NISS

# Robust Singular Value Decomposition

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

Technical Report Number 122
December, 2001

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

Douglas M. Hawkins[a,*], Li Liu[b], S. Stanley Young[c]

[a]*School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street NE, Minneapolis, MN 55455-0493, USA*
[b]*National Institute of Statistical Sciences, P.O. Box 14006, Research Triangle Park, NC 27709-4006, USA*
[c]*GlaxoSmithKline, Research Triangle Park, NC 27709, USA*

**Abstract**

The singular value decomposition of a rectangular data matrix can be used to understand the structure of the data and give insight into the relationships of the row and column factors. For example, the rows linked to the rows might be experimental conditions of temperature and the experimental conditions linked to the columns might pressure. In a biological setting the rows might be linked to tissues and the columns linked to genes. In experimentation, there might be aberrant values, outliers, or missing values that arise from flaws in the execution of the experiment so there is a need for singular value decomposition of data tables with missing values and outliers. Our idea is to use a sequential estimation of the eigenvalues and left and right eigenvectors that ignores missing values and is resistant to outliers. The benefit of our robust SVD is that data tables with experimental flaws, outliers and missing data, can be examined more easily.

*Keywords:* Singular value decomposition; Robust estimation; Alternating L1 regression; Outliers; Missing values; Biplot.

## 1. Introduction

The singular value decomposition (SVD) of a rectangular data matrix is a powerful tool in understanding its structure. SVD underpins such methods as the biplot (Bradu and Gabriel 1978), correspondence analysis (Greenacre 1984) and

* Corresponding author. Fax: 1-612-624-8868.
*E-mail addresses*: doug@stat.umn.edu (D.W. Hawkins), liliu@niss.org (L. Liu), ssy0487@gsk.com (S. S. Young).

principal component analysis. It is also sometimes used as a clustering method through the method of Q-mode analysis. An exposition of some variants of its use is given in Greenacre and Underhill (1982).

Let **X** be a matrix of order *nxp*. It is generally helpful to think of its rows are representing *n* cases and its columns as representing *p* variables. Typically p is much smaller than n, and we will use this framework for discussion. Note though that there is nothing to stop *p* exceeding *n*, as indeed is the case for a number of important problems in the analysis of microarray and chemometric problems.

The SVD of the matrix is

$$\mathbf{X} = \mathbf{A\Lambda B^T}$$

where **A** is the *nxp* column-orthogonal matrix of left eigenvectors; **Λ** the *pxp* diagonal matrix of eigenvalues, and **B** the *pxp* orthogonal matrix of right eigenvectors. Writing $\mathbf{a}_i$ and $\mathbf{b}_i$ for the $i^{th}$ left and right eigenvectors respectively, and $\lambda_i$ for the $i^{th}$ eigenvalue, the SVD can be written

$$\mathbf{X} = \sum_{i=1}^{p} \lambda_i \mathbf{a}_i \mathbf{b}_i^{T}$$

It is well known that if the summation is truncated to just *k* terms, the right hand side is the least-squares rank *k* approximation to *X* (see for example Greenacre and Undergill 1982).

Conventionally, the SVD is calculated through a principal component analysis of $\mathbf{X^TX}$. Since

$$\mathbf{X^TX = B\Lambda A^T A\Lambda B^T = B\Lambda^2 B^T}$$

this yields the right eigenvectors and the squares of the eigenvalues of **X**. Then **A** can be calculated from

$$\mathbf{A = XB\Lambda}^{-1}$$

If $n<p$, then it is computationally faster to carry out a PCA on $\mathbf{XX}^{\mathrm{T}}$; the calculations exactly parallel those sketched.

## 2. Alternating least squares approach

The conventional approach to the SVD requires that the matrix **X** be complete. If it has any missing elements, the calculation as sketched cannot be performed. An alternative iterative approach due to Gabriel and Zamir (1979) addresses this problem.

Its basis is that the leading eigen triple $\lambda_1$, $\mathbf{a}_1$, $\mathbf{b}_1$ is the least squares approximation to **X.** Their alternating least squares (ALS) algorithm works iteratively. It starts with an initial estimate of $\mathbf{a}_1$. Then estimates of the successive elements of $\mathbf{b}_1$ are found by regressing each column of **X** on $\mathbf{a}_1$ . The resulting coefficients are estimates of $\lambda_1 b_{j1}$. At the end of the cycle, scaling the vector estimate to unit length gives the estimate of the $\mathbf{b}_1$, and the scale factor estimates $\lambda_1$.

Having obtained the estimate of $\mathbf{b}_1$, the next step is to refine the estimate of $\mathbf{a}_1$. This is done by regressing each row of **X** on the estimated $\mathbf{b}_1$, and again normalizing the estimate.

Since each step of the ALS reduces the residual sum of squares from that obtained at the preceding step, the process converges. Gabriel and Zamir claimed that, part from recognizable degenerate solutions that can arise, the convergence is to the true leading eigenvalue / eigenvector triple.

Once the first eigen solution has been obtained, we replace **X** by the deflated matrix

$$\mathbf{X} - \lambda_1 \mathbf{a}_1 \mathbf{b}_1$$

and start the iteration anew to get the second eigen triple.

This ALS approach is stable and easy to program. While it is not particularly fast, it can be computationally attractive if both *n* and *p* are large, but only a few eigen solutions are needed.

## 3. The effect of outliers on the SVD and robust approaches

As the SVD is a least squares procedure, it is highly susceptible to outliers. In the extreme case, an individual cell, if sufficiently outlying, can draw even the leading principal component toward itself. It is therefore desirable to have some way of computing a robust SVD.

One obvious approach is through a robust principal component analysis. Instead of diagonalizing $\mathbf{X}^T\mathbf{X}$, use some resistant scatter matrix such as the minimum covariance determinant (MCD, Rousseeuw 1984, Hawkins and Olive 1999, Rousseeuw and Van Driessen 1999). This method though is not well suited to the problem or making the SVD resistant to a minority of outlying cells. Resistant scatter matrices are based on the premise of a data matrix, most of whose entire row vectors are ‘clean’, with a minority of arbitrary outlier vectors. But there are many applications in which the outliers are individual cells in otherwise-good rows. Consider for example a data matrix with 100 rows and 2000 columns, such as is commonly seen in microarray settings. Suppose that 2% of the cells are outlying, and that the outliers are located at random. Then in expectation only 13% of the rows will be outlier-free; far below the majority required for resistant covariance matrix approaches to succeed.

Galpin and Hawkins (1987) and Choulakian (2001) give another approach based on the variational properties of principal components. Classical PCA maximizes (minimizes) the variance of the linear combination of the variables subject to orthogonality with the preceding (succeeding) components, while this method replaces the objective function of variance with sum of absolute deviations and leads to linear or quadratic programming formulations for each eigenvalue and right eigenvector pair. The Galpin-Hawkins approach does not directly address the question of finding the left eigenvectors.

To solve the problem of modeling rectangular data arrays with a minority of outlier cells, we propose a robust method we call AL1-SVD, for alternating L1 Regression for SVD. It has a similar flavor to the Gabriel-Zamir ALS approach and, like that approach, can accommodate missing values in the data matrix. It differs from ALS in using a more resistant measure than least squares.

## 4. The alternating L1 regression algorithm

The leading eigenvalue and eigenvector pair has the property of minimizing the Euclidean norm of the unexplained portion of the data matrix. We will replace this by the criterion of minimizing the L1 norm – the sum of absolute values – of the unexplained portion of the data matrix, and implement it by alternating L1 (AL1) regressions.

*Algorithm*

- Start with an initial estimate of the leading left eigenvector $\mathbf{a}_1$
- For each column $j$, $j = 1, 2, \ldots, p$, fit the L1 regression coefficient $c_j$ by

$$\min \sum_{i=1}^{n} | x_{ij} - c_j a_{i1} |$$

- Calculate the resulting estimate of the right eigenvector $\mathbf{b}_1 = \mathbf{c} / ||\mathbf{c}||$, where $||.||$ refers to Euclidean norm.
- Using this estimate of the right eigenvector, refine the estimate of the left eigenvector. For each row $i$, $i$=1, 2, …, $n$, fit the L1 regression coefficient $d_i$ by

$$\min \sum_{j=1}^{p} | x_{ij} - d_i b_{j1} |$$

- Calculate the resulting estimate of the left eigenvector $\mathbf{a}_1 = \mathbf{d} / ||\mathbf{d}||$.
- Iterate to convergence.

The L1 regressions involved are trivial, reducing to finding weighted medians.

$$\min \sum_{i=1}^{n} | x_{ij} - c_j a_{i1} | \;=\; \min \sum_{i=1}^{n} | a_{1j} | . | x_{ij} / a_{1j} - c_j |$$

which is solved by the weighted median of the ratios $x_{ij}/a_{1j}$ with weights $|a_{1j}|$. If $a_{1j}$=0, then the term can be ignored as it does not contribute to the weight. Missing cells are omitted from the calculation of the median.

There is no very obvious choice of a starting estimate of the leading left eigenvector. One reasonable choice is to use the vector of median absolute values of the rows, rescaled to unit length.

Note that convergence of a sort is guaranteed in that each step of the AL1 reduces the fit criterion. However it is not automatic that the eigenvectors will converge to fixed values. The median of a set of values is not always unique (and this is also true of the weighted median), so care must be taken if you want to ensure a unique solution for the eigenvectors. The most transparent solution is to use for the weighted median the center of whatever range of values might be found to minimize the criterion. Since the criterion is convex, there will be at most one such range.

This alternation gives the first eigenvector pair. Once the criterion value has stabilized, the L1 eigenvalue is found through the minimization

$$\min \sum_{i=1}^{n} \sum_{j=1}^{p} \; | x_{ij} - \lambda_1 a_{i1} b_{j1} |$$

This is another single-parameter L1 regression, and is given by a weighted median.

For the second and subsequent of the SVD, we replace **X** by a deflated matrix obtained by subtracting the most recently found term in the SVD

$$\mathbf{X} \leftarrow \mathbf{X} - \lambda_k \mathbf{a}_k \mathbf{b}_k^T$$

## 5. Comparison with the conventional L2 SVD

Two differences from the usual SVD may be noted. One relates to orthogonality. In the conventional SVD, all the eigenvectors are orthogonal even if not explicitly imposed. In complete data problems, the vectors returned by the ALS algorithm will be orthogonal. Those returned by the AL1 algorithm are, in general, not orthogonal.

Another difference is that, in the L2 analysis of the conventional SVD, the successive eigen triples are found in descending order of eigenvalue. This is not necessarily the case with the AL1 algorithm; it is our common experience that a term with larger $\lambda$ may follow one with smaller, something that the 'power' property of the L2 SVD precludes.

## 6. Examples

The SVD has many uses; for our examples we will use one of them – using the SVD to construct a biplot to gain an understanding of the additivity or otherwise of an unreplicated two-way anova layout.

### *6.1 Simulation*

As explained by Bradu and Gabriel (1978), to construct a biplot we take a rank-2 approximation to the unreplicated two-way layout **X**

$$\mathbf{X} \approx LR^T$$

where $L$ and $R$ are the first two left and right eigenvectors of **X** rescaled by apportioning the first two eigenvalues between them. Then using **G** as bivariate coordinates allows us to plot a representation of the rows – this set of points is termed the 'row markers'. The column markers are similarly defined. An additive model for **X** can be inferred if the row and column markers are both straight lines, and lie at right angles. If they are lines but at an angle other than $90^o$, the diagnosis is of the Tukey 'one degree of freedom' model. A line for the row markers but not the column markers shows that there is a row regression model (Mandel 1969). The same interpretation frame applied to subsets of the markers can identify submatrices that satisfy particular simple models.

To illustrate biplotting with outlier-contaminated data, we generated a 10*10 additive data matrix $x_{ij} = \mu+\alpha_i+\beta_j+e$, where $\mu = 1$, $\alpha = -5,-4,\ldots, 4, 5$, $\beta = -5, -4, \ldots, 4, 5$ and the random noise term $e$ is $N(0,0.125)$. This data matrix is of approximate rank two after the overall mean is removed.

To obtain $L$ and $R$ for the generated data matrix, we use both regular SVD and robust SVD. The biplots of the data matrix are shown in Figure 1 and Figure 2. Notice that in both biplots, the row markers form a straight line, the column markers form a straight line and they are at a right angle to each other. This correctly diagnoses an additive model for the table.

Now we contaminate the data by adding four outliers (add 15 to four randomly chosen cells in the data table). The regular SVD and robust SVD are performed again on the data matrix with outliers to get $L$ and $R$. The new biplots are shown in Figure 3 and Figure 4. As we can see, we do not get straight lines from the biplot based on the regular SVD. The markers are scattered around and have no visible simple structure.  In contrast, the biplot based on the robust SVD is almost the same as before. Two straight lines are at a right angle, which indicates an additive model. This illustrates the advantage of the robust SVD, whose analysis is not or is less influenced by the outliers.

### *6.2 Rubber data*

This is the data on specific volume of rubber, which is analyzed by Bradu and Gabriel (1978), see Table 1. There are three factors, which are treatment of rubber (2 levels), temperature (4 levels) and pressure (6 levels), so it is a three way table. The data is not complete in that the 500 kg/cm$^2$ observations for the unvulcanized rubber at 0$^o$ centigrade is missing.

To illustrate the data with biplot, we consider two way table by combining temperature and treatment as rows. The two factors will be separated later.

We perform robust SVD on this dataset since the method can handle missing values. The rank two approximation gives a goodness of fit of $R^2$ = 99.995%, and the missing value is estimated to be 172.762 (as compared to Mandel’s estimate of 173, and Bradu and Gabriel’s estimate of 173.578).

The biplot of this rank two approximation is shown in Figure 5. As we can see, all the column markers (pressure levels) form a straight line, while all the row markers roughly form a line which is almost perpendicular to the line of column markers. This indicates an additive model will be relatively good.

On closer inspection, we see that the row markers divide naturally into two groups, each of which is very close to linear.  These two groups turn out to correspond to the treatments (peroxide cured or unvulcanized). The two treatment lines are parallel and almost at the right angles to the line of column markers. This verifies that there is a small treatment difference and a large temperature effect, and that the model is close to additive in pressure and the Kronecker product of temperature and treatment.

This conclusion is consistent with that of Bradu and Gabriel (1978). Now let us contaminate the data with a couple of outliers, and see what happens. Figure 6 shows the biplot based on the approximation by the robust SVD. Not surprisingly, the result is similar to that without outliers.  However, the biplot based on the regular SVD is not satisfactory, see Figure 7.  The markers are no longer straight lines and seem to have no simple pattern to them.  The outliers therefore have ruined the conventional L2-norm SVD, but have had no substantive impact on the AL1 SVD.

## 7. Conclusion

The SVD is a tool with many uses. Our two examples have illustrated the biplot, which is interesting as a method that uses the SVD directly to infer data structure, but Q-mode clustering, principal component analysis, and correspondence analysis also rely on it.

As a least squares method though, the SVD is non-robust, being highly sensitive to even modest numbers of outliers in the data array. Our alternating L1-norm approach leads to an outlier-resistant approach that also accommodates missing information. The latter alone is enough to make it attractive compared to other high breakdown methods that require complete data, but even more compelling is that is well suited to the types of outlier commonly seen in rectangular data arrays – that is outliers affecting isolated cells rather than entire rows or columns.

It is simple to implement, and while not particularly fast, executes quickly enough not to create an obstacle to its wider use.

## References

Bradu, D., Gabriel K.R., 1978. The biplot as a diagnostic tool for models of two-way tables. Technometrics, 20, 47-68.

Choulakian, V., 2001. Robust Q-mode principal component analysis in L1. Comput. Statist. Data Anal., 37, 135-150.

Demmel, J.W.,1997. Applied Numerical Linear Algebra. SIAM, Philadelphia, 361-386.

Gabriel, K. R., Zamir , S., 1979. Lower rank approximation of matrices by least squares with any choice of weights. Technometrics, 21, 489-498.

Galpin, J.S., Hawkins, D.M., 1987. $L_1$ estimation of a covariance matrix. Comput. Statist. Data Anal., 5, 305-319.

Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis. Academic Press, London.

Greenacre M.J., Underhill L.G., 1982. Scaling a data matrix in a low dimensional Euclidean space, In Hawkins DM (ed.). Topics in applied multivariate analysis. Cambridge University Press, Cambridge, 183-268.

Hawkins, D.M., Olive, D.J., 1999. Improved feasible solution algorithms for high breakdown estimation. Comput. Statist. Data Anal., 30, 1999, 1-11.

Mandel, J., 1969. A method for fitting empirical surfaces to physical or chemical data. Technometrics, 11, 411-429.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Amer. Statist. Assoc., 79, 871-880.

Rousseeuw, P.J., van Driessen, D.K., 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics, 41, 212-223.

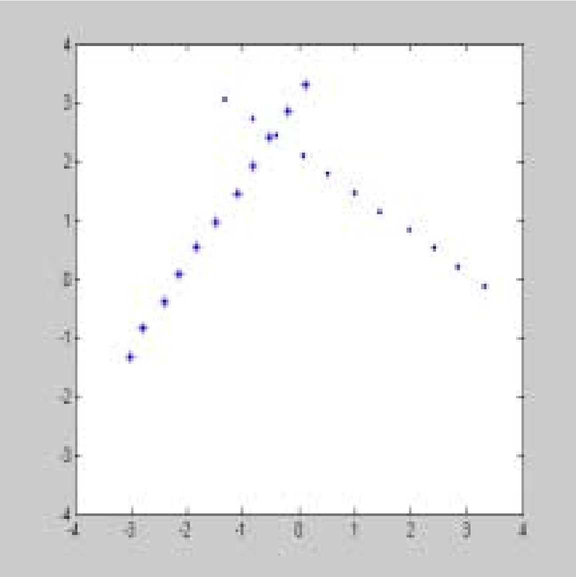


Fig. 1. Biplot based on the regular SVD without outliers.

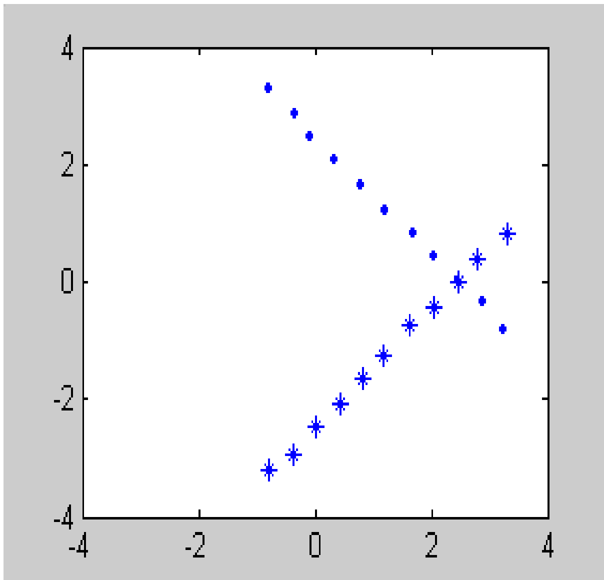


Fig. 2. Biplot based on the robust SVD without outliers.

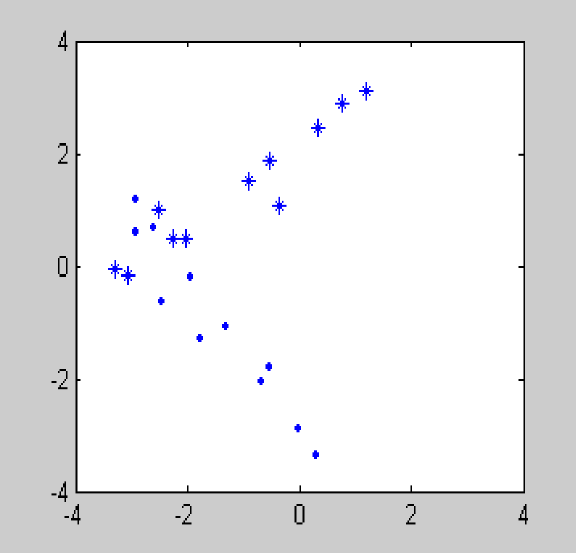


Fig. 3. Biplot based on the regular SVD with outliers.

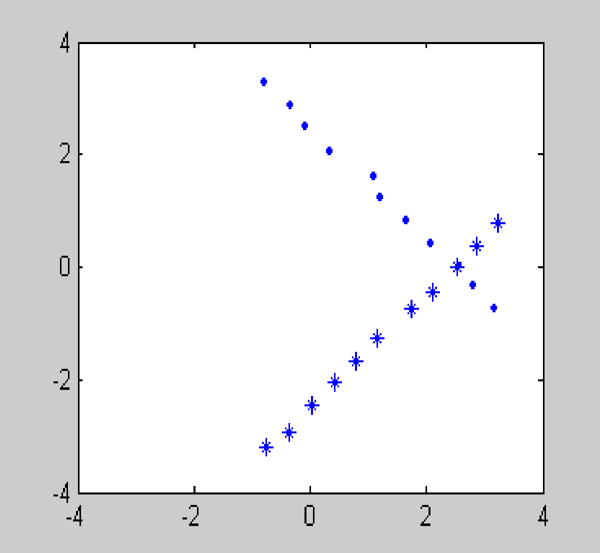


Fig. 4. Biplot based on the robust SVD with outliers.

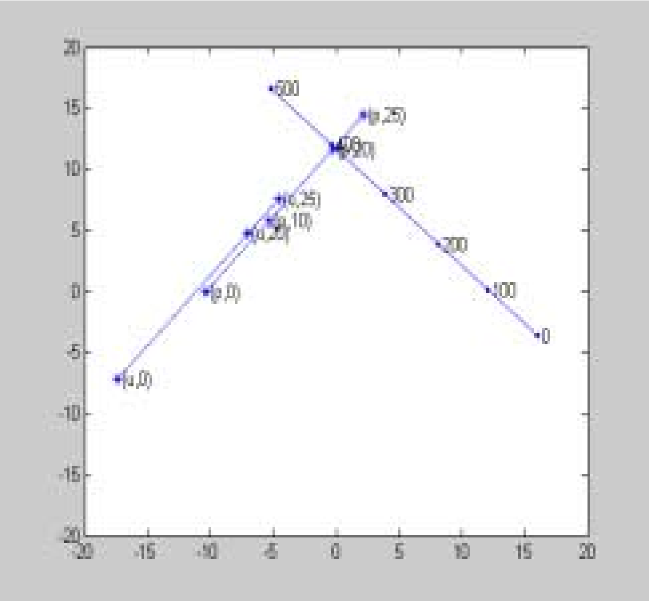
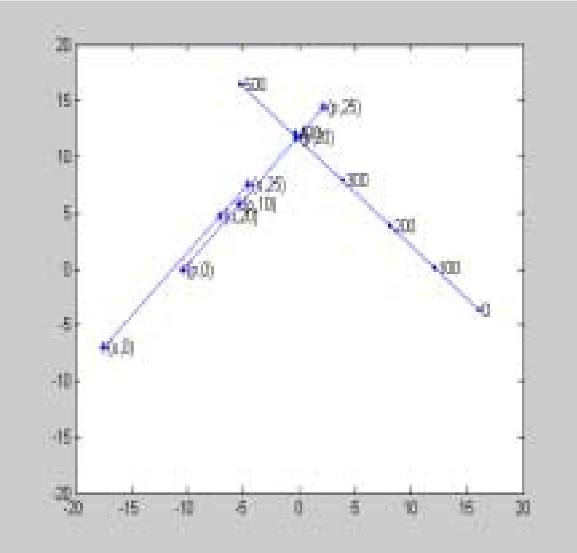


Fig. 5. The biplot of the rubber data based on the robust SVD.



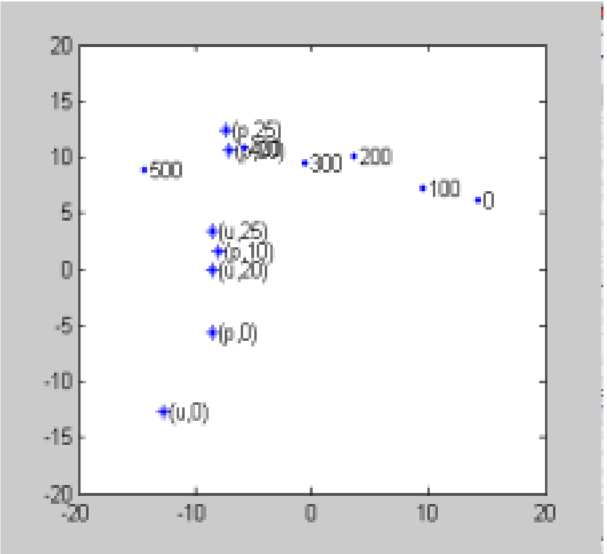


Fig. 6. The biplot of the rubber data with outliers based on the robust SVD.

Fig. 7. The biplot of the rubber data with outliers based on the regular SVD.

Table 1
Specific volumes of two rubbers

| | | Pressure ( kg/cm$^2$ ) | | | | | |
|---|---|---|---|---|---|---|---|
| Rubber | Temp ($^o$C) | 500 | 400 | 300 | 200 | 100 | 0 |
| Peroxide Cured | 0 | 137 | 178 | 219 | 263 | 307 | 357 |
| | 10 | 197 | 239 | 282 | 328 | 376 | 427 |
| | 20 | 256 | 301 | 346 | 394 | 444 | 498 |
| | 25 | 286 | 330 | 377 | 426 | 477 | 532 |
| Unvul-Canized | 0 | 54 | 93 | 136 | 179 | 225 | 272 |
| | 20 | ? | 218 | 264 | 314 | 364 | 417 |
| | 25 | 202 | 248 | 295 | 345 | 396 | 451 |

* To see the effect of the outliers, we replace 346 by 293 (peroxide cured, 20 $^o$C, 300 kg/cm$^2$), and 272 by 263 (unvulcanized, 0$^o$C, 0 kg/cm$^2$).