^2\Big] \\ &= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\Big[\theta^2\left(\sigma^2(\sigma^2+\tau^2) + \tau^2(\sigma^2+\tau^2)\right) \\ &\qquad - 2\tau^2(\sigma^2+\tau^2)X\theta \\ &\qquad + X^2\left(\tau^2(\sigma^2+\tau^2) - \sigma^2\tau^2\right)\Big] \\ &= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\Big[\theta^2(\sigma^2+\tau^2)^2 \\ &\qquad - 2(\sigma^2+\tau^2)\theta\cdot\tau^2X + \tau^4X^2\Big] \\ &= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\left((\sigma^2+\tau^2)\theta - \tau^2X\right)^2 \\ &= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}(\sigma^2+\tau^2)^2\left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X\right)^2 \\ &= \left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X\right)^2 \Big/ \frac{\sigma^2\tau^2}{\sigma^2+\tau^2} \end{aligned}$$

Therefore,

$$\pi(\theta|X) = \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}\, \exp\left[-\frac{\left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X\right)^2}{2\left(\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)}\right] \tag{27}$$

A.4 Proof of the fact that n independent observations from the Normal population $\mathcal{N}(\theta, \sigma^2)$ can be treated as a single observation from $\mathcal{N}(\theta, \sigma^2/n)$

Given the data y, $f(y|\theta)$ can be viewed as a function of θ. We then call it the *likelihood function* of θ for given y, and write

$$l(\theta|y) \propto f(y|\theta).$$

When y is a single data point from $\mathcal{N}(\theta, \sigma^2)$,

$$l(\theta|y) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta - x}{\sigma}\right)^2\right] = \exp\left[-\frac{1}{2\sigma^2}(\theta - x)^2\right], \tag{28}$$

where x is some function of y.

Now, suppose that $\vec{y} = (y_1, \ldots, y_n)$ represents a vector of n independent observations from $\mathcal{N}(\theta, \sigma^2)$. We can denote the sample mean by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The likelihood function of θ given such n independent observations from $\mathcal{N}(\theta, \sigma^2)$ is

$$l(\theta|\vec{y}) \propto \prod_i \exp\left[-\frac{1}{2\sigma^2}(y_i - \theta)^2\right] = \exp\left[-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right].$$

Also, since

$$\sum_{i=1}^{n}(y_i - \theta)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \theta)^2, \tag{29}$$

it follows that

$$\begin{aligned} l(\theta|\vec{y}) &\propto \underbrace{\exp\left[-\frac{1}{2\sigma^2}\sum_i (y_i - \bar{y})^2\right]}_{\text{const w.r.t. } \theta} \exp\left[-\frac{1}{2\sigma^2} n(\bar{y} - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2(\sigma^2/n)}(\theta - \bar{y})^2\right], \end{aligned} \tag{30}$$

which is a Normal function with mean $\bar{y}$ and variance σ^2/n. Comparing with (28), we can recognize that this is equivalent to treating the data $\vec{y}$ as a single observation $\bar{y}$ with mean θ and variance σ^2/n, i.e.,

$$\bar{y} \sim \mathcal{N}(\theta, \sigma^2/n). \tag{31}$$

Proof of (29):

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \theta)^2 &= \sum_i (y_i - \bar{y} + \bar{y} - \theta)^2 \\ &= \sum_i \left[(y_i - \bar{y})^2 + 2(y_i - \bar{y})(\bar{y} - \theta) + (\bar{y} - \theta)^2\right] \\ &= \sum_i (y_i - \bar{y})^2 + 2(\bar{y} - \theta)\sum_i (y_i - \bar{y}) + \sum_i (\bar{y} - \theta)^2 \\ &= \sum_i (y_i - \bar{y})^2 + 2(\bar{y} - \theta)\underbrace{\left(\sum_i y_i - \sum_i \bar{y}\right)}_{n\bar{y} - n\bar{y} = 0} + n(\bar{y} - \theta)^2 \\ &= \sum_i (y_i - \bar{y})^2 + n(\bar{y} - \theta)^2 \end{aligned}$$

A.5 Distribution of the Sum of two Independent Normal Random Variables

Let

$$\begin{aligned} X &\sim \mathcal{N}(0, \alpha^2) \\ Y &\sim \mathcal{N}(0, \beta^2) \end{aligned}$$

be two independent random variables.

Claim: $X + Y \sim \mathcal{N}(0, \alpha^2 + \beta^2)$
(We are only using this result for mean-0 Normal r.v.'s, although a more general result can be proven.)

Proof: (use moment generating functions)

$$\begin{aligned} m_X(t) &= \mathbf{E}\left(e^{tX}\right) = \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi}\alpha}\, e^{-\frac{1}{2\alpha^2}(x-0)^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}\alpha}\int_{-\infty}^{\infty} e^{-\frac{1}{2\alpha^2}\underbrace{[x^2 - 2\alpha^2 tx]}}\, dx \end{aligned} \tag{32}$$

Completing the square, we obtain

$$\begin{aligned} x^2 - 2\alpha^2 tx &= x^2 - 2(\alpha^2 t)x + (\alpha^2 t)^2 - (\alpha^2 t)^2 \\ &= (x - \alpha^2 t)^2 - (\alpha^4 t^2) \\ \frac{1}{\alpha^2}(x^2 - 2\alpha^2 tx) &= \big((x - \alpha^2 t)/\alpha\big)^2 - (\alpha^4 t^2)/\alpha^2 \\ &= \left(\frac{x - \alpha^2 t}{\alpha}\right)^2 - \alpha^2 t^2 \end{aligned} \tag{33}$$

Using the result of (33) in (32) yields

$$m_X(t) = \frac{e^{-\frac{1}{2}\left(-\alpha^2 t^2\right)}}{\sqrt{2\pi}\alpha}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x - \alpha^2 t}{\alpha}\right)^2} dx$$

$$\begin{aligned} \text{Let } y &= \frac{x - \alpha^2 t}{\alpha} \\ dy &= \frac{dx}{\alpha} \implies dx = \alpha\, dy \end{aligned}$$

With this substitution, we obtain

$$m_X(t) = \frac{e^{\frac{1}{2}\,\alpha^2 t^2}}{\sqrt{2\pi}\alpha} \cdot \alpha \underbrace{\int_{y=-\infty}^{\infty} e^{-\frac{1}{2}\,y^2}\, dy}_{\sqrt{2\pi}}$$

or

$$m_X(t) = e^{\frac{1}{2}\,\alpha^2 t^2} \tag{34}$$

Similarly

$$m_Y(t) = e^{\frac{1}{2}\,\beta^2 t^2} \tag{35}$$

To obtain the distribution of $X+Y$, it suffices to compute the corresponding moment generating function:

$$\begin{aligned} m_{X+Y}(t) &= \mathbf{E}\left(e^{t(X+Y)} \right) = \mathbf{E}\left(e^{tX}\ e^{tY} \right) \\ &= \mathbf{E}\left(e^{tX} \right) \mathbf{E}\left(e^{tY} \right) \quad \text{by independence of } X \text{ and } Y \\ &= m_X(t) \cdot m_Y(t) \\ &= e^{\frac{1}{2}\,\alpha^2 t^2} \cdot e^{\frac{1}{2}\,\beta^2 t^2} \quad \text{by (34) and (35)} \\ &= e^{\frac{1}{2}(\alpha^2+\beta^2)t^2}, \end{aligned}$$

which is a moment generating function of a Normal random variable with mean 0 and variance $\alpha^2+\beta^2$. Therefore,

$$X+Y \sim \mathcal{N}(0, \alpha^2+\beta^2). \tag{36}$$

A.6 Properties of the Chi-square Distribution

Let $X_1, X_2, \ldots, X_k$ be i.i.d.r.v.'s from standard Normal distribution, i.e.,

$$X_j \sim \mathcal{N}(0,1) \quad \forall j.$$

Then

$$\chi^2_k = X_1^2 + X_2^2 + \cdots + X_k^2 = \sum_{j=1}^{k} {X_j}^2$$

is a random variable from Chi-square distribution with k degrees of freedom, denoted

$$\chi^2_k \sim \chi^2_{(k)}.$$

It has the probability density function

$$f(x) = \begin{cases} \frac{1}{2^{k/2}\Gamma\,(k/2)}\, x^{k/2-1} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t}\, dt. \tag{37}$$

The result we are using is

$$\mathbf{E}\left(\frac{1}{\chi^2_k} \right) = \frac{1}{k-2} \quad \text{for } k > 2,$$

which can be obtained as follows:

$$\begin{aligned} \mathbf{E}\left(\frac{1}{\chi^2_k} \right) &= \int_{\mathcal{R}} \frac{1}{x}\, f(x)\, dx \\ &= \frac{1}{2^{k/2}\Gamma\,(k/2)} \int_0^\infty \frac{1}{x}\, x^{k/2-1} e^{-x/2}\, dx \\ &= \frac{1}{2^{k/2}\Gamma\,(k/2)} \int_0^\infty x^{k/2-2} e^{-x/2}\, dx \end{aligned} \tag{38}$$

Let

$$\begin{array}{rclcrcl} t &=& x/2 & \Longrightarrow & x &=& 2t \\ & & & & dx &=& 2dt \\ x &=& 0 & \Longrightarrow & t &=& 0 \\ x &=& \infty & \Longrightarrow & t &=& \infty \end{array}$$

$$\begin{aligned} &\int_0^\infty x^{k/2-2} e^{-x/2}\, dx \\ &= \int_{t=0}^\infty (2t)^{k/2-2}\, e^{-t}\, 2\, dt \\ &= 2^{k/2-2} \cdot 2 \int_0^\infty t^{k/2-2} e^{-t}\, dt. \end{aligned} \tag{39}$$

Let

$$\begin{array}{rclcrcl} u &=& e^{-t} & & dv &=& t^{k/2-2}\, dt \\ du &=& -e^{-t}\, dt & & v &=& \frac{t^{k/2-1}}{k/2-1} \qquad \text{for } k > 2 \end{array}$$

Integration by parts transforms (39) into

$$\begin{aligned} &= 2^{k/2-1} \left(\frac{1}{k/2-1} \underbrace{e^{-t} t^{k/2-1} \Big|_0^\infty}_{\longrightarrow 0} - \int_0^\infty \frac{1}{k/2-1}\, t^{k/2-1} \left(-e^{-t}\right)\, dt \right) \\ &= \frac{2^{k/2-1}}{k/2-1} \underbrace{\int_0^\infty t^{k/2-1} e^{-t}\, dt}_{\Gamma(k/2),\ \text{by (37)}} \\ &= \frac{2^{k/2-1}}{k/2-1}\, \Gamma(k/2) \end{aligned}$$

Substituting this result in (38) yields

$$\begin{aligned}\mathbf{E}\left(\frac{1}{\chi^2_k}\right) &= \frac{1}{2^{k/2}\Gamma(k/2)}\cdot\frac{2^{k/2-1}\Gamma(k/2)}{k/2-1}\\ &= \frac{1}{2\,(k/2-1)}\\ &= \frac{1}{k-2}\quad \text{for } k>2. \end{aligned}\tag{40}$$

A.7 Distribution of Sample Variance s^2

Let $X_j \sim \mathcal{N}(\mu,\sigma^2)$ for $j=1,\ldots,n$ be independent r.v.'s. We'll derive the joint distribution of

$$\frac{\sqrt{n}\,(\bar{X}-\mu)}{\sigma}\quad\text{and}\quad\frac{(n-1)\,s^2}{\sigma^2}.$$

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j-\bar{X}\right)^2\\ \frac{(n-1)\,s^2}{\sigma^2} &= \frac{n-1}{\sigma^2}\cdot\frac{1}{n-1}\sum_{j=1}^{n}\left(X_j-\bar{X}\right)^2\\ &= \sum_{j=1}^{n}\left(\frac{X_j-\bar{X}}{\sigma}\right)^2\end{aligned}$$

W.L.O.G. can reduce the problem to the case $\mathcal{N}(0,1)$, i.e., $\mu=0$, $\sigma^2=1$: Let $Z_j=(X_j-\mu)/\sigma$. Then

$$\begin{aligned}\bar{Z} &= \frac{1}{n}\sum Z_j=\frac{1}{n}\sum\left(\frac{X_j-\mu}{\sigma}\right)=\frac{1}{n}\left(\frac{\sum X_j}{\sigma}-\frac{\sum\mu}{\sigma}\right)\\ &= \frac{1}{n}\left(\frac{\sum X_j}{\sigma}-\frac{n\mu}{\sigma}\right)=\frac{1}{\sigma}\left(\frac{\sum X_j}{n}-\mu\right)=\frac{\bar{X}-\mu}{\sigma}\end{aligned}$$

and hence

$$\frac{\sqrt{n}\,\left(\bar{X}-\mu\right)}{\sigma}=\sqrt{n}\,\bar{Z}.\tag{41}$$

Also,

$$\begin{aligned}\frac{(n-1)\,s^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum\left(X_j-\bar{X}\right)^2\\ &= \frac{1}{\sigma^2}\sum\left((X_j-\mu)+\left(\mu-\bar{X}\right)\right)^2\\ &= \sum\left[\frac{X_j-\mu}{\sigma}-\frac{\bar{X}-\mu}{\sigma}\right]^2=\sum\left(Z_j-\bar{Z}\right)^2\end{aligned}\tag{42}$$

By (41) and (42), it suffices to derive the joint distribution of $\sqrt{n}\,\bar{Z}$ and $\sum_{j=1}^{n}\left(Z_j-\bar{Z}\right)^2$, where $Z_1,\ldots,Z_n$ are i.i.d. from $\mathcal{N}(0,1)$.

Let

$$P=\begin{pmatrix}\text{—}p_1\text{—}\\ \text{—}p_2\text{—}\\ \vdots\\ \text{—}p_n\text{—}\end{pmatrix}$$

be an $n\times n$ orthogonal matrix where

$$p_1=\left(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}\right)$$

and the remaining rows p_j are obtained by, say, applying Gramm-Schmidt to $\{p_1,e_2,e_3,\ldots,e_n\}$, where e_j is a standard unit vector in j^{th} direction in $\mathcal{R}^n$. Let

$$\begin{aligned}&\vec{Y}=P\vec{Z}\\ &= \begin{pmatrix}\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}}\\ \hline & & \vdots & \\ \hline\end{pmatrix}\begin{pmatrix}Z_1\\ Z_2\\ \vdots\\ Z_n\end{pmatrix}=\begin{pmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{pmatrix}\end{aligned}$$

Then

$$Y_1 = \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{n} Z_j \right) = \frac{1}{\sqrt{n}} n\bar{Z} = \sqrt{n}\, \bar{Z}. \tag{43}$$

Since P is orthogonal, it preserves vector lengths:

$$\begin{aligned} \|\vec{Y}\|^2 &= \|\vec{Z}\|^2 \\ \sum_{j=1}^{n} {Y_j}^2 &= \sum_{j=1}^{n} {Z_j}^2 \\ \implies \left(\sum_{j=1}^{n} {Y_j}^2 \right) - Y_1^2 &= \sum_{j=1}^{n} {Z_j}^2 - \left(\sqrt{n}\, \bar{Z} \right)^2 \quad \text{by (43)} \end{aligned}$$

Hence

$$\begin{aligned} \sum_{j=2}^{n} {Y_j}^2 &= \sum_{j=1}^{n} {Z_j}^2 - n\bar{Z}^2 = \sum_{j=1}^{n} {Z_j}^2 - 2n\bar{Z}^2 + n\bar{Z}^2 \\ &= \sum_{j=1}^{n} {Z_j}^2 - 2\bar{Z}(n\bar{Z}) + n\bar{Z}^2 \\ &= \sum_{j=1}^{n} {Z_j}^2 - 2\bar{Z} \left(\sum_{j=1}^{n} Z_j \right) + \sum_{j=1}^{n} \bar{Z}^2 \\ &= \sum_{j=1}^{n} \left(Z_j - \bar{Z} \right)^2 \end{aligned} \tag{44}$$

Since the Y_j's are mutually independent (by orthogonality of P), we can conclude that

$$\sum_{j=2}^{n} {Y_j}^2 = \sum_{j=1}^{n} \left(Z_j - \bar{Z} \right)^2$$

is independent of

$$Y_1 = \sqrt{n}\, \bar{Z}.$$

Also by orthogonality of P, $Y_j \sim \mathcal{N}(0,1)$ for $j = 1, \ldots, n$, so

$$\left(\sum_{j=2}^{n} {Y_j}^2 \right) \sim \chi^2_{(n-1)} \quad \text{(See Appendix A.6)}$$

and hence, by (42) and (44),

$$\frac{(n-1)\, s^2}{\sigma^2} \sim \chi^2_{(n-1)} \tag{45}$$

Since $\mathbf{E}\left(\chi^2_k\right) = k$, for $\chi^2_k \sim \chi^2_{(k)}$, we can see that

$$\mathbf{E}\left(\frac{(n-1)\, s^2}{\sigma^2} \right) = n - 1.$$

Also, since

$$\mathbf{E}\left(\frac{(n-1)\, s^2}{\sigma^2} \right) = \frac{n-1}{\sigma^2}\, \mathbf{E}\left(s^2\right),$$

we can conclude that

$$\mathbf{E}\left(s^2\right) = \frac{\sigma^2}{n-1} \cdot \frac{n-1}{\sigma^2}\, \mathbf{E}\left(s^2\right) = \frac{\sigma^2}{n-1} \cdot (n-1) = \sigma^2, \tag{46}$$

i.e., s^2 is an unbiased estimator of the variance σ^2.