NISS

Robust Singular Value Decomposition Analysis of Microarray Data

Li Liu, Douglas M. Hawkins, Sujoy Ghosh,
and S. Stanley Young

Technical Report Number 123
January, 2002

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Robust Singular Value Decomposition Analysis of Microarray Data

Li Liu[*], Douglas M. Hawkins[†], Sujoy Ghosh[‡], S. Stanley Young[‡]

[*]National Institute of Statistical Sciences, P.O. Box 14006, Research Triangle Park, NC 27709-4006; [†] School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street NE, Minneapolis, MN 55455; [‡] GlaxoSmithKline, Research Triangle Park, NC 27709

Abstract

In microarray data there are a number of biological samples, each assessed for the level of gene expression for a typically large number of genes. There is a need to examine this data with statistical techniques to help discern possible patterns in the data. Our method applies a combination of mathematical and statistical methods to progressively take the data set apart so that different aspects can be examined for both general patterns and for very specific effects. Unfortunately, these data tables are often corrupted with extreme values (outliers), missing values, and non-normal distributions that preclude standard analysis. We develop a robust analysis method to address these problems. The benefits of this robust analysis will be both the understanding of large-scale shifts in gene effects and the isolation of particular sample-by-gene effects that might be either unusual interactions or the result of experimental flaws. Our method requires a single pass, and does not resort to complex "cleaning" or imputation of the data table before analysis. We illustrate the method with a data set from the literature, where missing values, extreme values and non-normal distribution hamper standard methods.

Introduction

Biologists are using DNA microarrays to monitor the level of gene expression of biological samples. Thousands of genes are typically monitored on a few to tens of samples. In the near future, it is expected that there will be data sets of hundreds of samples. Patterns of gene expression can be used to determine co-regulated genes, suggest biomarkers of specific disease, and propose targets for drug intervention.

Microarray data present a number of challenges to statistical modeling. The size of the typical array – up to thousands of columns and perhaps hundreds of rows – defies easy graphical analyses. There may be severe distributional difficulties such as non-normal distributions, outliers (unusual data values), and numerous missing values. Common objectives are finding 'patterns' in the data, in particular

- clustering the biological samples (rows) into groups with similar expression profiles;
- clustering the genes (columns) into groups where the level of gene expression is similar in the samples.

One attractive way of clustering is a by-product of 'ordination.' Ordination involves finding suitable permutations of the rows (and perhaps of the columns) that lead to a steady

progression going down the rows (and perhaps across the columns). A clustering is given by placing vertical (and perhaps horizontal) dividing lines in the array to break it up into rectangular blocks within which the values are homogeneous.

Conversely, not all clustering methods are hierarchical, but if we cluster the rows and elect to do so with any hierarchical clustering method, the clustering induces an ordering of the rows ('dendrogram ordering'). Thus good ordination methods can lead to good clustering; any hierarchical clustering solution corresponds to a row ordination.

Method

The classical method of ordination is through the singular value decomposition. Write the expression data as an *nxp* array *X* with rows representing the *n* biological samples and columns representing the *p* genes. Approximate *X* with a bilinear form

$$x_{ij} = r_i c_j + e_{ij},$$

where r_i is a parameter corresponding to the i^{th} biological sample, c_j corresponds to the j^{th} gene, and e_{ij} is a 'residual'. This representation solves the ordination problem in that the rows can be ordered by their r_i values and the columns by their c_j values. Ordering the rows by *r* and the columns by *c* permutes the original data array to one in which we have high and low values in the corners and medium values in the middle, leading to an informative display.

Subsequently, grouping together those rows whose r_i are similar will give clusters of biological samples. Grouping the columns with similar c_j will give clusters of genes. To the extent that the residuals are small, the ordination and subsequent clustering will be unique.

Standard practice is to remove 'uninteresting structure' such as a grand mean, or even the row or column means from *X* prior to attempting the approximation. This is more of an implementation detail than a central aspect of the method.

The conventional method of getting this bilinear approximation is from the singular value decomposition (SVD) of *X,* Healy (1). It is well known that the leading term of the SVD provides the bilinear form with the best least squares approximation to *X*. The SVD is often found by performing a principal component analysis on X'X and XX'.

The conventional SVD however has some serious deficiencies. First, being a least squares method, it is highly susceptible to outlier values in the array *X*. Such outliers are an accepted fact of life when dealing with microarray data, where a sprinkling of entries are found to be very large or small. Second, finding the SVD through a principal component analysis of *X'X* requires that all elements of *X* be observed. This goes counter to a second reality of microarray data, which is that missing values are a routine feature of the experimental data.

Alternating least squares There is a standard remedy for the second of these deficiencies – the Gabriel-Zamir alternating least squares algorithm (2). This begins with a tentative

estimate of the column factors c_j. These are used to provide a matching scaling for the rows. Regarding

$$x_{ij} = r_i c_j + e_{ij}$$

as a regression of the i^{th} row of X on the column factors identifies r_j as the coefficient of a no-intercept regression. Fitting this regression row by row using all non-empty cells then leads to an estimate of the row factors r_i. Then switching roles, we take the row factors r_i as given, and use regression of all non-empty cells in exactly the same way to calculate fresh estimates of the column factors c_j. Thus the r_i and c_j are determined without removal of entire rows or entire columns or resort to imputation.

Alternating robust fitting (ARF). The ALS method is effective in solving the missing information problem. But it does nothing about the sensitivity to outliers. Solving the outlier issue however can be done by a simple change in the regression method that lies at the heart of the Gabriel-Zamir algorithm. Instead of using ordinary least squares (OLS) to carry out the alternating regressions, we can use an outlier-resistant regression method such as L1 (refer to Hawkins et al. (3) for details), weighted L1 (refer to Croux et al. (4) for details), least trimmed squares (refer to Ukkelberg et al. (5) for details), or more generally an M-estimation method. In this paper, we use the L1 method proposed by Hawkins et al. (3). The resulting algorithm then is to take the model

$$x_{ij} = r_i c_j + e_{ij,}$$

Use any convenient initial values for the column factors c_j (or optionally for the row factors) and then apply a robust no-intercept linear regression algorithm to alternately use the c_j to refine the estimates of the r_i; and the r_i to refine the estimates of the $c_{j.}$

Using any M-estimation criterion, each of these alternating regressions will lead to a reduction in the regression criterion, so the algorithm will converge.

Properties of the ARF fit. Broad properties of the ARF bilinear fit follow at once. The method handles missing information routinely, without requiring a separate 'fill-in' step. And it is impervious to a minority of outlier cells. Outliers will, of course, create a problem for the ARF, as with almost any conceivable method, if they constitute the majority of the elements of any row or column.

Clustering of the rows and columns. As already noted, sorting the rows by their r_i values will create a natural ordination. This can be turned into a clustering of k groups of biological samples by finding 'breakpoints' $b(0)=0 < b(1) < b(2)<b(k-1)<b(k)=n$ and allocating to cluster h those genes which, in the reordering, have index $b(h-1)<i<=b(h)$. The breakpoints need to be chosen so that the biological samples within each cluster have r_i values as similar as possible. This can be made operational by the criterion that the pooled sum of squared deviations of the r_i broken down into the k clusters should be a minimum. Exact algorithms for finding breakpoints to attain this minimum are given by Steel and Venter (6), and by Hawkins (7). Similarly, applying the optimal segmentation algorithm to the column factors c_j can cluster the genes.

Relationship to other clustering approaches

A common approach to clustering genes has been through (dis)similarity indices between the rows of X – for example the Euclidean distance between rows as a dissimilarity measure, or their correlation as a similarity measure. These measures can then be used in any convenient dissimilarity-based cluster method such as average linkage. If we look at this approach through the bilinear approximation

$$x_{ij} = r_i c_j + e_{ij}$$

(now regarded as an identity, without regard to quality of fit and trivially true for vectors r and c), we see that the squared Euclidean distance between any two rows i and k can be written

$$D_{ik}^2 = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2 = \sum_{j=1}^{p} (r_i c_j + e_{ij} - r_k c_j - e_{kj})^2$$
$$= (r_i - r_k)^2 \sum_{j=1}^{p} c_j^2 + \sum_{j=1}^{p} (e_{ij} - e_{jk})^2 + 2(r_i - r_k) \sum_{j=1}^{p} (r_i - r_k) c_j (e_{ij} - e_{kj})$$

Now, specialize the interpretation of this identity. Suppose that the bilinear term $r_i c_j$ has captured all the 'structure' in the sample-gene association, and all that is left is statistically independent measurement noise – not necessarily small – as is the case when the SVD is used to get the bilinear approximation. Consider the three terms comprising D_{ik}^2: the first is recoverable from the bilinear approximation and the last is zero by statistical independence; the center term is made up simply of measurement noise and it can not contribute usefully to the clustering. In fact it will have the effect of degrading the clustering. Theory therefore implies that a well-fitting bilinear approximation to the matrix X should give a better picture of the biological sample differences through its row factors r_i than can be found directly using the Euclidean distances between rows. A similar conclusion applies to using the correlation between pairs of row profiles.

There is yet another consideration favoring use of the ARF for clustering. If the matrix contains outliers, these outliers contaminate the Euclidean distances (and the correlation coefficients) between rows. However it is a consequence of the robust fits used in the ARF that the outliers do not contaminate the r_i or c_j substantially.

NCI60 data

We will illustrate the bilinear fit using the ARF and the subsequent segmentation of the genes and biological specimens using the well-studied NCI60 data set. The dataset is from the National Cancer Institute. The cDNA microarrays were used to assess gene expression profiles in 60 human cancer cell lines used in a drug discovery screen (8). The cell lines were derived from tumors with the following different sites of origin: breast (BR), central nervous system (CNS), colon (CO), leukemia (LE), melanoma (ME), non-small-cell-lung-carcinoma (LC), ovarian (OV), prostate (PR) and renal (RE). The reference sample was

prepared by pooling equal mixtures of mRNA from 12 of the cell lines. For each of the 60 cell lines, labelled cDNA was synthesized by reverse transcription from the cell mRNA with the red dye (CY5), and from the reference mRNA with the green dye (CY3). After competitive hybridization, laser scanner was used to measure the intensities of the red signals and the green signals, and the ratios were computed.

This data set can be downloaded from http://discover.nci.nih.gov/nature2000/. There are 1416 (1376 ESTs+ 40 known) genes, 60 cell lines, and each cell in the table represents the gene expression levels expressed as log (ratio), where the ratio is the red/green fluorescence ratio. There are about 2.4% missing values in the data. To give more details, all the rows (cell lines) have at least 1 missing value, and the median number of missing values is 13. As to the columns (genes), 850 out of 1416 genes have at least 1 missing values. It is important to handle the missing values properly.

Figure 1A shows the image of the unordered data matrix, where the rows correspond to the cell lines and the columns correspond to the genes, the white cells represent the missing values, the black cells represent the unchanged genes (log ratio =0), the green cells represent the under expressed genes (negative log ratio), the red cells represent the over expressed genes (positive log ratio), and the intensity of the color corresponds to the value of the log ratio. In this image, we cannot see any clear patterns.

Our goal is to cluster similar genes together and similar cell lines together, and at the same time we hope that the clusters will not be influenced by outliers, missing values or non-normal distribution of the noises in the data. Graphically, we hope to form blocks of reds and greens.

We used the ARF to fit a bilinear approximation to the activity matrix after subtracting out a grand mean. Using the resulting bilinear approximation to order the rows and the columns of X leads to the display of Figure 2. Note that visually, the ordination has been highly successful in rearranging the matrix so as to give blocks of high and low values in the corners and in-between values in the remainder of the array.

Next, we applied the segmentation algorithm to the row factors r_i to segment the biological samples, and to the column factors c_j to segment the genes. Tables 1 and 2 show the results of fitting various numbers of clusters. In an exploratory statistical analysis such as this, we do not need a rigorous answer to the question of the number of genuine clusters, but guidance comes from the variance explained by breaking the factors into 2, 3,…, groups. In both tables, the major explained variability is attained once three clusters are formed, and so we will use this as our working solution, dividing the columns (cell lines) into three segments, cell lines 1-28, 29-54 and 55-60. If we look at Figure 2A, we can see that the last six rows (cell lines 55-60) are indeed different from the rest, and the cell line names corresponding to the last six rows (cell lines 55-60) are the six 'LE' (leukemia). So the bilinear fit gives a good separation of leukemia from other cell lines. Similarly, based on Table 2, we can divide the genes into three segments, genes 1-462, 463-1096 and 1097-1416.

There is clearly an enrichment for leukemia cell lines (samples 55-60) in one of the cell line groups. In these leukemia cell lines, some genes are over expressed (shown in red) and some are under expressed (shown in green) as shown in Figure 2A. Among the genes found to be significantly over expressed in the leukemia lines are the transcription factor NFATx (gene 1398), a homolog of v-Myc (gene 1405), and cyclin dependent protein kinase (gene 1414). We now provide a summary of some of the literature for these genes.

NFAT(gene 1398): In the T-cell lymphoma cell lines EL4 and Jurkat, the NFAT proteins are reported to be activated by the oncogene Tpl-2 and may contribute to the molecular mechanism of the oncogenicity of Tpl-2 (9). Myc (gene 1405): In addition to the cell lines employed in this study, over expression of the myc oncogene was also noted in a multiple myeloma cell line, NCU-MM-1, and correlated with aggressive tumor growth of human TUR leukemia cells (10-11). Gene 1414: There is a significant amount of research on p27(Kip1), a cyclin-dependent kinase inhibitor, on the progression of hematological malignancies. Down regulation of p27 (functionally analogous to an up regulation of cyclin-dependent kinase, as seen in this study) has been shown to promote survival and cell cycle progression of T-cell acute lymphoblastic leukemia cells (12). Other cyclin-dependent kinase regulators such as flavopiridol induce growth arrest and apoptosis in chronic B-cell leukemia lines (13). In a separate study, cyclin dependent kinase 4 along with cyclin D1 were found to be the most important prognostic factors for children with acute lymphoblastic leukemia (14).

A literature search on the most down regulated genes in leukemia cell lines (as observed in this study) does not produce as clear picture as for the up regulated genes. It appears that in many cases, the expression of the gene depends upon the type and stage of leukemia. Some examples include laminin (gene 50) and caveolin (gene 55), covered next.

Lamin (gene 50): Indirectly, expression of the 67-kDa laminin receptor in acute myeloid leukemia cells mediates adhesion to laminin and is frequently associated with monocytic differentiation; thus loss of lamin or lamin receptor would inhibit differentiation and possibly maintain the transformed phenotype, (15). Caveolin (gene 55): Caveolin proteins are not detected in peripheral blood cells or blood cell lines which is consistent with down regulation in leukemic cell lines in this study. It is only in certain states of cell activation that caveolin-1 has been reported to be expressed in adult T cell leukemia cell lines (16).

Finding additional structure

The bilinear fit produced by the ARF does not necessarily exhaust all structure in the matrix X. As with the conventional, non-robust least-squares SVD, we can remove the initial bilinear fit from X to get the initial residual matrix ($x_{ij} - r_i c_j$) and apply the ARF to this matrix to get a second pair of matching row and column factors which may be segmented, just as were the leading pair.

Doing so does indeed uncover additional biologically meaningful structure, as shown in Figure 2B. The segmentation algorithm (refer to Table 4), suggests two segments for the cell lines: 1-51, and 52-60. The cell line names of the last 9 rows, show that they are the

seven 'ME' (melanoma) and two 'BR-MDA' (breast) cell lines. It is interesting that the two breast cell lines are grouped together with all the melanoma. It was pointed out (8) that the two breast cell lines (MDA-MB435 and MDA-N) were derived from a single patient with breast cancer, and have been treated as breast cancer cell lines. But it is likely that the patient had a co-existing occult melanoma. Corresponding to this cell line segmentation, we divide the genes into three segments, genes 1-481, 482-1124, and 1125-1416.

Subtracting this second bilinear term and repeating the ARF gives a third component. Segmentation divides the cell lines into 3 segments, cell lines 1-10 (which includes all the colon cancers), 11-35 and 36-60 (which includes the CNS and renal cancers). Similarly, we can divide the genes into three segments, genes 1-345, 346-971 and 972-1416. Thus, we divide the whole matrix into 9 homogenous blocks.

As we can see, the three components give three sets of ordering of genes and cell lines, and represent different aspects of the gene expression data. This cannot be achieved by one single ordering of genes and samples.

Finding outliers, filling in estimates of missing values, and smoothing

A strength of our method is that it does not require complete information, and is not affected by a minority of outliers. Outliers can be identified automatically by looking at the final residuals after removal of the three structural components. A simple outlier model might be that most of the residuals follow a normal distribution, but that some small number are 'wild'. A probability plot shows that, rather than this simple two-bin model, the residuals follow a heavy-tailed distribution. If we so wish, we can flag those readings that seem particularly anomalous. For example, finding the median absolute deviation (MAD) of all the residuals, under a normal distribution values more than 6*MAD in absolute value, a cutoff equivalent to 4 standard deviations, should be extremely rare. In the actual data however some 2% are outside this range. This is a red-flag warning against the use of non-robust methods (17).

Enriching notation slightly, write r_{im} and c_{jm} for the row and column factors given in the m^{th} bilinear pair fitted. Then for any missing cell ij, we could predict the missing value by $\sum_m r_{im} c_{jm}$. Another possible use of the ARF fits is to replace the entire matrix by the rank-m approximation given by using this missing value fill-in for all cells for purposes of other statistical analyses or displays. The potential attraction of this approach is that it would largely remove the impact of outlier cells as well as avoiding gaps in the matrix.

Discussion

We have proposed an analysis based on a variant of the SVD that is largely impervious to outliers and missing information. This can be used for ordination and display of the microarray, and also for segmentation.

The microarray example illustrates the usefulness of this method, where the cell lines of the same origin are grouped together, and some genes found are confirmed by previous

literature. The outlier detection points out some possible outliers. They may be experimental mistakes, or specific gene actions that deserve further study.

Acknowledgements

We thank Alan Karr and Jerry Sacks for helpful discussions.

References

1. Healy, M. J. R. (1986) Matrices for Statisticians. Clarendon Press, Oxford, 64-66.
2. Gabriel, K. R., Zamir , S. (1979). Technometrics, 21, 489-498.
3. Hawkins, D. M., Liu, L., Young, S. S. (2001) NISS Technical Report 122 (www.niss.org/downloadabletechreports.html).
4. Croux, C., Filzmoser, P., Pison, G. and Rousseeum, P.J. (2002) Preprint ISRO-Universite libre de Bruxelles 1999/140, to appear in Statistics and Computing.
5. Ukkelberg, A and Borgen, O. (1993) Analytica Chimica Acta, 277, 489-494.
6. Venter, J. H. and Steel, S. J. (1996) Comput Stat and Data Anal, 22, 481-504.
7. Hawkins, D.M. (2000) Comp. Statistics and Data Anal., 37, 323-341.
8. Ross, D., Scherf, U., Eisen, M. B., Perou, C. M., Spellman, P., Iyer, V., Jeffrey, S.S., Rijn, M., Waltham, M., Pergamenschikov, A., et al. (2000) Nature Genetics, 24:227-234.
9. Tsatsanis, C., Patriotis, C., Tsichlis, P. N. (1998) Proc. Natl. Acad. Sci., 95(7), 3827-3832.
10. Iida, S., Hanamura, I., Suzuki, T., Kamiya, T., Kato, M., Hayami, Y., Miura, K., Harada, S., Tsuboi, K., Wakita, A., et al. (2000) International Journal of Hematology, 72(1), 85-91.
11. Hass, R and Lopez-Guerroro, J.A. (1997) International Journal of Cancer, 72(6), 1113-1116.
12. Barata, J.T., Cardoso, A.A., Nadler, L. M., Boussiotis, V.A. (2001) Blood, 98(5), 1524-1531.
13. Konig, A., Schwartz, G.K., Mohammad, R.M., Al-Katib, A., Gabrilove, J.L. (1997) Blood, 90(11), 4307-4312.
14. Volm, M., Koomagi, R., Stammler, G., Rittgen, W., Zintl, F., Sauerbrey, A. (1997) International Journal of Cancer, 74(5), 508-512.
15. Montouri, N., Selleri, C., Risitano, A.M., Raiola, A.M., Ragno, P., Del Vecchio, L., Rotoli, B., Rossi, G. (1999) Clinical Cancer Research, 5(6), 1465-1472.
16. Hatanaka, M., Maeda, T., Ikemoto, T., Mori, H., Seya, T., Shimizu, A.(1998) Biochem.Biophys.Res. Comm., 253(2), 382-387.
17. Huber, P. J. (1981) Robust statistics. Wiley, New York.

Table 1. The segmentation of cell lines in component 1

No.	Explained SS	Change	Break points
2	78.6103		54 60
3	92.4562	13.8459	28 54 60
4	96.1626	3.7064	25 52 55 60
5	98.4368	2.2742	24 51 54 58 60
6	100.3947	1.9579	15 33 51 54 58 60

Table 2. The segmentation of genes in component 1

No.	Explained SS	Change	Break points
2	72.1226		835 1416
3	84.1818	12.0592	462 1096 1416
4	88.6127	4.4309	277 758 1179 1416
5	90.7224	2.1097	211 582 988 1256 1416
6	92.0583	1.3359	188 516 865 1132 1333 1416

Table 3. Description of genes from the 1st and 2nd components.

Gene index	Place 1st comp	Description
7	1398	SID W 245450, Human transcription factor NFATx mR-99, complete cds
29	1405	MYC V-myc avian myelocytomatosis viral oncogene homolog Chr.8
76	1414	Human cyclin-dependent protein ki-99se mR-99, complete cds Chr.12
845	50	LAMC1 Laminin, gamma 1 (formerly LAMB2) Chr.1
874	55	SID 472015, Homo sapiens caveolin-2 mR-99, complete cds

Gene index	Place 2nd comp	Description
1315	1393	SID W 486432, Dopachrome tautomerase (dopachrome delta-isomerase, tyrosine- related protein)
1345	1402	H.sapiens mR-99 for Gal-beta(1-3/1-4)Glc-99c alpha-2.3-sialyltransferase Chr.11
1043	11	SID W 429859, Villin 2 (ezrin)
248	25	ANX3 Annexin III (lipocortin III) Chr.4
262	26	DESMOPLAKIN I AND II Chr.6

Table 4. The segmentation of cell lines in component 2

No. of segments	Explained Sum of Squares	Change	Break points
2	84.3986		51 60
3	88.5639	4.1653	28 51 60
4	89.7414	1.1775	14 39 51 60
5	90.6048	0.8634	14 39 51 55 60
6	91.1050	0.5002	6 28 44 51 55 60

Table 5. The segmentation of genes in component 2

No. of segments	Explained Sum of Squares	Change	Break points
2	59.3728		802 1416
3	75.4268	16.0540	481 1124 1416
4	81.6079	6.1811	304 779 1202 1416
5	84.6040	2.9961	227 602 994 1284 1416
6	86.4369	1.8329	148 420 779 1111 1310 1416

Table 6. The segmentation of cell lines in component 3

No. of segments	Explained Sum of Squares	Change	Break points
2	56.2300		19 60
3	76.1226	19.8926	10 35 60
4	79.5835	3.4609	10 34 55 60
5	82.0603	2.4768	9 19 36 56 60
6	83.6325	1.5722	2 10 19 36 56 60

Table 7. The segmentation of genes in component 3

No. of segments	Explained Sum of Squares	Change	Break points
2	65.5843		722 1416
3	81.0274	15.4431	345 971 1416
4	86.6599	5.6325	236 687 1098 1416
5	89.5089	2.8490	128 408 781 1133 1416
6	91.2666	1.7577	119 348 687 989 1261 1416

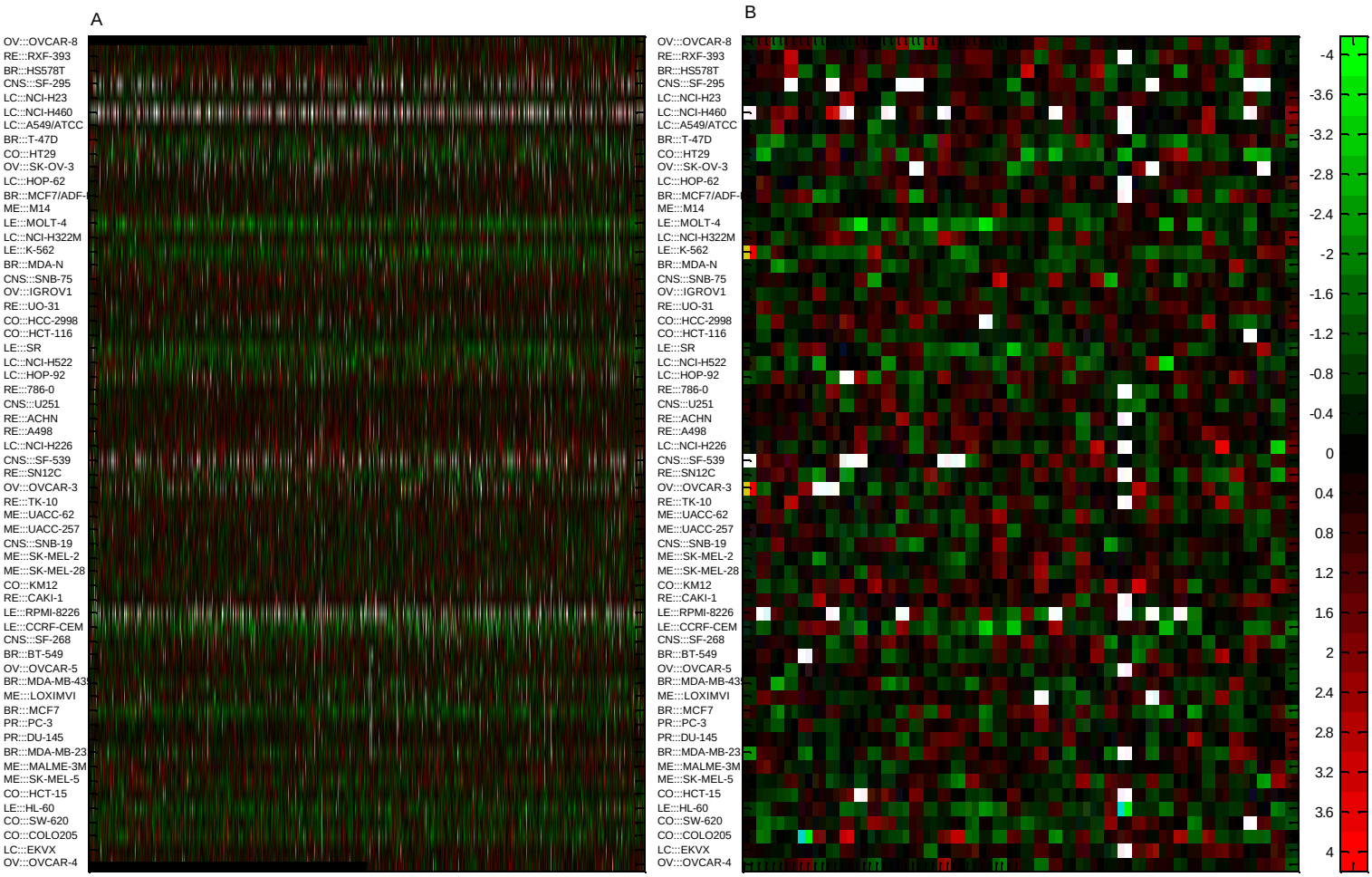

Fig. 1. (A)The unordered gene expression data matrix of human tumor cell lines. (B)Outliers identified in the image of the residuals : The missing values are colored white, and the outliers are colored yellow (higher than expected) or blue (lower than expected). To be able to view the figure clearly, we only select 40 columns to illustrate here.

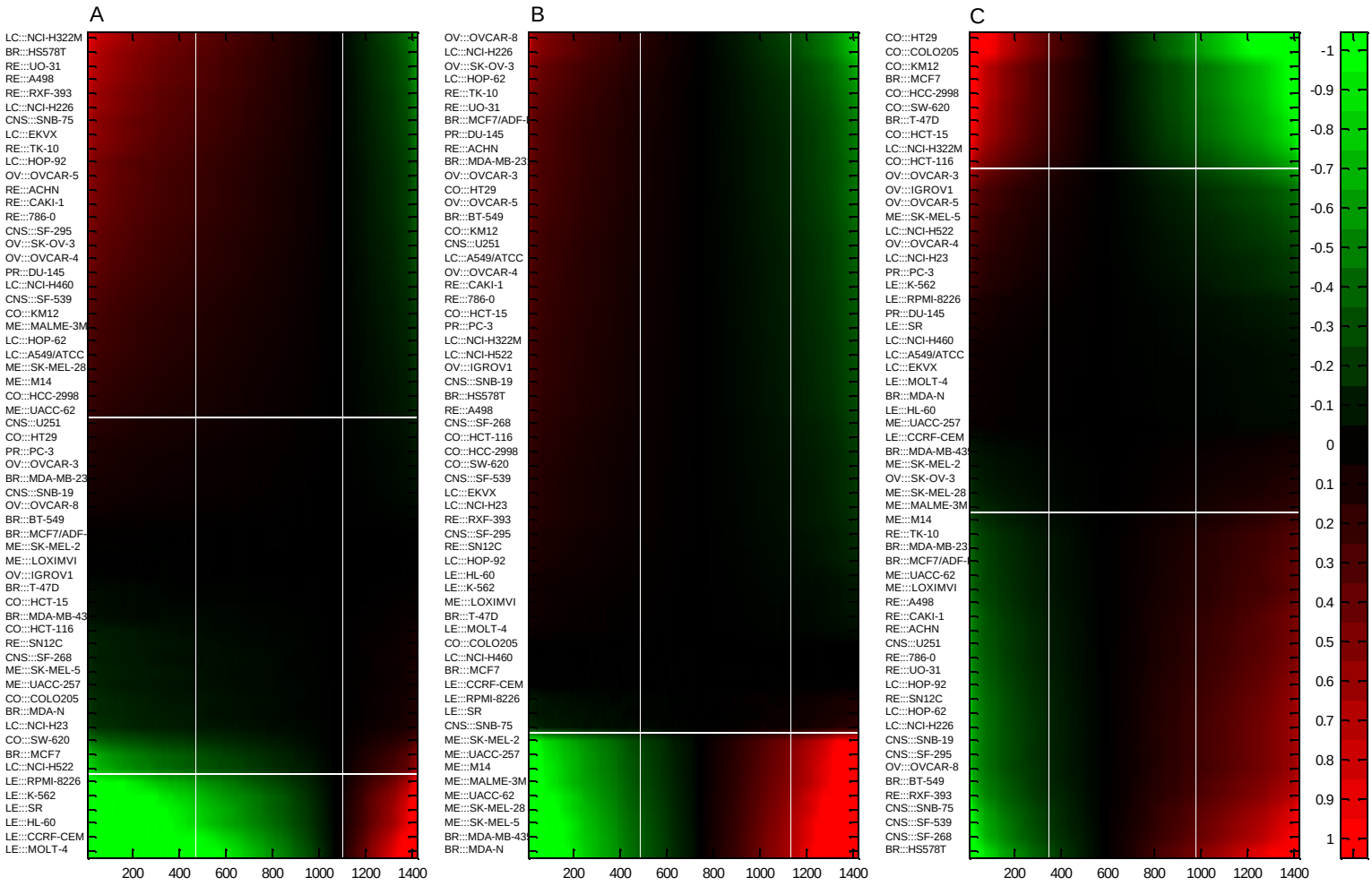

Fig. 2. (A) The first SVD component. (B) The second SVD component. (C) The third SVD component.

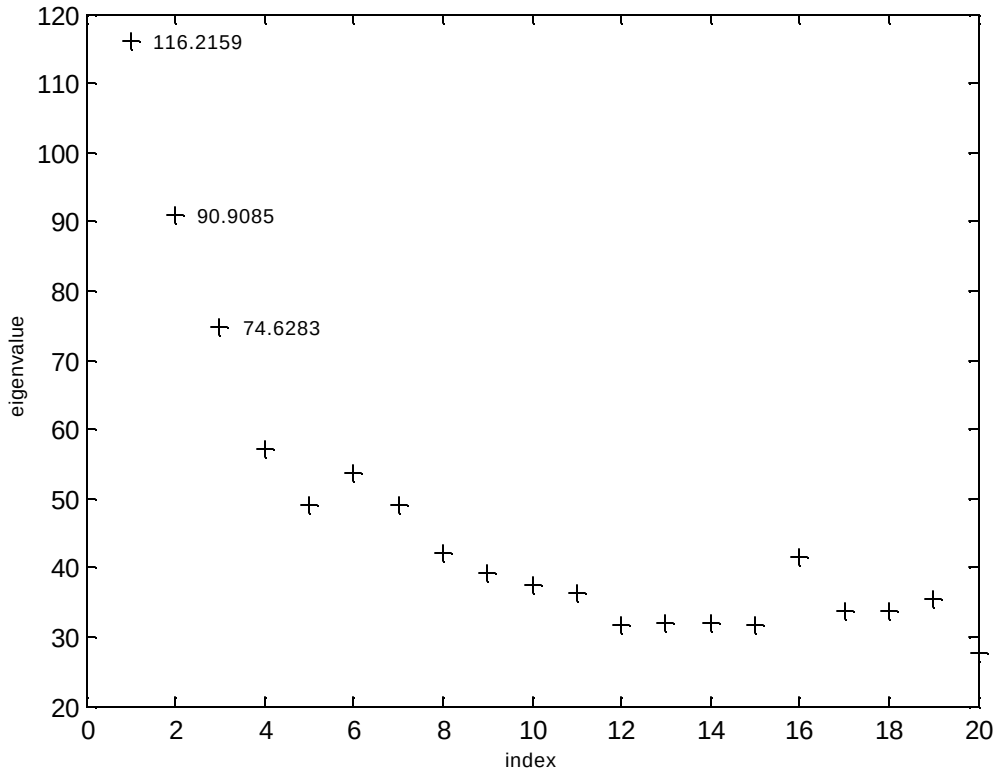

Fig. 3. The plot of the eigenvalues

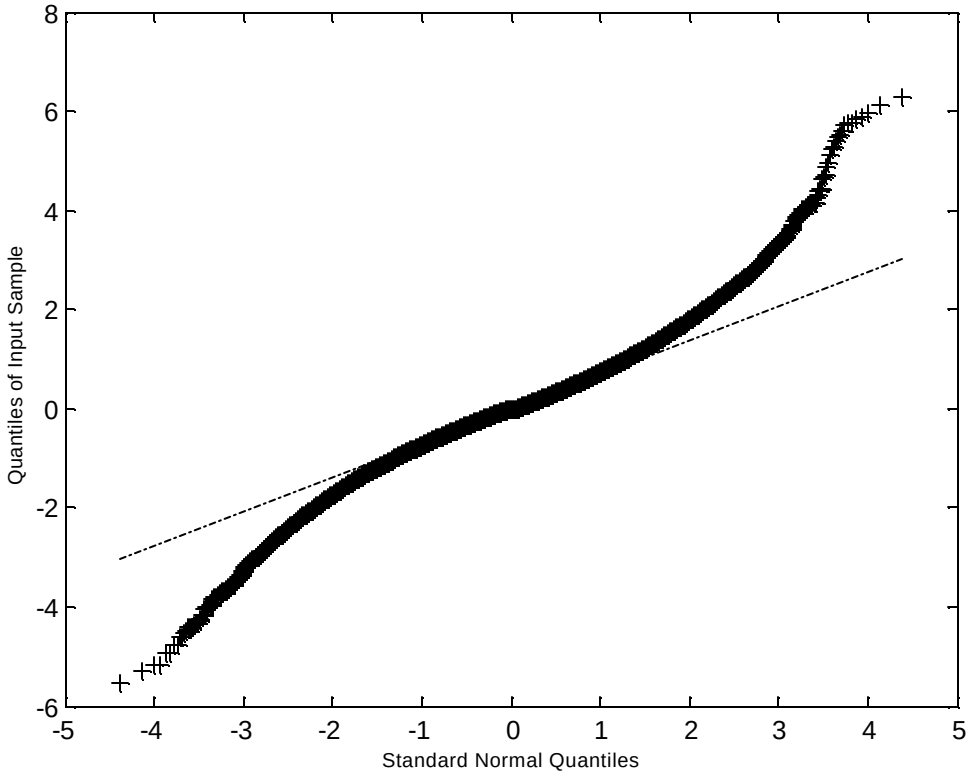

Fig. 4. The QQ plot of the residuals