PRINCIPAL COMPONENTS ANALYSIS TO SUMMARIZE MICROARRAY EXPERIMENTS: APPLICATION TO SPORULATION TIME SERIES

Soumya Raychaudhuri[*], Joshua M. Stuart[*], and Russ B. Altman[Ψ]
Stanford Medical Informatics
Stanford University, 251 Campus Drive, MSOB X-215, Stanford CA 94305-5479
{sxr, stuart, altman} @smi.stanford.edu

The enormous amount of data produced by microarray experiments can be unwieldy. A given series of microarray experiments produces observations of differential expression for thousands of genes across multiple conditions. These large data sets can be summarized with principal components analysis (PCA), a statistical technique that allows the key variables (or combinations of variables) in a multidimensional data set to be identified. Principal components analysis determines those key variables in the data that best explain the differences in the observations. Here we show the utility of applying PCA to expression data, where the experimental conditions are the variables, and the gene expression measurements are the observations. Thus, each component defines a linear combination of the experimental conditions that can be used to distinguish genes parsimoniously. Examination of the components also provides insight into what underlying factors are actually being measured in the experiment. We applied PCA to the publicly released yeast sporulation data set (Chu et al. 1998). In that work, 7 different measurements of gene expression were made over time. PCA on the time-points suggests that much of the observed variability in the experiment can be summarized in just 2 components—i.e. 2 variables capture most of the information. These underlying factors appear to represent (1) overall induction level and (2) change in induction level over time. A visualization of our results is made available (http://www.smi.stanford.edu/projects/helix/PCArray).

1 Introduction

The study of gene expression has been greatly facilitated by the application of the recently developed DNA microarray technology (Schena et al. 1995). DNA microarrays measure the expression of thousands of genes simultaneously. The anticipated flood of biological information produced by these experiments will open new doors into genetic analysis (Lander 1999). Expression patterns have already been used for a variety of inference tasks. For example, microarray data has been used to identify gene clusters based on co-expression (Eisen et al. 1998, Michaels et al. 1998), define metrics that measure a gene's involvement in a particular cellular event or process (Spellman et al. 1998), predict regulatory elements (Brazma et al. 1998), and reverse engineer transcription networks (D'Haeseleer et al. 1999, Liang

[*] These authors contributed equally to this communication.
[Ψ] To whom correspondence should be addressed.

et al. 1998). The success of these methods relies on the integrity of the expression data. Both experimental noise and non-independence among a set of experimental conditions may lead the inferential process astray. Considering or eliminating either of these complicating factors is non-trivial and can be over-looked in data analysis.

DNA microarrays consist of single-stranded DNA fragments affixed to a solid support (Chee et al. 1996, Chen et al. 1998, Duggan et al. 1999, Schena et al. 1995). Each spot on the microarray consists of a population of identical DNA fragments that represent one particular gene. To measure expression, the total RNA of a cell is harvested and labeled with fluorescent nucleotide tags during reverse transcription to make fluorescent probes. Commonly, two cell populations are used—cells under control and experimental conditions. The probes are then placed on the chip and permitted to hybridize with the target fragments on the corresponding spot. The intensity of the spot is approximately proportional to the probe and hence mRNA concentration. In a typical experiment, two colors (red and green) are used to measure expression of the experimental population relative to the control. Equal total mRNA probe concentrations are used to query the microarray and intensity ratios between the colors are calculated and reported as data (Schena et al. 1995).

Principal Components Analysis (PCA) is an exploratory multivariate statistical technique, originally introduced by Pearson (Basilevsky 1994, Everitt & Dunn 1992, Pearson 1901). Given m observations on n variables, the goal of PCA is to reduce the dimensionality of the data matrix by finding r new variables, where r is less than n. Termed principal components, these r new variables together account for as much of the variance in the original n variables as possible while remaining mutually uncorrelated and orthogonal. The goal is to reduce dimensionality while filtering noise in the process, making the data more accessible for visualization and analysis. Because each principal component is a linear combination of the original variables, it is often possible to ascribe meaning to what the components represent. For example, if we timed several people running in the 50m, 200m, 800m, and 3200m races, the component accounting for the most variability might represent overall fitness and the component accounting for the next most variability might distinguish the sprinters from the long-distance runners. Principal components analysis has been used in a wide range of biomedical problems. Two recent examples include unsupervised signal detection in functional magnetic resonance images (Lai & Fang 1999) and visualization of the genomic similarity among several different cell populations (Franklin et al. 1999).

We demonstrate the utility of PCA to the analysis of gene expression data by application to a published microarray experiment (Chu et. al. 1998). This experiment includes expression measurements on sporulating yeast cells taken at different time points. Analysis of this data by the same group identified 7 clusters for classifying key genes. These clusters were defined by the approximate times during which members are up-regulated.

2 Methods

Given a matrix of expression data, *A*, where each row corresponds to a different gene and each column corresponds to one of several different conditions to which the cells were exposed. The a_{it} entry of the matrix contains the i^{th} gene's relative expression ratio with respect to a control population under condition *t*. In an effort to equalize the influence of induction and repression on subsequent analysis, we applied the natural log transform to all ratios (Eisen et al. 1998). Up-regulated genes have a positive log expression ratio, while down-regulated genes have a negative log expression ratio.

We chose not to normalize the conditions to norm 0, variance 1 as recommended by certain text books of multivariate statistics (Everitt & Dunn 1992). This normalization is recommended when attempting PCA on measurements that may not be comparable to each other; range magnitudes may artificially weight components. Since the log ratios included in the analysis are comparable, no further preprocessing was necessary.

To compute the principal components, the *n* eigenvalues and their corresponding eigenvectors are calculated from the *n*x*n* covariance matrix of conditions. Each eigenvector defines a principal component. A component can be viewed as a weighted sum of the conditions, where the coefficients of the eigenvectors are the weights. Each of the *n* components can be calculated for a given gene :

$$a'_{ij} = \sum_{t=1}^{n} a_{it} v_{tj}$$

Where v_{tj} is the t^{th} coefficient for the j^{th} principal component; a_{it} is the expression measurement for gene *i* under the t^{th} condition. *A'* is the data in terms of principal components. Since *V* is an orthonormal matrix, *A'* is a rotation of the data from the original space of observations to a new space with principal component axes.

The variance accounted for by each of the components is its associated eigenvalue; it is the variance of a component over all genes. Consequently, the eigenvectors with large eigenvalues are the ones that contain most of the information; eigenvectors with small eigenvalues are uninformative.

Determining *r*, the true dimensionality of the data, and eliminating noisy components is often *ad hoc* and many heuristics exist. Eliminating low variance components, while reducing noise, also discards some valuable information. We chose to use one criterion, in common use in multivariate statistics, that discards all

components accounting for less than (70/*n*)% of the overall variability. Inspection of the coefficients of the remaining components can suggest what the component is measuring. However, such methods contain a certain amount of inherent subjectivity (Everitt & Dunn 1992).

The **Matlab**™ software package (The MathWorks, Inc., Natick, MA) was used to conduct most of our calculations.

3 Results

The data for this analysis was obtained from a publicly accessible web site[1]. The data contains expression ratios for 6118 known or predicted genes from *Saccharomyces cerevisiae*. The data was collected by plating cells on nitrogen deficient medium and measuring expression for each gene at 7 different time points (0hrs, 0.5hr, 2hrs, 5hrs, 7hrs, 9hrs, 11.5hrs) during sporulation. Thus, the matrix to be analyzedhas 6118 rows of genes and 7 columns of condtions corresponding to each of the measured time points. Table 1 reports the mean, median, and variance of each time point from the sporulation data. The means and medians are slightly negative but quite close to zero. Also note the relatively low variance of the t=0 time point; this is reassuring since the initial population should be similar to the background population.

<< TABLE 1 >> <<TABLE 2>>

Our analysis of the sporulation data series indicates that we can summarize the data with just two variables. Table 2 contains all 7 principal components and their corresponding eigenvalues. Figure 1 is a plot of the eigenvalues of the components. Two eigenvalues lie above the 10% (70/7) cutoff, suggesting two dimensions for the sporulation data. The first two principal components account for over 90% of the total variability; including the third component accounts for almost 95%. We include discussion of the third component for the sake of completeness. The meaning of these components can be distilled from their respective coefficients.

<< FIGURE 1>>

The first component represents a weighted average and distinguishes genes by their average overall expression. Ignoring the t=0 coefficient (it has negligible magnitude), it can be seen in Figure 2A that the remaining coefficients are positive (see also Table 2). The coefficients are proportional to the variance of the time points they are associated with (correlation = 0.97). The first component is an average expression weighted by the information content (i.e. variance) of a particular experiment. Genes with highly positive values along this component are

[1] http://cmgm.stanford.edu/pbrown/sporulation/index.html

up-regulated during sporulation, whereas genes with highly negative values are down-regulated.

<< FIGURE 2 >>

The second component represents change in expression over time; it distinguishes genes by their first derivatives. In Figure 2B the coefficients linearly increase with time from negative to positive values. Again, the exception to the rule is the low variance t=0 observation which has a negligible coefficient. Consider a gene *i* that is repressed (negative log expression ratio) in the early time points and highly induced (positive log expression ratio) in the final time points. The coefficient multiplied by the log expression score will be positive for each time point. Gene *i*'s value along the second component, a'_{i2}, is large and positive since every product in the sum is positive. Alternatively the second component for a gene that is induced early and repressed later will be large and negative. The expression scores are multiplied with coefficients of the opposite sign, yielding a large negative score. This component is positive for genes whose relative expression increases through time, and negative for those whose relative expression decreases; it measures positive trend in expression.

The third component measures concavity—notice the parabolic nature of the coefficients in Figure 2C (again ignore the negligible t=0 coefficient). Consider a gene *i* that is expressed at background level in the early and middle time points, but induced in the final time points—it has an expression profile that is concave up. Since the only non-zero expression levels occur at the final time points, only the later negative coefficients contribute to the sum, a'_{i3}. Consequently this gene will have a negative third component. Alternatively consider a gene with a similar profile, but that is expressed in the middle time points also (concave down); in this case the middle time points with positive coefficients increase the score along this component. The score of the second gene will be less negative.

In a sense, these vectors are decomposing a gene's expression pattern into Taylor series terms. The first component is the constant term, the second is the first derivative, and the third is the second derivative.

The first two components account for over 90% of the variance allowing most of the information to be visualized in two dimensions. All yeast genes are plotted in Figure 3 against the first two principal components; an ellipse enclosing 95% of the genes is drawn to distinguish between high and low variance genes. The genes appear to be distributed in a unimodal bivariate distribution. The data has been made available as a VMRL source at http://www.smi.stanford.edu/projects/helix/PCArray; the user can quickly navigate through two or three dimensional component space. Each data point is linked to its corresponding entry in the Saccharomyces Genome Database (Cherry et al. 1998).

<< FIGURE 3 >>

4 Discussion

Our results with the sporulation data indicate that PCA can be successful in finding a reduced set of variables that are useful for understanding the experiment. Since the data analyzed is a time series, it is reassuring that PCA identifies basic temporal patterns, such as magnitude, change, and the concavity of overall expression as the important features that characterize genes. Application of PCA to the publicly available cell division cycle data[2] reveals that PCA can also identify periodic patterns in time series data (Spellman et al. 1998). For example, this data reveals a 110 min period for the cdc15 synchronized experiment, consistent with the cell cycle duration.

The transcription factor NDT80 is key to the induction of many genes expressed in the middle of the sporulation process (Xu et al. 1995). The original dataset also includes measurements of gene expression for a NDT80 knockout microarray experiment and an ectopic NDT80 over-expression experiment. Including these extra experiments in the analysis results in coefficients for the first two components that are consistent with our understanding of the phenotype of these cells. In particular, the NDT80 knockout experiment traps cells in an early stage of sporulation; correspondingly, the coefficients in the first two components are most similar to the T=2 hour coefficients from the sporulation time series. In addition, the NDT80 over-expression data yields coefficients most similar to the T=11 hour coefficients. Since NDT80 is a sporulation promoting factor, the effects of over-expression may cause a phenotype that mimics a late time point.

Reduction of dimensionality in the sporulation data aids in data visualization; we can immediately see the unimodal quality of the sporulation data (Figure 3). The unimodal distribution of expression in the most informative two dimensions suggests the genes do not fall into well-defined clusters.

In the initial presentation of the data the investigators used clustering techniques to identify several gene classes relevant to sporulation: “metabolic”, “early I”, “early II”, “middle early”, “middle”, “middle late”, and “late” (Chu et al. 1998). For each class a canonical expression profile was calculated from a set of sample genes. These classes are plotted in Figure 4A; each ellipse in the plot represents a class. The location and dimensions of each ellipse was calculated from the average and standard deviation of the sample genes of the class. They are drawn so that approximately 68% (+/- 1SD in both dimensions) of the genes in the class

[2] http://genome-www.stanford.edu/cellcycle

are enclosed; in Figure 4B they are drawn to enclose 95% (+/- 1.96SD) of the genes in the class.

<< FIGURE 4 >>

An approximate understanding of a class's expression dynamic can be obtained quickly by looking at its location in space. For example, genes occupying the lower right quadrant (high PCA1, low PCA2) are up-regulated early but return to background later in sporulation. These genes have expression levels that decrease over time but maintain a high overall expression level relative to the control. Examples of these genes are ZIP1 (synaptonemal complex formation), IME2 (meiosis regulator), and HOP1 (homologous chromosome pairing), classified as "early I" or "metabolic" genes.

Exploring other quadrants can rapidly identify genes of potential interest. Genes with low overall expression levels that decrease over the course of sporulation can be found in the lower left quadrant. Many genes involved in metabolic or catabolic processes such as ERG6 (ergosterol synthesis), FBP1 (gluconeogenesis), and SAM2 (methionine biosynthesis) are found in this quadrant. Genes in the upper left are initially repressed and return to normal. Many of these genes are involved in protein synthesis. Examples include ISF1 (RNA splicing), BAP3 (valine transporter), and DBP3 (RNA helicase). The early repression may correspond with the cells' initial cessation of protein synthesis and growth; the renewed expression may function to pack the maturing spores with translation machinery (Chu et al. 1998). The reader is encouraged to further explore the genes in our visualization of the data.

Principal components analysis is often used as a preprocessing step to clustering (Everitt 1993). However, our work suggests that clustering genes with certain expression data sets may be inappropriate. In particular in Figure 4A the genes are not located in clusters - rather they are spread throughout this space. Focusing on the upper right quadrant in Figure 4B, it can be seen that the clusters presented in the original publication have a considerable amount of overlap. For unimodal or other smoothly varying distributions, distinctions drawn by clustering methodologies maybe more confusing than helpful. In particular, these clusters highlight the potential biases used in analyzing clusters using traditional cognitive categories. This observation corroborates the investigators' finding that the clusters are somewhat arbitrary; many genes were found to have high correlation with multiple cluster representatives (Chu et al. 1998). Perhaps it is more useful to ask what a particular gene's neighbors are rather than asking which cluster it is in.

Principal components analysis is a robust statistical technique that can be useful in analysis of microarray data. Techniques for dimensional reduction, such as PCA, offer promise for overcoming the difficulties in conceptualizing microarray data.

Acknowledgements

The authors wish to thank Raynee Chiang for her assistance in web development and visualization. S.R. is supported by NIH training grant GM-07365; J.M.S. is supported by NIH training grant LM-07033. This work was also supported by NIHLM06244, NSF DBI-9600637 and a grant from the Burroughs-Wellcome Foundation.

References

A. Basilevsky. Statistical Factor Analysis and Related Methods, Theory and Applications. 1994 John Wiley & Sons, New York, NY.

A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. "Predicting gene regulatory elements in silico on a genomic scale" *Genome Research* 8, 1202-1215 (1998)

M. Chee, R. Yang, E. Hubbel, A. Berno, X. C. Huang, D. Stern, J. Winkler, D.J. Lockhart, M.S. Morris,and S.P.A. Fodor. "Accessing Genetic Information with High-Density DNA Arrays" *Science* 274, 610-614 (1996)

J.J.W. Chen, R. Wu, P.-C. Yang, J.-Y. Huang, Y.-P. Sher, M.-H. Han, W.-C. Kao, P.-J. Lee, T.F. Chiu, F.Chang, Y.-W. Chu, C.-W. Wu, and K. Peck. "Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection" *Genomics* 51, 313-324 (1998)

J.M. Cherry, C. Adler, C. Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y. Jia, G. Juvik, T.Y. Roe, M. Schroeder, S. Weng, and D. Botstein. "SGD : Saccharomyces Genome Database" *Nucleic Acids Research* 26, 73-39 (1998)

S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz. "The transcriptional program of sporulation in budding yeast". *Science* 282, 699-705 (1998)

P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. "Linear modeling of mRNA expression levels during CNS development and injury" *Pacific Symposium on Biocomputing* 4, 41-52 (1999)

D.J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J.M. Trent. "Expression profiling using cDNA microarrays" *Nature Genetics* 21, 10-14, (1999)

M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. "Cluster analysis and display of genome-wide expression patterns" *Proc Natl Acad Sci U S A* 95, 14863-8 (1998)

B.S. Everitt. Cluster Analysis. 1993 John Wiley & Sons, New York, NY.

B.S. Everitt and G. Dunn. Applied Multivariate Data Analysis. 1992 Oxford University Press, New York, NY.

R.B. Franklin, D.R. Taylor, and A.L. Mills. "Characterization of microbial communities using randomly amplified polymorphic DNA" *J Microbiol Mtds* 35, 225-35 (1999)

S.H. Lai and M. Fang. "A novel local PCA-based method for detecting activation signals in fMRI" *Magn Reson Imaging* 17, 827-36 (1999)

E.S. Lander. "Array of hope" *Nature Genetics* 21, 3-4 (1999)

S. Liang, S. Fuhrman, R. Somogyi. "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures" *Pacific Symposium on Biocomputing* 3, 18-29 (1998)

G.S. Michaels, D.B. Carr, M. Askenazi, S. Fuhrman, X. Wen, and R. Somogyi. "Cluster analysis and data visualization of large-scale gene expression data" *Pacific Symposium on Biocomputing* 3, 42-53 (1998)

K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Phil. Mag.* 2, 559-572 (1901)

M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray" *Science* 270, 467-470 (1995)

P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Fucher. "Comprehensive Identification of Cell Cylce-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization." *Molecular Biology of the Cell* 9, 3273-3297 (1998)

L. Xu, M. Ajimura, R. Padmore, C. Klein, and N. Kleckner. "NDT80, a meiosis-specific gene required for exit from pachytene in Saccharomyces cerevisiae" *Mol Cell Biol* 15, 6572-6581 (1995)

Table 1. Summary of the experimental data collected by Chu and his colleagues (1998). The table contains average relative expression ratios after application of a natural log transform.

Time point	T=0	T=.5	T=2	T=5	T=7	T=9	T=11
Median	-0.122	-0.182	-0.104	-0.166	-0.095	-0.104	-0.131
Mean	-0.119	-0.214	-0.096	-0.119	-0.007	-0.032	-0.025
Variance	0.029	0.369	0.269	0.428	0.737	0.552	0.596

Table 2. Results of PCA on the sporulation time series data. The values in the columns are coefficients of the principal components that are related to each of the experimental time points. The eigenvalues express the variance of a principal component over all genes. Principal component 1 and 2 contain over 90% of the total variance in the data.

Projection On condition	Principal Components						
	1	2	3	4	5	6	7
T = 0	-0.0072	-0.0116	-0.0631	-0.2166	0.0764	-0.7433	0.625
T = .5	0.2076	-0.7524	-0.5373	0.2606	0.1545	-0.0683	-0.0756
T = 2	0.2358	-0.4925	0.3296	-0.5935	-0.453	0.1713	0.0803
T = 5	0.3975	-0.1156	0.5612	-0.002	0.5919	-0.2532	-0.3151
T = 7	0.554	0.0862	0.1869	0.4959	-0.1112	0.2889	0.5559
T = 9	0.4671	0.2517	-0.153	0.1169	-0.5413	-0.4488	-0.4324
T = 11	0.4671	0.3273	-0.4748	-0.5229	0.3307	0.254	0.044
Eigenvalue	2.2928	0.401	0.1322	0.0594	0.0406	0.0288	0.025
% variance	76.9 %	13.5 %	4.4 %	2.0 %	1.4 %	1.0 %	0.8 %

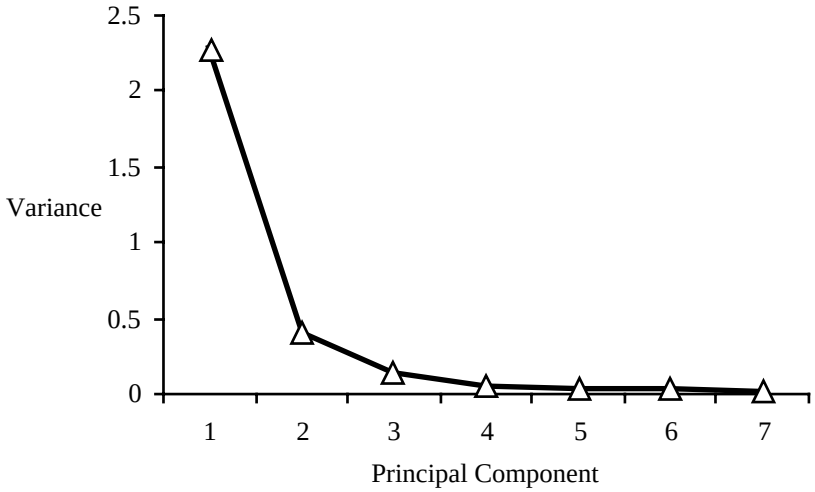

Figure 1. Plot of eigenvalues of the principal components. Most of the variance in the sporulation data set is contained in the first two principal components.

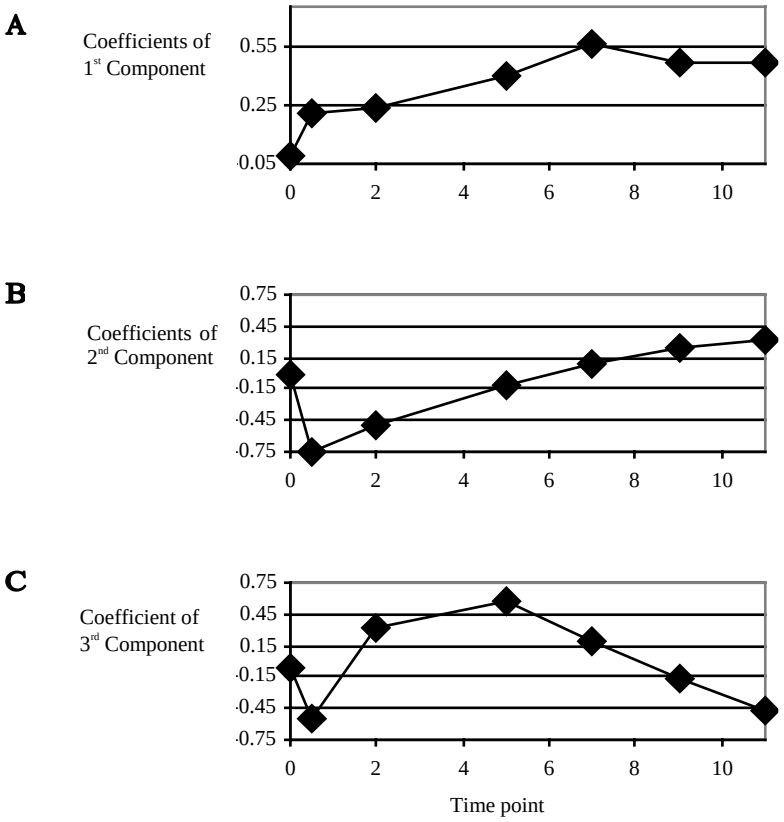

Figure 2. Plots of the coefficients of the first three principal components. Each coefficient indicates the weight of a particular experiment in the principal component. The first principal component has all positive coefficients, indicating a weighted average. The second principal component has negative values for the early time points and positive values for the latter time points, indicating a measure of change in expression. The third coefficient captures information about the concavity in the expression pattern over time.

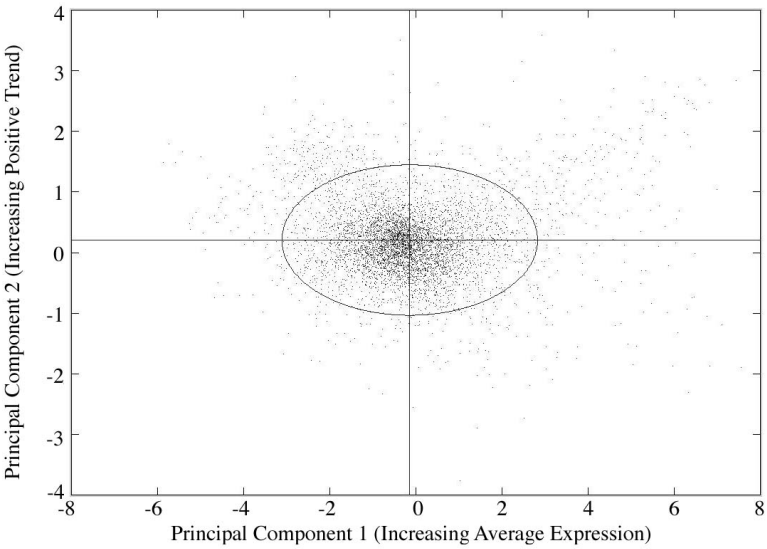

Figure 3. The rotated and dimensionally reduced expression data. All yeast genes are plotted on to the first and second principal components. The first principal component is a measure of total average expression, the second is a measure of increasing expression with respect to time. The ellipse at the center contains 95% of the genes.

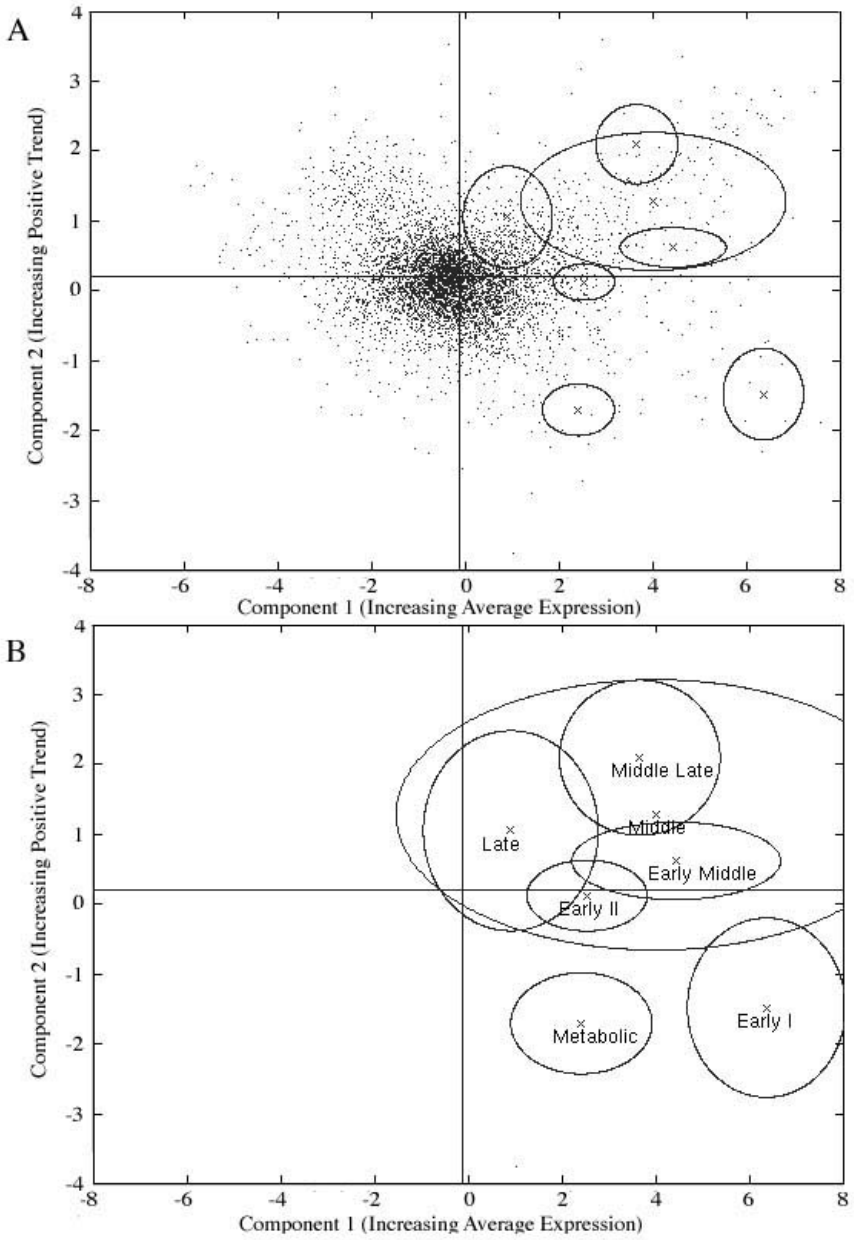

Figure 4. A. All genes plotted with respect to first and second principal components. Ellipses represent clusters identified in the original publication of the sporulation data. Ellipses are drawn to include 68% of the genes in the cluster. B. Ellipses are labelled using labels reported by the original investigators (Chu et al. 1998) and drawn to include 95% of genes in the cluster.