AN ITERATED VERSION OF THE SINGULAR VALUE DECOMPOSITION FOR THE JOINT ANALYSIS OF TWO HIGH-DIMENSIONAL DATA SETS

A Thesis Presented to The Faculty
of the Department of Mathematics
University of Hawaii, Manoa

In Partial Fulfillment of the Requirements
for the Degree Master of Arts

by Ashkan Zeinalzadeh

Date: August, 2013

ABSTRACT

In this work, we developed a new computational algorithm for the integrated analysis of high-dimensional data sets based on the Singular Value Decomposition(SVD). We developed an iterative version of the Singular Value Decomposition (ISVD) that jointly analyzes two data matrices to identify signals that correlate the rows of two matrices. The ISVD has been validated on simulated and real genomic data sets, and results show that the algorithm is able to sequentially detect multiple signals that were embedded in high levels of background noise. Results on real DNA microarray data from normal and tumor tissue samples indicate that the ISVD detects signals that are biologically relevant to the initiation and progression of liver cancer.

Table of Contents

CHAPTER 1

Introduction

Technology advances now enable the cost-effective acquisition of multiple types of large-scale, high-dimensional data (i.e., big data) from a common set of tissue samples. Examples of big data include global gene expression measurements, common and rare variants in the genome, sequenced RNA from whole tran-scriptomes, genome-wide methylation patterns, and image data based on Positron Emission Tomography (PET). The joint analysis of different types of big data should provide a deeper and more focused understanding of the molecular mechanisms underlying liver and other cancers that might not be apparent from analyzing one type of data in isolation [NHL11], [Alt07] and [BSe07]. Unfortunately, the size of the data sets and joint analysis of multiple data types has resulted in an "analysis bottleneck" that has slowed the discovery of genes and pathways associated with the image-based phenotypes. As the cost of acquiring big data plummets, we can expect to see an increased demand for the joint analysis of multiple data types, which should result in novel biomarkers and therapeutic targets, and deeper mechanistic insight into Hepatocellular Carcinoma (HCC) and liver disease.

Although correlation does not imply causation in high-throughput studies, identifying statistically significant correlations is an important first step in any data reduction step that attempts to transform high-dimensional data into a small set of variables that characterizes the underlying biology of cancer and its clinical outcomes. The joint analysis of two or more data matrices imposes constraints on the space of possible solutions that results in lists of biologically significant genes and pathways that are much more compact than those obtained from the analysis of any one data set alone. Such parsimony is more likely to lead to highly specific therapeutic targets and predictive signatures that are generalizable to the general population. The data obtained from the measurement of m variables in n samples can be represented as a $m \times n$ data matrix, A, composed of m rows and n columns of real numbers. Here, the ith row of A represents the n-dimensional profile of the ith variable over n samples, while the jth column of A represents the m-dimensional profile of the jth sample over m variables. For big data sets, the dimensionality of A is highly skewed where m is much greater than n (i.e., $m >> n$). Typically, the integrated analysis of two data matrices, A and B, means that the rows of each matrix are analyzed separately for correlations with a given phenotype of interest. The results from each analysis are then manually combined using known biology and other meta-level knowledge (e.g., genomic collocation, participation in a common signaling pathway) to identity a relatively small number of rows in each matrix that are mutually correlated. Although this multi-step approach achieves the primary goal of reducing the number of variables that must be considered for follow-up studies, it is labor intensive, error-prone, and sub-optimal from an analytical perspective.

1.1 Main Contribution

In this work we introduce an approach that involves the use of matrix factorizations such as the Singular Value Decomposition (SVD) to simultaneously analyze two matrices by iteratively indentifying cross-correlations between them. The Iterative Singular Value Decomposition (ISVD) jointly analyzes two data matrices by automatically indentifying any correlations that may exist between the rows of the two matrices, or the rows and columns of any one of the matrices. Since big data sets contain many distinct signals that associate the measured variables and samples, repeated application of the ISVD on re-normalized data iteratively detects these signals as orthogonal correlations between the two matrices or within a single matrix. This approach is computationally efficient, noise-resistant, allows for multivariate phenotypes, results in compact sets of significant variables, and can be shown to be near-optimal in quite general settings. The joint analysis of multiple data sets based on the ISVD algorithm could lead to the discovery of novel diagnostics and personalized treatment options for patients suffering from cancer that would likely be missed by analyzing individual data sets in isolation. Implementing ISVD for more than two data sets should be straightforward. Just stack more than two matrices and proceed as before.

Consider a $19,700 \times 24$ data matrix, A, of gene expression values and a 6×24 data matrix, B, of PET kinetic parameters obtained from a common set of 24 normal and tumor tissue samples that were jointly analyzed using the ISVD algorithm. The multi-dimensional PET phenotype represented by the rows of B was used to supervise the analysis of A and a statistically significant expression signature composed of 517 genes was identified that were highly correlated with PET tracer dynamics (please refer to Section 2.3.5 for statistically significant). Further analysis using Ingenuity Pathway Analysis (IPA) software identified 19 genes in the farnesoid X receptor/retinoid X receptor (FXR/RXR) signaling pathway. IPA is described in sequence.

IPA is a comprehensive library of the most current knowledge available in the literature on known genes, normal and disease processes and signaling and metabolic pathways. IPA profiling provided a means of evaluating different analysis methods based on their ability to produce gene sets enriched for biology associated with prior knowledge about the experimental conditions of a given study. IPA is a proprietary knowledge-base of genes, gene networks, signaling pathways, and associated biological functions extracted and manually curated from current and past publications in field of molecular biology. IPA is typically used to infer the biology underlying a small set of genes that characterize the changes in a clinical or phenotypic variable of interest (e.g., tumor versus normal).

Figure 1.1 shows a row-clustered heatmap of the 19 FXR/RXR genes with columns grouped by tumor status (i.e., normal vs. tumor), and rows annotated by gene symbols. Note that many of the 19 genes are known to play crucial roles in bile acid and lipid metabolism in the context of HCC. Note also that overall FXR/RXR expression was suppressed on a subset of HCC samples relative to the other normal and tumor samples, and this pattern was highly correlated with serum alpha feto-protein (AFP) levels suggesting a possible link between bile acid metabolism and HCC that could have diagnostic and/or prognostic value. The above example illustrates how the supervised analysis of a big data set (matrix A) by a complex, multi-dimensional phenotype (matrix B) using the ISVD algorithm could lead to novel diagnostics and personalized treatment options for patients suffering from HCC.

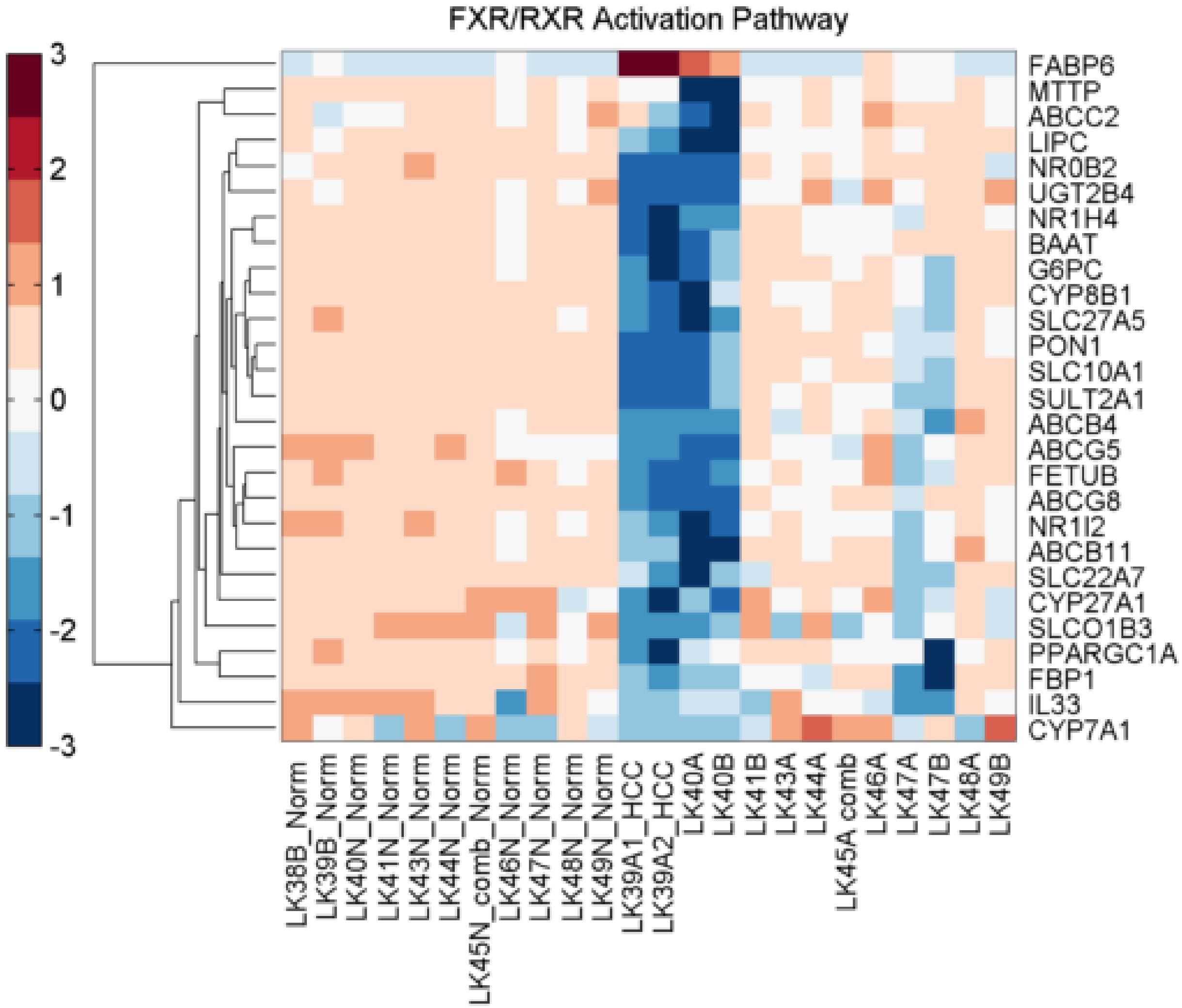

Fig. 1.1: Row-clustered heatmap of 19 genes in the FXR/RXR signaling pathway that were highly correlated with PET kinetic parameters based on ISVD analysis. The rows of the heat map represent genes in the FXR/RXR pathway clustered by similarity of expression profile. The columns represent samples grouped by normals (columns $1-11$), cholangiocarcinomas (columns $14-15$) and HCCs (columns $12-13$ and $16-24$). The colored entries of the heatmap represent normalized expression levels for a given gene in a specific sample. .

CHAPTER 2

Iterated Singular Value Decompositions

2.1 Singular Value Decomposition

The singular value decomposition of a real $m \times n$ matrix A is the decomposition of A into the product of three matrices of the form

$$A = UDV^T$$

where U is an $m \times m$ real unitary matrix, D is an $m \times n$ diagonal matrix with non-negative real entries, and V is an $n \times n$ real unitary matrix. The diagonal entries of D are called the singular values of A, and columns of U and V are called left and right singular vectors of A respectively. The left-singular vectors are eigenvectors of AA^T and right-singular vectors are eigenvectors of A^TA. Non-zero singular values of A are same as the square roots of non-zero eigenvalues of AA^T (which are the same as the non-zero eigenvalues of A^TA). Based on the SVD, the matrix A can be written as an outer product of orthogonal rank one approximations

$$A = \sum_{i=1}^{\text{rank}(A)} (u_i v_i^T) d_i$$

where u_i and v_i are the ith columns of U and V respectively; d_i is the ith diagonal element of D.

2.2 Generalized Singular Value Decomposition

The papers [Loa76], [PS81], and [Fri05] extended the SVD to two matrices $A_{m\times n}$ and $B_{p\times n}$. Note that A and B have the same number of columns, but the numbers of rows can be different. The columns of A can be viewed as vectors of m measurements in n samples based on a given technology platform, and similarly, the columns of B represent vectors of p measurements in the same set of n samples using a different technology. We define the joint analysis of the data pair A and B to mean the detection of rows of A and B that are highly correlated with a $n-$dimensional signal x that is defined on the columns of A and B. Let $q = \min(m + p, n)$. Presentation of GSVD is given as follows:

$$A = \underbrace{U}_{m\times m}\underbrace{C}_{m\times q}\underbrace{X^T}_{q\times n},$$

$$B = \underbrace{V}_{p\times p}\underbrace{S}_{p\times q}\underbrace{X^T}_{q\times n},$$

$$C^T C + S^T S = I,$$

where $U_{m\times m}$ and $V_{p\times p}$ are unitary matrices; C and S are nonnegative diagonal matrices and diagonal entries of C and S are non-increasing and non-decreasing respectively; X is an $n \times q$ matrix; I is an identity matrix. . If u_j, v_j, and x_j represent the jth columns of U, V, and X, then u_j and v_j can be viewed as correlation detectors of the signal x_j that is common to A and B. That is, we can detect the rows of A and B that are mutually correlated with the common signal x_j by applying thresholds to the column vectors u_j and v_j. Note that the above procedure will result in n distinct signals that link the rows of A and B via the common signals x_j.

Unfortunately, the GSVD is computationally demanding, which precludes the direct application of the GSVD to high-dimensional data matrices. Moreover, correlations of interest are typically sparse and embedded in enormous amounts of noise, which again precludes direct application of the GSVD to real world data. Hence, we developed the Iterative SVD (ISVD) which iteratively computes rank one approximations of the standard SVD that is both computationally tractable and robust to background noise and clutter. The SVD approximation is then used to detect correlations between A and B, after which, the data is orthogonalized with respect to the rank one approximation and the ISVD is applied again to the orthogonalized data. This iterative process is repeated with appropriate weighting and testing for statistical significance until the noise floor of the data is reached (please refer to Section 2.3.5 for statistically significant).

2.3 Iterative Singular Value Decomposition

Specifically, given the rank one approximation to the data we form the residual data and apply rank one approximation to the residual data pair. This procedure is iterated until the largest singular value of residue falls below a certain threshold that is near zero. This algorithm is based on the SVD of the stacked matrix $\begin{pmatrix} A \\ B \end{pmatrix}$.

$$SVD\begin{pmatrix} A \\ B \end{pmatrix} = \underbrace{P}_{(m+p)\times(m+p)} \; \underbrace{D}_{(m+p)\times n} \; \underbrace{Q^T}_{n\times n}$$

Define q_i (p_i) to be the ith column of matrix Q (P). Let $P(1:m,i)$ denote column i and rows from 1 to m of matrix P and $P(m+1:m+p,i)$ denote column i and rows from $m+1$ to $m+p$ of matrix P. The rank one approximation to SVD of $\begin{pmatrix} A \\ B \end{pmatrix}$ is given as

$$A_1 = \underbrace{P(1:m,1)}_{P_{11}} d_1 q_1^T,$$

$$B_1 = \underbrace{P(m+1:m+p,1)}_{P_{21}} d_1 q_1^T,$$

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \begin{pmatrix} P_{11} d_1 q_1^T \\ P_{21} d_1 q_1^T \end{pmatrix} = \begin{pmatrix} P_{11} \\ P_{21} \end{pmatrix} d_1 q_1^T = p_1 d_1 q_1^T,$$

$$A_1 = P_{11}/|P_{11}|.|P_{11}|.d_1 q_1^T = U_1.c_1.X_1^T,$$

$$B_1 = P_{21}/|P_{21}|.|P_{21}|.d_1 q_1^T = V_1.s_1.X_1^T,$$

with

$$c_1^2 + s_1^2 = 1,$$

$$X_1^T X_1 = d_1^2.$$

We are interested to find the support of signals in A and B. The vectors $U = \frac{P_{11}}{\|P_{11}\|}$ and $V = \frac{P_{21}}{\|P_{21}\|}$ in the rank one approximation of A and B can be viewed as detectors of the signal in A and B, respectively. That is, the components of U and V with large absolute magnitude correspond to the rows in A and B that are highly correlated with $X = d_1 q_1^T$. Note that X is common in the matrix factorization of A and B. The indices of U and V with large absolute magnitude are the signal support for A and B respectively. We apply the universal threshold proposed by Donoho and Johnstone to detect components of U and V with large absolute magnitude (Section 2.3.3).

Now we subtract A_1 and B_1 from A and B respectively and get the rank one approximation to SVD of the pair $\begin{pmatrix} A - A_1 \\ B - B_1 \end{pmatrix}$. It is shown in Appendix 4.1 that

$$SVD \begin{pmatrix} A - A_1 \\ B - B_1 \end{pmatrix} = \sum_{i=2}^{n} p_i d_i q_i^T.$$

The rank one approximation to SVD of $\begin{pmatrix} A - A_1 \\ B - B_1 \end{pmatrix}$ is given as A_2 and B_2 respectively:

$$A_2 = \underbrace{P(1:m,2)}_{P_{12}} d_2 q_2^T,$$

$$B_2 = \underbrace{P(m+1:m+p,2)}_{P_{22}} d_2 q_2^T,$$

$$A_2 = P_{12}/|P_{12}|.|P_{12}|.d_2 q_2^T = U_2.c_2.X_2^T,$$

$$B_1 = P_{22}/|P_{22}|.|P_{22}|.d_2 q_2^T = V_2.s_2.X_2^T,$$

with

$$c_2^2 + s_2^2 = 1,$$

$$X_2^T X_2 = d_2^2,$$

$$X_2^T X_1 = 0.$$

Similarly to the first iteration, the signals in $U = \frac{P_{12}}{\|P_{12}\|}$ and $V = \frac{P_{22}}{\|P_{22}\|}$ are detected. The sequential application of the ISVD is continued until the maximum singular value of

$$\begin{pmatrix} A - A_1 - A_2 - \cdots - A_k \\ B - B_1 - B_2 - \cdots - B_k \end{pmatrix}$$

drops below a predefined threshold indicating that the noise floor of the data has been reached. The number of iterations is at most $rank\begin{pmatrix} A \\ B \end{pmatrix}$. Therefore, the ISVD algorithm is computationally efficient and enhances signal detection by rank one approximation. To get the largest singular value and singular vector of A we can use the MATLAB command $eig(A^T A, 1)$. For further information about the algorithm of finding a few eigenvalue of a large matrix please refer to [LS96].

2.3.1 Residue in Iterative Singular Value Decomposition

If we iterate the preceding algorithm to the full rank of the stacked matrix $\begin{pmatrix} A \\ B \end{pmatrix}$, we obtain a decomposition

$$\begin{pmatrix} A \\ B \end{pmatrix} = \sum_{k=1}^{n} p_k \ d_k \ q_k^T.$$

Let us consider the Frobenius norm of the difference $\begin{pmatrix} A \\ B \end{pmatrix} - \begin{pmatrix} A_1 \\ B_1 \end{pmatrix}$. Recall that the Frobenius norm is given by

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} \mid a_{ij} \mid^2} = \sqrt{trace(A^T A)}.$$

Since

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = p_1 \ d_1 \ q_1^T,$$

we have

$$\|\begin{pmatrix} A \\ B \end{pmatrix} - \begin{pmatrix} A_1 \\ B_1 \end{pmatrix}\|_F^2 = trace\Big\{\Big(A^T - A_1^T \ \ B^T - B_1^T\Big)\begin{pmatrix} A - A_1 \\ B - B_1 \end{pmatrix}\Big\}$$

$$= trace\Big\{\Big(A^T - A_1^T \ \ B^T - B_1^T\Big)\begin{pmatrix} A - A_1 \\ B - B_1 \end{pmatrix}\Big\}$$

$$= trace\Big\{(A^T - A_1^T)(A - A_1)\Big\} + trace\Big\{(B^T - B_1^T)(B - B_1)\Big\}$$

$$= \|A - A_1\|_F^2 + \|B - B_1\|_F^2 = \sum_{k=2}^{n} d_k^2.$$

2.3.2 Convex Weighting of ISVD

The data matrices are first weighted using an appropriate convex linear parameter $0 < \lambda < 1$. Then, the ISVD algorithm is applied to the stacked matrix $\begin{pmatrix} \lambda A \\ (1-\lambda)B \end{pmatrix}$. Here, λ determines how A and B are jointly analyzed. The case $\lambda \sim 1$ implies that A supervises the analysis of B and the focus is on detecting significant variables in B correlated with X. The case $\lambda \sim 0$ implies B supervises the analysis of A and the focus is on detecting significant variables in

A correlated with X. $\lambda \sim 0.5$ implies both matrices equally supervise the analysis and the focus is on detecting signal support in both A and B.

The common X of A and B is synthesized from information in both A and B. Since X controls what signals are detected in the rows of both matrices, it would be nice to able to control the form of X by emphasizing A over B (i.e., X is influenced by A more than B) or by emphasizing B over A (i.e., X is influenced by B more than A). Convex weighting is one way of controlling what data matrix controls the form of X. Convex weighting is required if one wants to do a supervised analysis of one matrix by another. For example, if B is weighted heavily, then X will be based mostly on information in B. In particular, the analysis of the rows of A will be based on X, which is based mostly on B, hence the analysis of A is supervised by B. If A and B have very different number of rows, then the matrix with the smaller number of rows should be weighted heavier than the bigger matrix since the signal is less sparse in the smaller matrix.

2.3.3 Donoho-Johnstone Universal Threshold

The Donoho-Johnstone universal threshold for detectors U and V are given as $t_u = MAD_U\sqrt{2\log(|U|)}$ and $t_v = MAD_V\sqrt{2\log(|V|)}$ respectively. Here, $|U|$ and $|V|$ are the length of vectors U and V respectively, and MAD_U and MAD_V are the median absolute deviation (MAD) of U and V respectively. Note that MAD_U and MAD_V are different since their MAD estimates will differ. It is proved in [DJ94] and [Don95] that the maximum of any N iid samples with normal distribution $N(0;\sigma^2)$ will be smaller than $\sigma\sqrt{2\log(N)}$ with the probability approaching one as N increases. The large rows of U and V contribute the most to the $\begin{pmatrix} A_1 \\ B_1 \end{pmatrix}$ approximation, because the second part is the same X^T for both. The vector components of U and V with magnitudes that exceed their thresholds, correspond to the rows of A and B that are correlated. This threshold value is simple to implement and asymptotically optimal in minmax sense as the length of detectors approach infinity.

2.3.4 Flowchart for iterative SVD

The flowchart for iterative SVD is given in figure 2.1. We use Donoho-Johnstone universal threshold to detect the support signals in U and V ([DJ94], [Don95]).

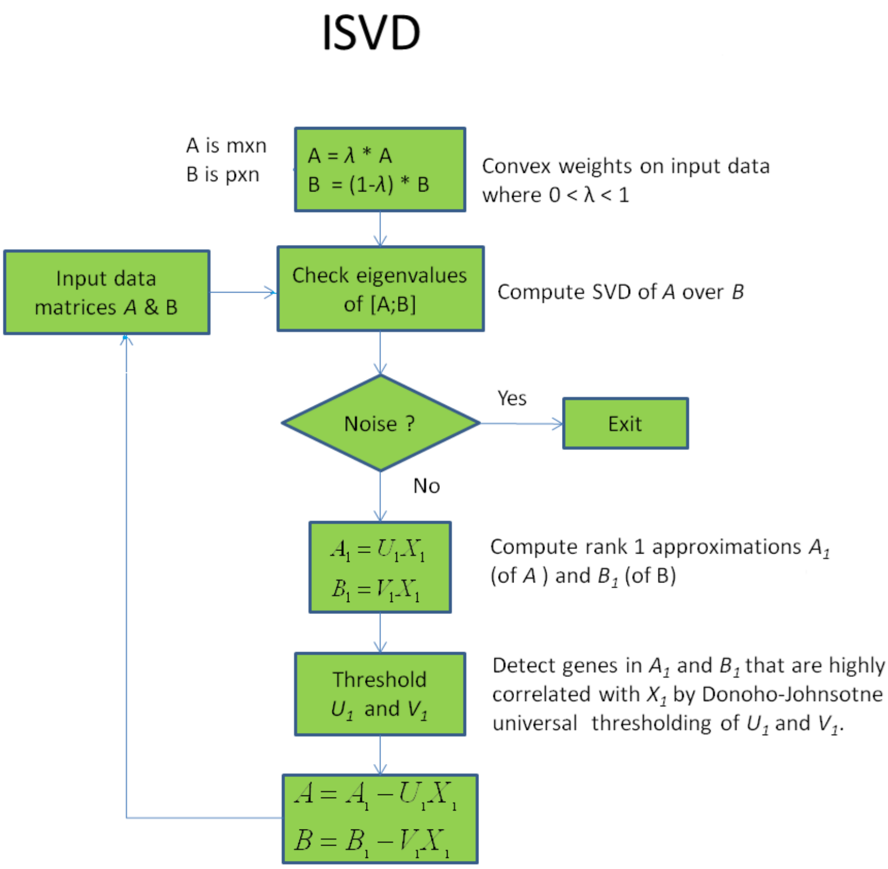

Fig. 2.1: Flowchart for Iterative SVD.

2.3.5 Statistically significance

When analyzing real-world data, each signal of cross-correlation (SCC) between A and B detected for a given iteration of the ISVD is tested for statistical significance using resampling methods. Basically, the rows of A and B are randomly permuted or bootstrapped to destroy the signal detected in the original data. The ISVD is then applied to the resampled data to obtain and store an estimate of the size of the support in both A and B for a resampled signal SCC. This process is repeated many thousands of times to form a distribution of support size for the resampled SCC. The detected signal from realworld data is called statistically significant if less than 5 percent of detected signals from resampled data have larger support size.

CHAPTER 3

Simulations and Conclusion

3.1 Simulation

3.1.1 Receiver Operating Characteristic (ROC) curve

Receiver Operating Characteristic (ROC) curves are used to assess the performance of signal detection algorithms [Met06]. We will use them here for the ISVD algorithm. Define

Sensitivity=Probability(Finding the true signal | There exists signal in data)

Specificity=Probability(Not finding the signal | There is no signal in data)

A ROC curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1–specificity) for a range of possible thresholds. An ROC curve shows different things:

- The tradeoff between sensitivity and specificity (any increase in sensitivity will be associated by a decrease in specificity).
- The closer the curve to a unit step function, the more accurate the signal detection.
- The closer the curve to the 45-degree line, the less accurate the signal detection.
- The slope of the tangent line at a point gives the likelihood ratio (LR) for that value of the threshold.

- The area under the curve is a measure of accuracy of signal detection.

3.1.2 Example

The ISVD algorithm has been tested and validated on both simulated and real data sets. In simulated data, the ISVD successfully detected a series of random and deterministic signals that were embedded in high-dimensional random Gaussian matrices. Three distinct signals were embedded in two 5000×50 random matrices A and B, where each signal was supported by randomly selected rows in each matrix. Signal 1 was a random signal supported by 100 randomly selected rows in A and B. Signal 2 was a step function with value -1 on columns $1 - 25$ and $+1$ on columns $26 - 50$ representing a signal for differential expression supported by 100 randomly selected rows in A and B. Finally, Signal 3 was a random signal supported by 100 randomly selected rows in A only. The other 4900 rows of A and B consisted of only random noise. The ISVD was iteratively applied to the data pair (A,B) with $\lambda = 0.5$. The detected signals in three iterations and analysis are given in figures 3.1, 3.2, 3.3, 3.4, 3.5, 3.6 and 3.7.

The Receiver Operator Curve (ROC) analyses shown in figures 3.8, 3.9 and 3.10 quantifies detection performance for the three simulated signals embedded in the data matrices A (blue ROC) and B (red ROC) over three iterations of the ISVD. For each ROC, we are interested in the Area Under the Curve (AUC), which is equal to the probability of correctly discriminating between a randomly selected row containing signal and a randomly selected noise-only row from a given matrix. We see that the ISVD detects only one signal, where for example, in the first iteration, we see the ROC in the top panel in figure 3.8 exhibiting a high AUC of 0.99 indicating that Signal 1 was successfully detected in iteration 1, while the middle and bottom panels corresponding Signals 2 and 3, respectively, have AUCs near 0.5 indicating that detection of these signals was no better than chance. In figure 3.9 and 3.10, we see

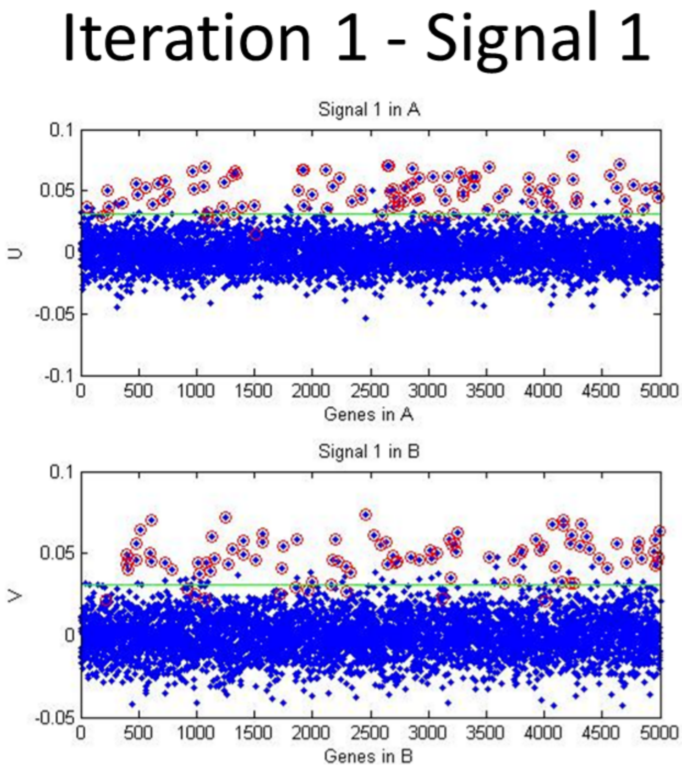

Fig. 3.1: Output of U and V detectors for iteration 1 with support for signal 1 superimposed in red. Top panel (Bottom panel): the most of the support rows of signal 1 in A (B) were detected since they exceeded the Donoho-Johnstone treshold (in green).

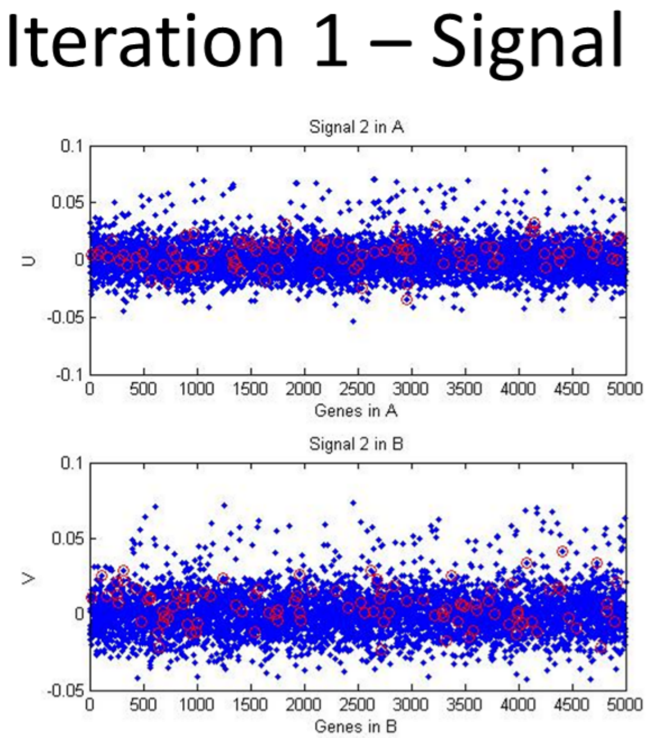

Fig. 3.2: Output of U and V detector output for iteration 1 with support of signal 2 highlighted in red . Note that all of the support of signal 2 in either A or B would NOT be detected since in each case, all signal 2 support are embedded deep in the noise floor. Hence, signal 2 is NOT detected on the 1st iteration of ISVD since it is NOT the strongest of the three simulated signals.

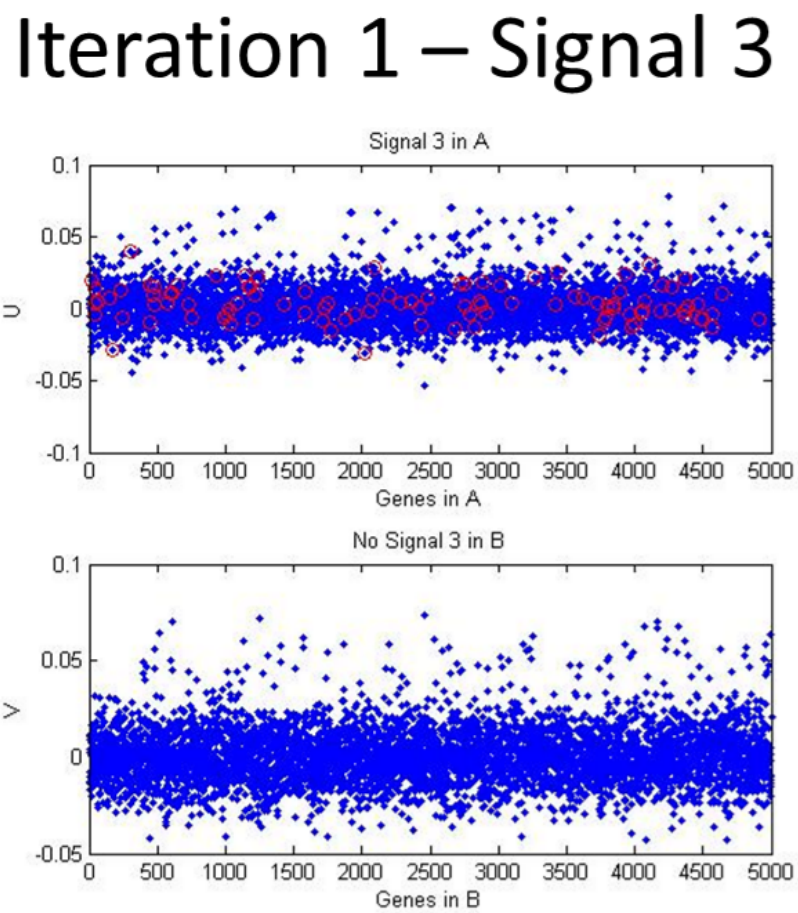

Fig. 3.3: Same U and V detectors for iteration 1 with signal 3 highlighted in red. Top Panel: The support of signal 3 are embedded deep in the noise floor of U, so none of these support would be detected by universal thresholding. Bottom Panel: Signal 3 was not embedded in B, so no support were highlighted in the output of V.

Iteration 2: Signal 1

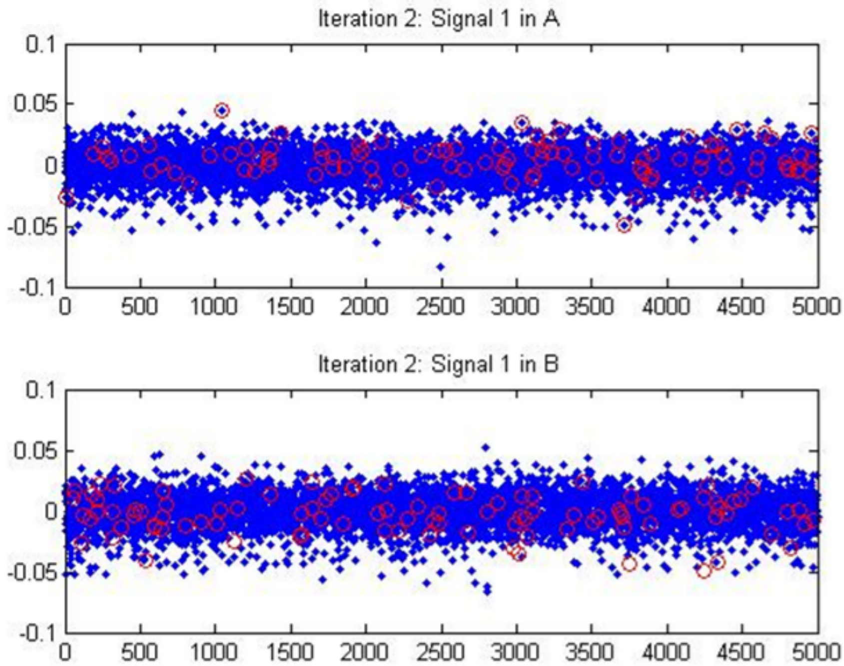

Fig. 3.4: Note that signal 1 (in red) was not detected in U or V since it was removed from the both A and B by the orthogonalization step of the IGSVD algorithm.

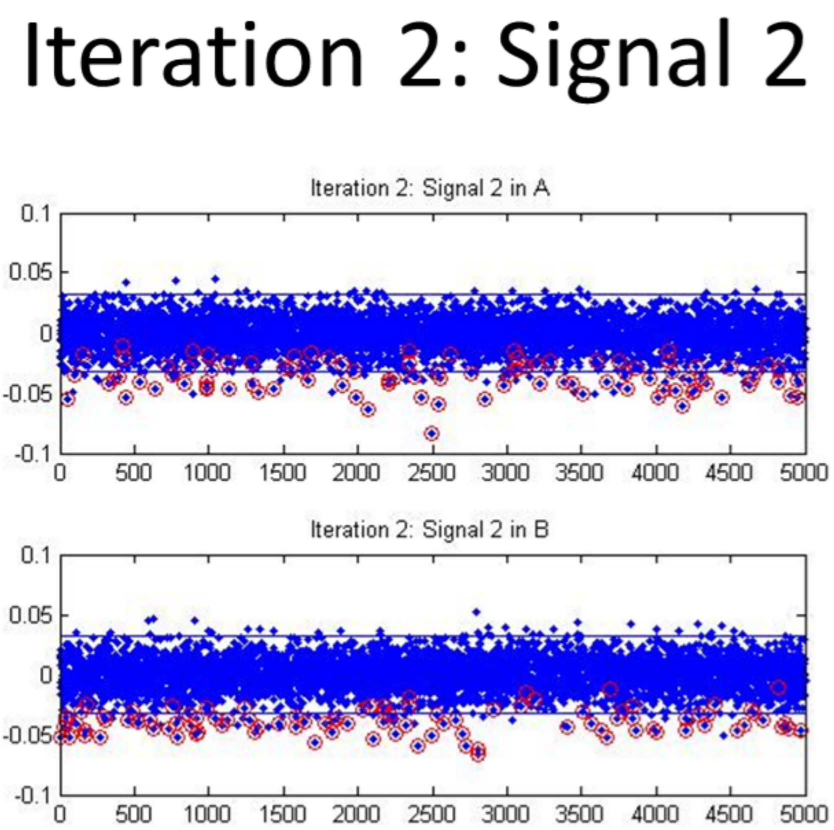

Fig. 3.5: U and V detector output for iteration 2 with signal 2 superimposed in red. Note that signal 2 was clearly detected in U and V based on the universal thresholds for each detector (blue horizontal lines).

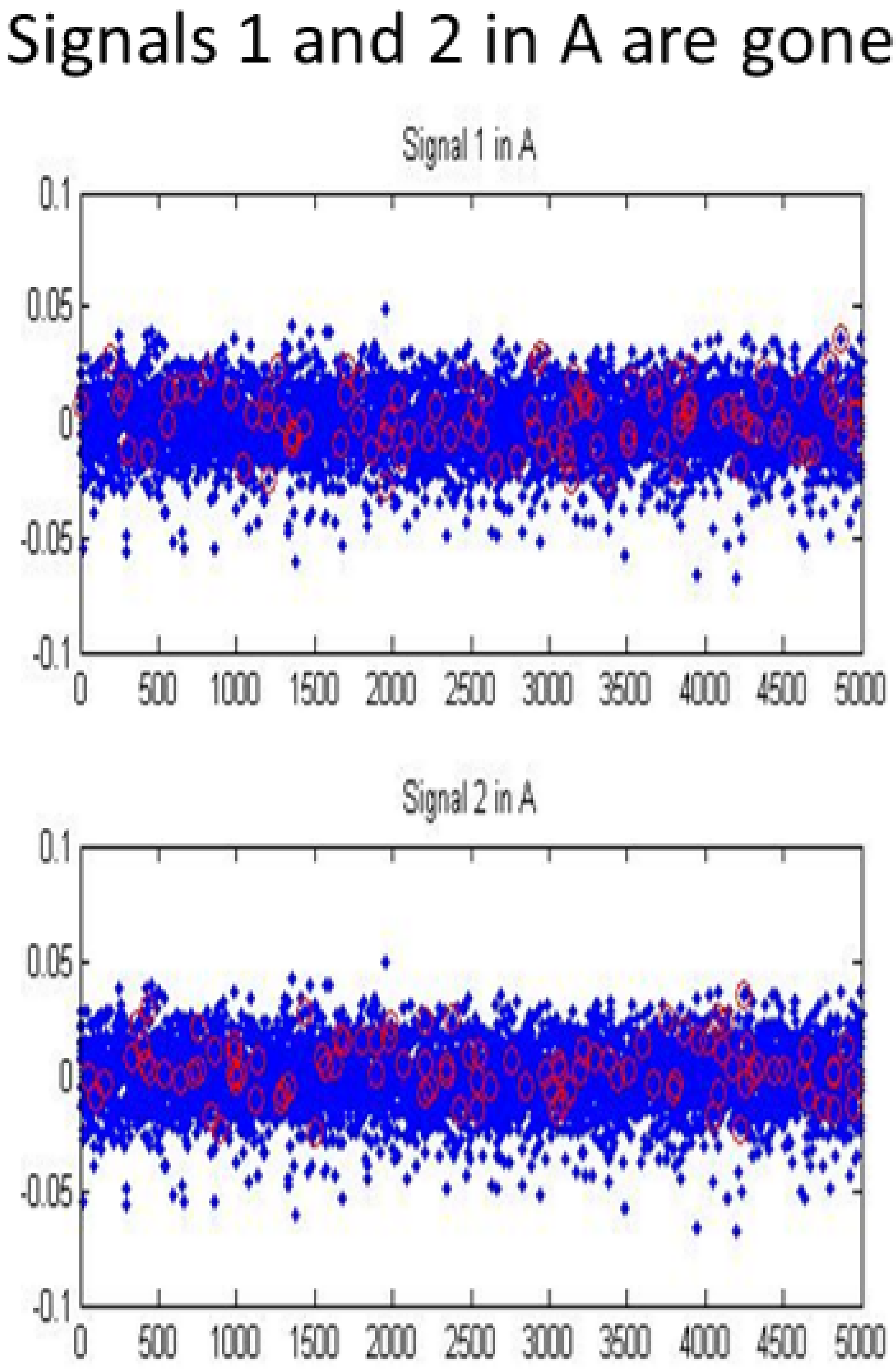

Fig. 3.6: U detector out for iteration 3 with support for signals 1 (top panel) and 3 (lower panel) superimposed. Note in each case, the signals would NOT be detected by universal thresholding since their support are deeply embedded in the noise floor.

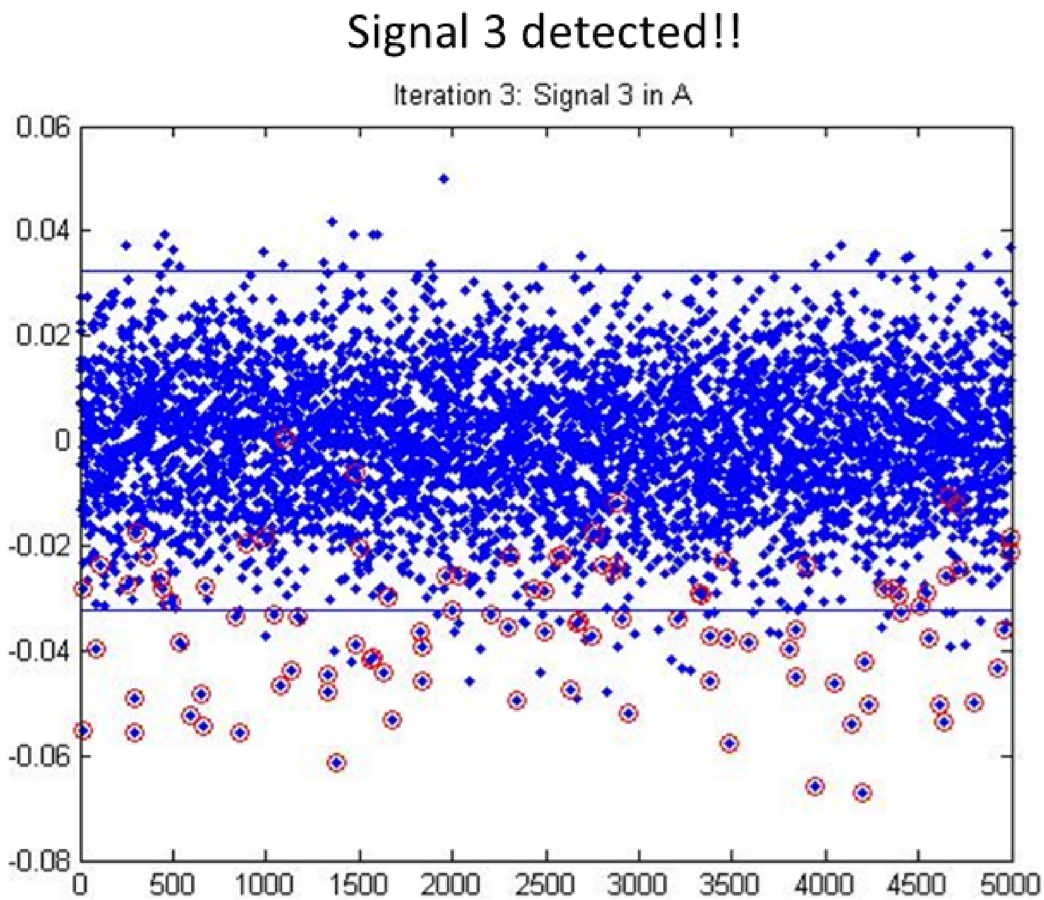

Fig. 3.7: U detector output for iteration 3 with signal 3 superimposed (in red). Note that most (but not all) of the support for signal 3 were detected in A.

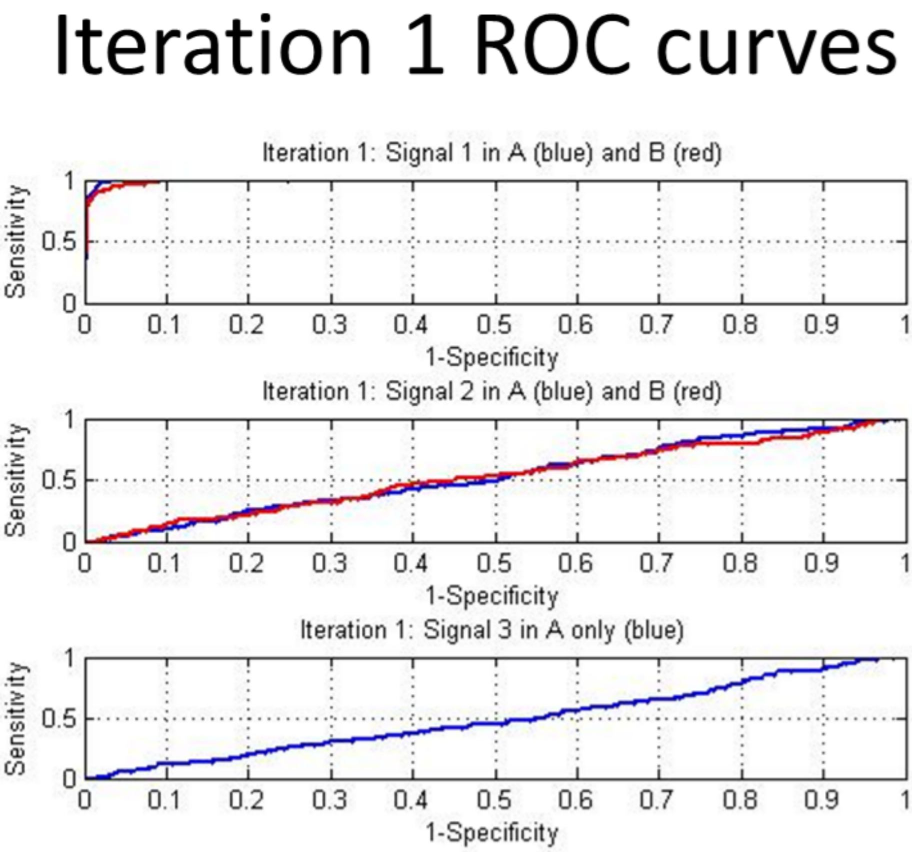

Fig. 3.8: ROC analysis confirms that the 1st iteration of IGSVD detected signal 1 in A and B, but not signals 2 and 3.

Iteration 2 ROCs

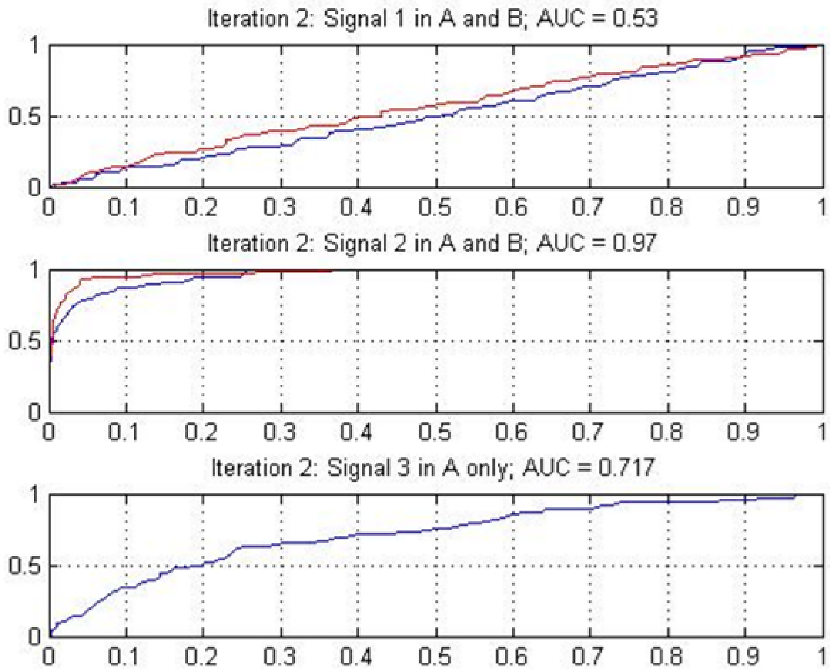

Fig. 3.9: ROC analysis confirms that the 2st iteration of IGSVD detected signal 2 in A and B, but not signals 1 and 3.

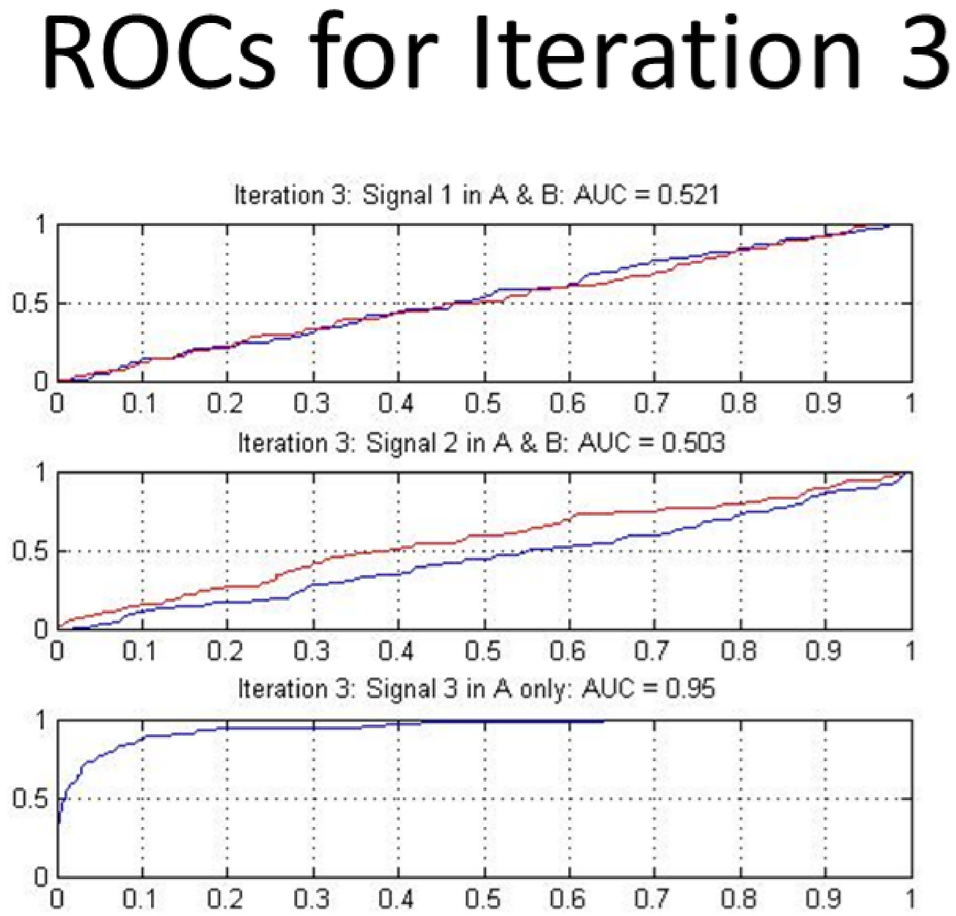

Fig. 3.10: ROC analysis confirms that the 3st iteration of IGSVD detected signal 3 in A and B, but not signals 1 and 2.

that the ISVD successively detected the step signal (AUC $= 0.97$) and the random signal unique to A (AUC $= 0.95$) in iterations 2 and 3, respectively. Figures 3.1, 3.5 and 3.7 plot the components of the U and V "detectors" of X in A (top panel) and B (bottom panel), respectively, obtained at each iteration of the ISVD. In figure 3.1 for example, we see most of the support for Signal 1 (red points) exceeds the threshold (in green) for both U and V, which represents a ROC operating point that is (asymptotically) optimal in a minmax sense for detecting Signal 1. Here, the blue points represent noise-only rows in A and B. Finally, in figure 3.7, we display a single scatter plot for the U detector for X in data matrix A, since Signal 3 is unique to A, only. Again, we see a majority of the support for Signal 3 exceeding the threshold. Note also that not all of the support for Signals 1, 2 and 3 are detected by the ISVD since no detector is perfect in the real world.

3.2 Real Data

As for real data, an example was presented in figure 1.1, that described the joint analysis of a $19,700 \times 24$ expression data matrix A and a 6×24 matrix of PET kinetic parameters extracted from a common set of 24 HCC and normal tissue samples. Here, λ was set to 0.01 so that the ISVD analysis of A was supervised by B since the number of genes greatly exceeded the number of PET kinetic parameters.

A second example involves the joint analysis of a $480,000 \times 46$ methylation data matrix A (Infinium $450K$ Methylation Array) and a $18,392 \times 46$ expression data matrix B (Whole-Genome DASL HT Assay) obtained from a set of 46 fresh liver tissue samples, of which 22 were HCC and 24 were adjacent-normal. The columns of A and B were grouped and aligned by the normal samples first (columns $1-24$) followed by the tumors (columns $25-46$). The first iteration of ISVD on $\begin{pmatrix} A \\ B \end{pmatrix}$ resulted in the correlated component where $4,911$ rows of A corresponding to methylation loci were highly correlated with 602 rows of B corresponding to genes. Panel A of figure 3.11 shows the common signal X shared by A and B, and clearly X is associated with tumor status because of its step-like profile over the tumor and normal samples. Note that ISVD found this signal without any supervision from an external data source other than what was contained in A and B. Panel B of figure 3.11 shows a heatmap of the expression data matrix B composed of the top 602 genes that were highly correlated with X. Functional profiling of these genes using IPA showed significant enrichment for genes, pathways and biological functions associated with HCC, liver disease and liver function.

3.3 Conclusion

The ISVD algorithm is able to detect signals that are common or unique to two data sets A and B. Signals are detected sequentially based on signal strength. Very weak signals with small singular values are detectable since noise background is systematically reduced relative to signal strength. Note that the ISVD is computationally feasible in situations where the GSVD is not. It may also be theoretically more meaningful than the GSVD. Based on the simulations result unlike ISVD, GSVD will not provide us distinct detectors for independent signals.

data2.png

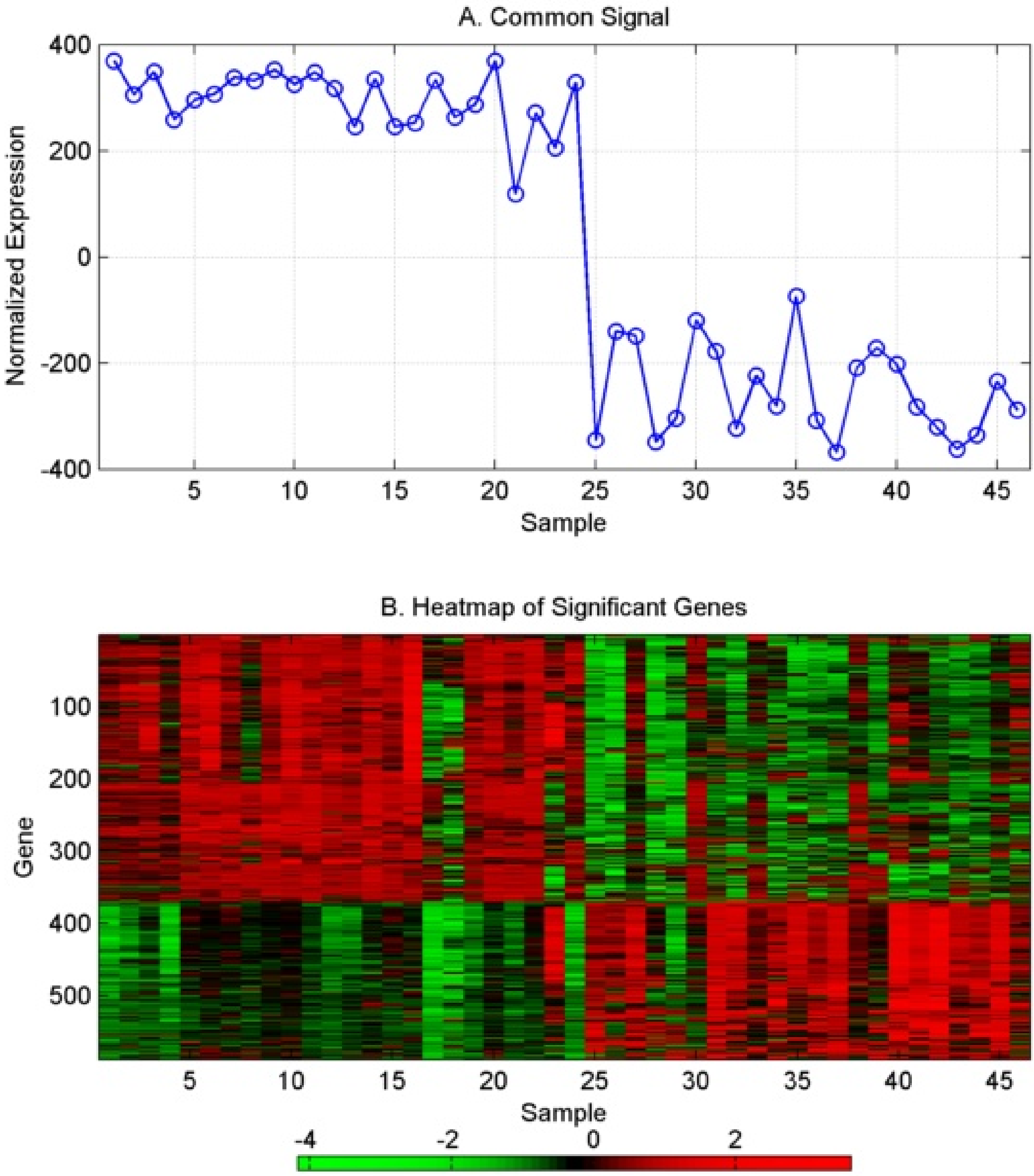

Fig. 3.11: ISVD analysis of whole-genome expression and methylation data from a set of 46 normal and HCC liver tissue samples. (A). Signal X shared by gene expression and methylation data. (B). Heatmap of 602 genes that were highly correlated with the signal X common to A and B.

CHAPTER 4

Appendix

4.1 Appendix

Lemma:

Let the SVD of the stack matrix be

$$G = \begin{bmatrix} A \\ B \end{bmatrix} = PDQ^T = \sum_{i=1}^{r} p_i d_i q_i^T.$$

Then, splitting apart the first column of P, we can write

$$\begin{aligned} A_1 = p_{11} d_1 q_1^T &= \frac{p_{11}}{\|p_{11}\|} \cdot \|p_{11}\| \cdot d_1 q_1^T \\ &= u_1 \cdot c_1 \cdot x_1^T \\ B_1 = p_{21} d_1 q_1^T &= \frac{p_{21}}{\|p_{21}\|} \cdot \|p_{21}\| \cdot d_1 q_1^T \\ &= v_1 \cdot s_1 \cdot x_1^T \end{aligned}$$

satisfying

$$\begin{aligned} c_1^2 + s_1^2 &= 1 \\ x_1^T x_1 &= d_1^2 \\ u_1^T u_1 = v_1^T v_1 &= 1. \end{aligned}$$

Let

$$H = \begin{bmatrix} A - A_1 \\ B - B_1 \end{bmatrix} = \sum_{i=2}^{r} p_i d_i q_i^T.$$

which latter equality holds since

$$\begin{bmatrix} A_1 \\ B_1 \end{bmatrix} = p_1 d_1 q_1^T.$$

The claim is that

- the nonzero eigenvalues of HH^T are $p_2, \ldots, p_r$,
- the nonzero eigenvalues of $H^T H$ are $q_2, \ldots, q_r$,
- with the corresponding eigenvalues $d_2^2, \ldots, d_r^2$,

so that the above representation of H is its singular value decomposition.

Proof: The proof is in the calculation

$$\begin{aligned} HH^T p_2 &= (\sum_{i=2}^{r} p_i d_i q_i^T)(\sum_{j=2}^{r} q_j d_j p_j^T) p_2 \\ &= (\sum_{i=2}^{r} p_i d_i q_i^T) q_2 d_2 \\ &= d_2^2 p_2 \end{aligned}$$

and similarly $H^T H q_2 = d_2^2 q_2$, etc.

The algorithm proceeds to find

$$
\begin{aligned}
A_2 = p_{12} d_2 q_2^T &= \frac{p_{12}}{\|p_{12}\|} \cdot \|p_{12}\| \cdot d_2 q_2^T \\
&= u_2 \cdot c_2 \cdot x_2^T \\
B_2 = p_{22} d_2 q_2^T &= \frac{p_{22}}{\|p_{22}\|} \cdot \|p_{22}\| \cdot d_2 q_2^T \\
&= v_2 \cdot s_2 \cdot x_2^T
\end{aligned}
$$

satisfying

$$
\begin{aligned}
c_2^2 + s_2^2 &= 1 \\
x_2^T x_2 &= d_2^2 \\
u_2^T u_2 = v_2^T v_2 &= 1 \\
x_1^T x_2 &= 0.
\end{aligned}
$$

The implementation and interpretation remain the same as before. □

Bibliography

[Alt07] O. Alter. Genomic signal processing: From matrix algebra to genetic networks. In *Microarray Data Analysis: Methods and Applications*, pages 17–58. Humana Press Inc, 2007. 1

[BSe07] M. Dunning B. Stranger, M. Forrest and etal. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315:848–853, 2007. 1

[DJ94] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika Trust*, 81(3):425–455, 1994. 12, 13

[Don95] David L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995. 12, 13

[Fri05] Shmuel Friedland. A new approach to generalized singular value decomposition. *SIAM J. Matrix Analysis Applications*, 27(2):434–444, 2005. 7

[Loa76] Charles F. Van Loan. Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, 13(1):76–83, 1976. 7

[LS96] R.B. Lehoucq and D.C. Sorensen. Deflation techniques for an implicitly re-started arnoldi iteration. *SIAM J. Matrix Analysis and Applications*, 17:789–821, 1996. 10

[Met06] Charles E. Metz. Receiver operating characteristic (roc) analysis: A tool for quantitative evaluation of observer performance and imaging systems. *J Am Coll Radiol*, 3:413–422, 2006. 15

[NHL11] P. Shah N. Huang and C. Li. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in Bioinformatics*, 13(3):305–336, 2011. 1

[PS81] C. C. Paige and M. A. Saunders. Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18(3):398–405, 1981. 7