Singular Value Decomposition and Principal Component Analysis (PCA) I

Prof. Ned Wingreen

MOL 410/510

Microarray review

Data per array: $\sim 10\,000$ genes, $I_i^{(\text{green})}, I_i^{(\text{red})}$. With $X \sim 100$ arrays, this yields $\sim 1\,000\,000+$ data points!

The expression matrix has entries of the form $\log_2\left(\frac{I_{ij}^{(\text{green})}}{I_{ij}^{(\text{red})}}\right)$:

$$X = \begin{pmatrix} & & \vec{a}_j \\ & & \downarrow \\ \vec{g}_i & \rightarrow & \end{pmatrix},$$

where the i^{th} row $\vec{g}_i$ of the m rows is the transcriptional response of the gene i, and where the j^{th} column $\vec{a}_j$ of the n columns is the expression profile assay j. This is an $m \times n$ matrix, where the number of rows $m \sim 10\,000$ and the number of columns $n \sim 100$.

Example

Assays taken every 10 minutes after addition of some compound (nutrient, poison, signal molecule) — how to characterize the response of the genes? For any

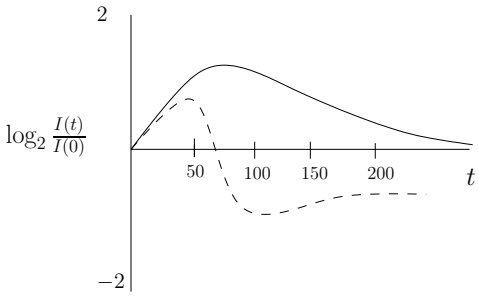

Figure 1: Noisy genes

individual gene, it may be impossible to see the signal through the noise. Individual genes may respond with a superposition of different patterns and noise. How to extract the important patterns from the noise? We will use SVD and

PCA to analyze the expression matrix X. But first we need to learn (or review) some linear algebra, so we can manipulate matrices like X.

Linear Algebra

Matrices

$$m \times n \text{ matrix } A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \text{ the "transpose" } A^T = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ a_{12} & & \\ \vdots & & \\ a_{1m} & & \end{pmatrix},$$

where the transpose interchanges columns and rows. If $m = n$, then A is called "square".

Matrix multiplication

If A is a $m \times n$ matrix and B is a $n \times l$ matrix, then AB is a $m \times l$ matrix where

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

We can think of this as dot products of vectors:

$$AB = \begin{pmatrix} \vec{a}_1 \rightarrow \\ \vec{a}_2 \rightarrow \\ \vdots \\ \vec{a}_m \rightarrow \end{pmatrix} \begin{pmatrix} \vec{b}_1 & \vec{b}_2 & \cdots & \vec{b}_l \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}$$

$$(AB)_{ij} = \vec{a}_i \cdot \vec{b}_j = \vec{a}^T \vec{b}.$$

Identity matrix

$$\text{square matrix } I = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & 0 & \\ & & 1 & & & & \\ & & & \ddots & & & \\ & & & & 1 & & \\ & 0 & & & & 1 & \\ & & & & & & 1 \end{pmatrix}.$$

I is the "identity matrix" because

$$AI = A$$
$$IA = A,$$

for all matrices A.

Inverse of a square matrix

$$AA^{-1} = A^{-1}A = I.$$

Note that $(A^{-1})^{-1} = A$. Finding the inverse of a matrix is slow in practice. In general:

$$A^{-1} = \frac{1}{|A|}(C_{ij})^T = \frac{1}{|A|}\begin{pmatrix} C_{11} & C_{21} & \cdots \\ C_{12} & \ddots & \\ \vdots & & C_{ji} \end{pmatrix},$$

where $|A| = \det A$. Therefore A^{-1} is $1/detA$ times the transpose of the cofactor matrix (aka the "adjoint of A").

$$C_{ij} = (-1)^{i+j}M_{ij},$$

where the minor M_{ij} is the determinant of the submatrix obtained from A by deleting row i and column j.

Determinants of square matrices

Example 2×2 matrix:

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc.$$

The determinant of larger matrices can be defined by induction. Example) 3×3 matrix:

$$\det\begin{pmatrix} i & j & k \\ a & b & c \\ d & e & f \end{pmatrix} = i(bf - ce) - j(af - cd) + k(ae - bd).$$

More generally,

$$|A| = \det A = \sum_{j=1}^{n} a_{ij}(-1)^{i+j}M_{ij},$$

where as above M_{ij} is the determinant of the submatrix formed by A without row i and column j:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & & (\text{submatrix "11"}) & \\ \vdots & & & \end{pmatrix}, \text{ so that } |A| = a_{11}M_{11} - a_{12}M_{12} + \ldots$$

Example: Inverse of a 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$A^{-1} = \frac{1}{|A|}\begin{pmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \end{pmatrix} = \frac{1}{|A|}\begin{pmatrix} M_{11} & -M_{21} \\ -M_{12} & M_{22} \end{pmatrix}$$
$$= \frac{1}{|A|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Check:

$$\begin{aligned} A^{-1}A &= \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ &= \frac{1}{ad-bc}\begin{pmatrix} ad-bc & 0 \\ 0 & ad-bc \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I \qquad \checkmark \end{aligned}$$

Linear independence

$$c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3 + \cdots + c_n\vec{v}_n = 0 \Leftrightarrow c_1 = c_2 = \cdots = c_n = 0.$$

If n vectors are not linearly independent, then they span a subspace of dimension $< n$. Eg. Two vectors $\vec{v}_1$ and $\vec{v}_2$ are linearly dependent if and only if they are multiples of each other: $\vec{v}_1 = a\vec{v}_2$. In this case $\vec{v}_1 + \vec{v}_2$ only span a 1 dimensional

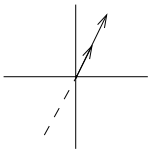

Figure 2: Linearly dependent vectors

subspace.

Orthogonality

Two vectors are orthogonal if $\vec{x} \cdot \vec{y} = 0$.

$$\vec{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \vec{y} = \begin{pmatrix} 4 \\ -2 \end{pmatrix}: \qquad \vec{x} \cdot \vec{y} = 1 \cdot 4 + 2(-2) = 0.$$

Orthogonal vectors are "perpendicular".

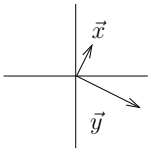

Figure 3: Orthogonal vectors

Eigenvalues and eigenvectors of symmetric matrices

"Eigen" means characteristic, so these are also sometimes called "characteristic" values and vectors. Eigenvalues and vectors are defined and related by the equation

$$A\vec{v} = \lambda\vec{v},$$

where $\vec{v}$ is an eigenvector and λ is an eigenvalue. (A consequence of this definition is that if $\vec{v}$ is an eigenvector, then so is $c\vec{v}$.)

To find eigenvalues, we use the theorem

$$(A - \lambda I)\vec{v} = 0 \Leftrightarrow \det(A - \lambda I) = 0.$$

This follows from a fact which is good to know: $\det X = \prod_i \lambda_i$. Similarly, $\operatorname{tr} X = \sum_i \lambda_i$.

We can therefore find eigenvalues by solving

$$\det(A - \lambda I) = \left| \begin{pmatrix} a_{11} - \lambda & a_{12} & & \cdots \\ a_{21} & a_{22} - \lambda & & \\ & \vdots & \ddots & \\ & & & a_{nn} - \lambda \end{pmatrix} \right| = 0,$$

which is an n^{th} order polynomial in λ, with n "roots", i.e. solutions for λ (these are either real or complex conjugate pairs).

Once a λ_i is known, $A\vec{v} = \lambda_i \vec{v}$ gives a set of linear equations which can be solved for the associated eigenvector $\vec{v}_i$.

Diagonalization

If A is an $n \times n$ matrix with n distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and associated eigenvectors $\vec{v}_i$, then let

$$U = \begin{pmatrix} \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ \downarrow & \downarrow & & \downarrow \end{pmatrix},$$

where the eigenvectors $\vec{v}$ are the columns of U. Now

$$\begin{aligned} AU &= A(\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n) \\ &= (A\vec{v}_1, A\vec{v}_2, \ldots, A\vec{v}_n) \\ &= (\lambda_1\vec{v}_1, \lambda_2\vec{v}_2, \ldots, \lambda_n\vec{v}_n) \\ &= U \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} = UD, \end{aligned}$$

where D is the diagonal matrix of eigenvalues. Manipulating the expression $AU = UD$ yields the identity

$$A = UDU^{-1}.$$

This is the eigen-decomposition theorem, and it is an example of a similarity transform.

When A is a symmetric matrix, i.e., when $A = A^T$, then it is possible to make the $\vec{v}$s orthonormal, in which case

$$U^{-1} = U^T,$$

so that

$$A = UDU^T.$$

(Principal Component Analysis (PCA) is equivalent to diagonalization of a particular symmetric matrix — the covariance matrix.)

The eigendecomposition allows some nice results:

$$A^2 = AA = (UDU^{-1})(UDU^{-1}) = UDDU^{-1} = UD^2U^{-1} = U \begin{pmatrix} \lambda_1^2 & & & \\ & \lambda_2^2 & & \\ & & \ddots & \\ & & & \lambda_n^2 \end{pmatrix} U^{-1},$$

and in the same way

$$A^m = UD^mU^{-1} = U \begin{pmatrix} \lambda_1^m & & & \\ & \lambda_2^m & & \\ & & \ddots & \\ & & & \lambda_n^m \end{pmatrix} U^{-1}.$$

So if we known the eigenvalues of A, we immediately know the eigenvalues of A^m. In particular,

$$A^{-1} = (UDU^{-1})^{-1} = (U^{-1})^{-1}D^{-1}U^{-1} = UD^{-1}U^{-1}.$$

It is easy to check that

$$A^{-1}A = (UD^{-1}U^{-1})(UDU^{-1}) = UD^{-1}DU^{-1} = UU^{-1} = I \qquad \checkmark$$

Therefore

$$A^{-1} = U \begin{pmatrix} 1/\lambda_1 & & & \\ & 1/\lambda_2 & & \\ & & \ddots & \\ & & & 1/\lambda_n \end{pmatrix} U^{-1}.$$

Singular Value Decomposition (SVD)

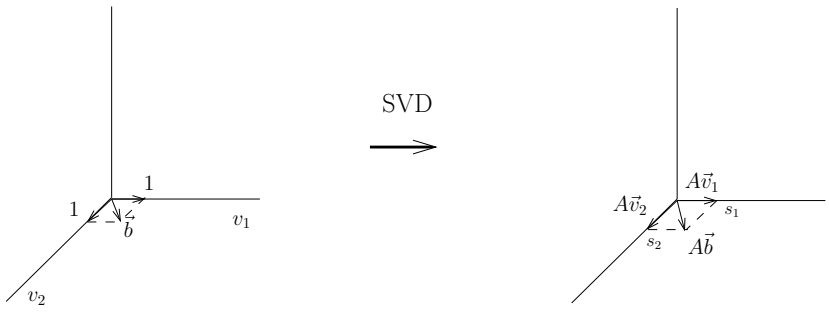

Figure 4: Singular Value Decomposition (SVD)

SVD is a generalization of diagonalization for non-symmetric matrices.
If A is an $m \times n$ matrix, then $\exists U, D, V$:

$$A = UDV^T,$$

where U is an $m \times m$ matrix with orthonormal columns and V is an $n \times n$ matrix with orthonormal rows. D is an $m \times n$ diagonal matrix, where

$$D = \begin{pmatrix} S_1 & & & \\ & S_2 & & \\ & & \ddots & \\ & & & S_n \end{pmatrix}, \qquad S_1 \geq S_2 \geq \cdots \geq S_n \geq 0.$$

These S_i are the singular values.

SVD and PCA II

Prof. Ned Wingreen

MOL 410/510

Let's see how we can first apply PCA and then SVD to extract information about gene regulation from a gene expression matrix. Recall

$$X = \begin{pmatrix} & \vec{a}_j & \\ & \downarrow & \\ \vec{g}_i & \rightarrow & \end{pmatrix},$$

where the i^{th} row $\vec{g}_i$ of the m rows is the transcriptional response of the gene i, and where the j^{th} column $\vec{a}_j$ of the n columns is the expression profile assay j.

Consider the expression profile for each assay. This will be the vector

$$\vec{a}_k = (x_{1k}, x_{2k}, \ldots, x_{mk}).$$

Each assay is therefore associated with a point in the m-dimensional space of genes. The graph, of course, shows only 2 dimensions of a m-dimensional space,

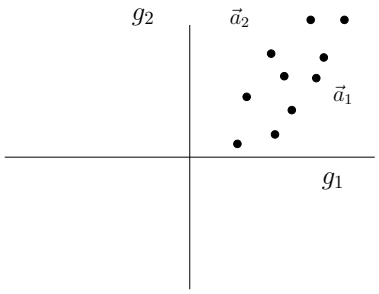

Figure 1: Gene space

but one can immediately see that there is a correlation between the expression of gene g_1 and g_2. How can we find correlations among multiple genes? E.g. how would we know if the expression of genes g_1, g_2, g_{37}, and g_{64} are all correlated? We can learn this from PCA and/or SVD.

First, center the data by subtracting out the mean value for each gene:

$$y_{ik} = x_{ik} - \overline{x}_i, \text{ where } \overline{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ik},$$

the average having been taken over the n assays. This yields the vector $\vec{a}'_k = \vec{a}_k - \overline{\vec{x}} = (y_{1k}, y_{2k}, \ldots, y_{mk})$.

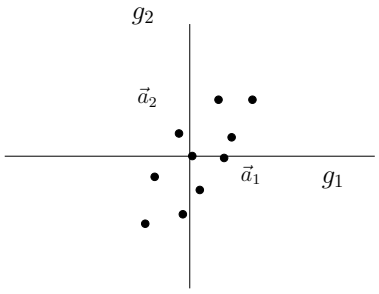

Figure 2: Re-centered gene space

Notice in this example that even after subtracting out the means, the expression values for genes 1 and 2 are correlated over the set of assays. Biologically, this suggests that genes 1 and 2 are coregulated. How can we quantify this? In particular how can we recognize if many sets of genes are coregulated (or counter-regulated)?

Graphically, we'd like to find the directions in gene space (weighed combinations of genes) that best capture the observed correlations in the data.

How do we do this mathematically, allowing for many correlated genes? Answer — these directions are the principal components of a particular matrix, the covariance matrix. With n assays, this takes the form:

$$\begin{aligned} C_{ij} &= \frac{1}{n}\sum_{k=1}^{n} y_{ik}y_{jk} = \frac{1}{n}\sum_{k=1}^{n}(x_{ik}-\overline{x}_i)(x_{jk}-\overline{x}_j) \\ &= E[(x_i-\overline{x}_i)(x_j-\overline{x}_j)] = E[x_ix_j] - \overline{x}_i\overline{x}_j \\ &= \mathrm{cov}(x_i,x_j). \end{aligned}$$

From the covariance matrix we can also define

$$\mathrm{cor}(x_i,x_j) = \frac{\mathrm{cov}(x_i,x_j)}{\sigma_i\sigma_j},$$

the "correlation coefficient" of x_i and x_j (often called r), which satisfies $-1 \leq r \leq 1$.. If x_i and x_j are independent,

$$\mathrm{cov}(x_i,x_j) = E[x_ix_j] - \overline{x}_i\overline{x}_j = \overline{x}_i\overline{x}_j - \overline{x}_i\overline{x}_j = 0.$$

Two other easy relations are

$$\begin{aligned} \mathrm{cov}(x_i,x_i) = E[(x_i-\overline{x}_i)^2] = \sigma_i^2 \qquad & (\mathrm{cor}(x_i,x_i) = r = 1) \\ \mathrm{cov}(x_i,-x_i) = -\sigma_i^2 \qquad & (\mathrm{cor}(x_i,-x_i) = r = -1). \end{aligned}$$

The covariance matrix is symmetric:

$$\mathrm{cov} = \begin{pmatrix} \mathrm{cov}(x_1,x_1) & \mathrm{cov}(x_2,x_1) & \cdots \\ \mathrm{cov}(x_1,x_2) & & \\ \vdots & & \end{pmatrix}.$$

Diagonalizing the covariance matrix is called a "Principal Component Analysis."

$$\text{cov} = UDU^T = \begin{pmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m \\ \downarrow & \downarrow & & \downarrow \end{pmatrix} \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{pmatrix} \begin{pmatrix} \vec{u}_1 & \rightarrow \\ \vec{u}_2 & \rightarrow \\ \vdots & \\ \vec{u}_m & \rightarrow \end{pmatrix},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. The $\vec{u}$s are called orthonormal eigenvectors called the "principal component vectors" and the λs, the eigenvalues of the covariance matrix, are called "principal component values."

The eigenvectors give the variance of the data along the principal component directions, and the covariance between these directions is zero. So $\vec{u}_1$ gives the direction in which the data is most "stretched" and λ_1 is the variance of the data in this direction.

Notice that

$$\text{Total variance of data} = \sum_i \text{cov}(x_i, x_i) = \text{tr cov} = \sum_k \lambda_k$$

where we've used the fact that the trace of a square matrix is the sum of its eigenvalues.

In terms of gene expression, $\vec{u}_1$ gives the weighted set of genes that are most strongly coregulated (or counter-regulated!), $\vec{u}_2$ gives the second strongest set, etc.

A plot of the eigenvalues λ_i is order from largest to smallest may reveal a transition from signal to noise:

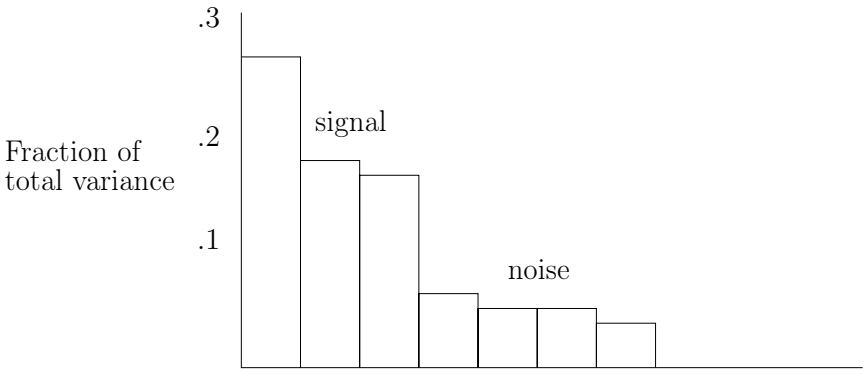

Figure 3: Eigenvalue plot

Dimension reduction

If the PCA separates signal from noise, we can "clean up" data by considering projection of data onto principal components. For each assay:

$$\vec{a}_k = (x_{1k}, x_{2k}, \ldots, x_{mk}) = \vec{a}'_k + \overline{\vec{x}},$$

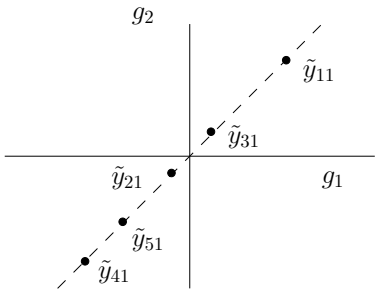

Figure 4: Values of $\tilde{y}$

we can find the projection of $\vec{a}'_k$ onto the principal components:

$$\begin{aligned}\tilde{y}_{k1} &= \vec{u}_1 \cdot \vec{a}'_k \\ \tilde{y}_{k2} &= \vec{u}_2 \cdot \vec{a}'_k \\ &\vdots\end{aligned}$$

Keeping only the first few values of $\tilde{y}$ for each assay accounts for most of the signal in the data.

Plotting the assay data projected onto the first few principal components can also reveal correlations beyond linear. Sometimes, $\tilde{y}_1$ and $\tilde{y}_2$ will show no additional correlations (Figure 5), but sometimes correlations will be apparent even though $\langle \tilde{y}_1, \tilde{y}_2 \rangle = 0$ (Figure 6).

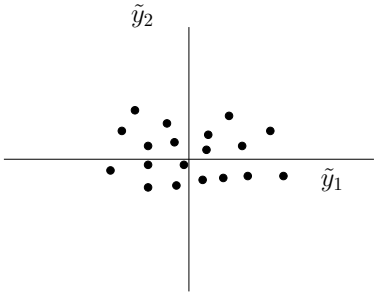

Figure 5: Uncorrelated

Other applications of PCA in biology:

- Molecular dynamics — reconstructing flexible modes of biomolecules from snapshots
- Synthetic lethality
- Immunology — antigen-antibody
- Residue-residue potentials
- Almost any large data set...

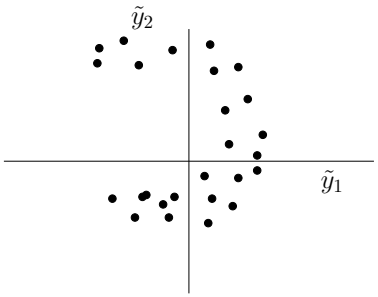

Figure 6: Correlated

SVD

By summing over assays to produce the covariance matrix, we have thrown away information. Imagine that the assays were taken at fixed time intervals — how can we recapture the time dependence of the gene expression?

Any one gene may have a noisy time course of expression:

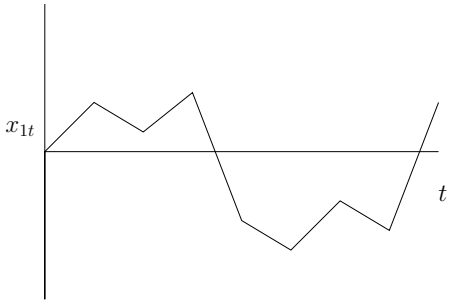

Figure 7: Single gene expression over time

But maybe there is a coherent response that contributes a little bit to many genes:

This coherent signal will contribute to the correlation of many genes, and so these genes will *co-vary*. So these genes will form a principal component in the PCA of the covariance matrix.

So after performing a PCA, imagine putting time labels back on the data: By finding the principal components and plotting the projections on these components for each assay vs. time, we reconstruct the coherent signal hidden in noisy data. SVD allows us to do this in one step:

$$m \times n \text{ matrix } X = UDV^T$$

$$= \begin{pmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m \\ \downarrow & \downarrow & & \downarrow \end{pmatrix} \begin{pmatrix} S_1 & & & \\ & S_2 & & \\ & & \ddots & \\ & & & S_n \end{pmatrix} \begin{pmatrix} \vec{v}_1 & \rightarrow \\ \vec{v}_2 & \rightarrow \\ \vdots & \\ \vec{v}_n & \rightarrow \end{pmatrix},$$

where $S_1 \geq S_2 \geq \cdots \geq S_n \geq 0$. In this decomposition, the $\vec{u}$s represent correlated sets of genes and the $\vec{v}$s represent coherent patterns of genes expression.

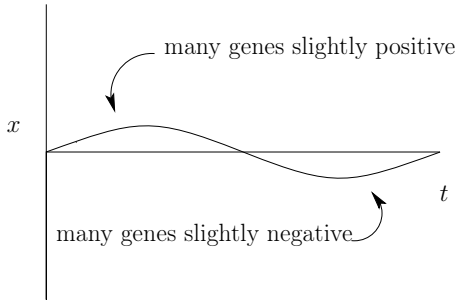

Figure 8: Overall x signal

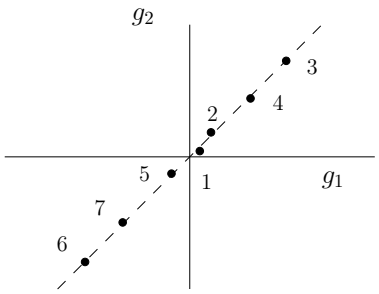

Figure 9: Relabeled data

The singular values S are simply related to the principal components of the covariance matrix:

$$\lambda_k = S_k^2.$$

The matrix constructed from the leading l singular values, and the associated right $\{\vec{v}\}$ and left $\{\vec{u}\}$ singular vectors is the best rank l approximation to X, that is, choosing

$$X^{(l)} = \sum_{k=1}^{l} \vec{u}_k S_k \vec{v}_k^T$$

minimizes

$$\sum_{i,j} |x_{ij} - x_{ij}^{(l)}|^2.$$

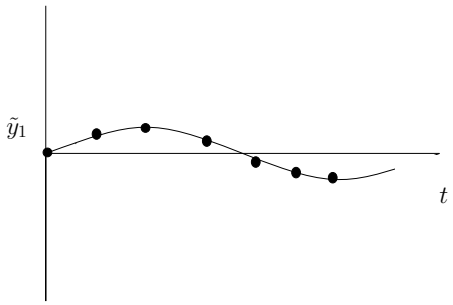

Figure 10: Explicit time plot