Computers & Geosciences 58 (2013) 19–28

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Modeling by singular value decomposition and the elimination of statistically insignificant coefficients

G.V. Haines, R.A.D. Fiori *

Geomagnetic Laboratory, Natural Resources Canada, Ottawa, Ontario, Canada

ARTICLE INFO

Article history:
Received 12 June 2012
Received in revised form
6 March 2013
Accepted 25 April 2013
Available online 3 May 2013

Keywords:
SVD
Stepwise regression
Numerical analysis

ABSTRACT

Singular value decomposition (SVD) has numerical advantages over other least squares modeling techniques because it requires the summation of basis functions only, rather than of their squares and products. It also transforms the original independent variables to an orthogonal system of variables, thus exposing issues of collinearity and singularity. The SVD approach by itself, however, is simply a decomposition of this original matrix of independent variables, and does not refer to observations affected by errors. With no information on observational errors, it does not include a method for rejecting model coefficients that have little statistical significance. Eliminating singular values to reduce model dimensionality in the least squares application of SVD can thus be done on the basis of statistical error tests, a procedure not directly available to many other applications of the SVD method. A statistical backward elimination procedure applied directly to the transformed SVD principal components compares well with a stepwise procedure applied to the original untransformed coordinates, allowing advantage to be taken of the numerical superiority of SVD. On the other hand, it is important to understand that the approaches taken by SVD and ordinary least squares (OLS) in handling singularities are quite different, and in these cases can lead to different solutions. Analyses of several singular and near-singular least squares matrices in the literature, as well as two real-world examples of modeling electric field, demonstrate the similarities and differences between the two least squares approaches, and the benefit of a statistical rejection procedure in both of them.

1. Introduction

Many programs now use singular value decomposition (SVD) to obtain least squares model coefficients by means of principal component analysis. The SVD approach is recognized to be a very stable solution to the least squares problem, and provides information on the singularity or near-singularity of the least squares sums of squares and products matrix, and on the collinearity or near-collinearity of the model variables. It is therefore easy to eliminate principal components based on very small singular values resulting from imprecise calculations based on the computer word size available. However, SVD is fundamentally a matrix decomposition and therefore does not in itself look at the errors in the dependent variable or allow one to eliminate variables based on a statistical test. In fact, in many applications using SVD, observational errors do not even play a major role. SVD is, or at least was originally, primarily aimed at the solution of a system of linear equations. Of course, the normal equations of linear least squares are also a system of linear equations, but they have statistical ramifications as well. The observational equations have an attached error, and the solution of the system depends on the statistical properties of that error. One can, for instance, determine the scatter or standard error of estimate about the model, and use that to estimate the error in the model coefficients. One may then decide, on the basis of a statistical test, that a certain subset of the modeled coefficients is significant (at a given level of significance), whereas the others are not. Using a procedure such as this provides a statistically meaningful way of choosing which singular values are not appropriate for inclusion in the model and which therefore can be eliminated to reduce the model dimensionality. By eliminating insignificant coefficients, dimensionality reduction gives a more accurate standard error of estimate.

Least squares programs that do not use SVD solve a system of normal equations which involve squares and products of basis functions rather than simply the basis functions themselves as in SVD. For that reason, the SVD approach is numerically superior to that of the traditional least squares approach. However, many of these traditional programs allow for statistical coefficient rejection procedures, and therefore respond well (or can easily be modified to respond well) to singularities and other issues of poor data or data design. Their modeling philosophy is different from that of SVD, though, in that they attempt to minimize the size of the

* Corresponding author. Tel.: +1 6138375137.
E-mail address: Robyn.Fiori@NRCan-RNCan.gc.ca (R.A.D. Fiori).

0098-3004/$ - see front matter
http://dx.doi.org/10.1016/j.cageo.2013.04.021

model rather than estimate coefficients of every possible variable initially considered in the design. This paper will briefly review the two approaches and discuss through examples the similarities and differences of the two techniques. The traditional least squares program used for comparison purposes in this paper is called STEPREG and uses a stepwise regression procedure for coefficient rejection introduced by Efroymson (1960), and is applied to the original data coordinates. A second program is described in this paper which uses a backward elimination procedure, and is applied to the transformed principal coordinates of SVD. Coefficient rejection in this case is done by eliminating the corresponding singular values.

The mathematics of singular value decomposition and its relationship to least squares analysis are well explained in detail by many authors, including Forsythe et al. (1977), Lawson and Hanson (1974), Menke (1989), Nash (1990), and Rawlings et al. (1998).

In this paper, Sections 2 and 3 describe singular value decomposition and its application to the method of least squares, respectively. The stepwise regression procedure for eliminating statistically insignificant coefficients in ordinary least squares (OLS) is described in Section 4. The analysis of principal components obtained by means of SVD is presented in Section 5, and a statistical backward elimination procedure that is applied directly to the SVD principal component coefficients to determine which singular values should be zeroed is suggested in Section 6. Illustrative examples of the issues and techniques described in this paper and a comparison of the SVD and OLS methods are presented in Section 7. Throughout this paper, references are made to specific routines, which are required for running the examples described in Section 7. Fortran programs and the data used in Section 7 may be obtained from Robyn Fiori (Robyn.Fiori@NRCan-RNCan.gc.ca).

2. Singular value decomposition (SVD)

The (reduced or thin) singular value decomposition (SVD) of an $n \times m$ real matrix $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{1}$$

where $\mathbf{U}$ is an $n \times m$ matrix with orthonormal columns, $\mathbf{S}$ is an $m \times m$ diagonal matrix of the singular values of $\mathbf{X}$, and $\mathbf{V}$ is an $m \times m$ orthogonal matrix (Golub and Van Loan, 1996, Section 2.5.4; Miller, 2002, Section 2.2; Press et al., 1992, Section 2.6; Rawlings et al., 1998, Section 2.8). The superscript T indicates the matrix transpose. In modeling work, $\mathbf{X}$ is the design matrix, m is the number of basis functions, n is the number of observations, and usually $n \geq m$. (This latter condition is discussed in Section 3; however, when $n < m$ Eq. (1) is still valid but $m-n$ columns of $\mathbf{U}$ are not normalized but zero, as are the corresponding singular values.) Since $\mathbf{U}$ is column orthonormal $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m$, the $m \times m$ identity matrix, and since $\mathbf{V}$ is orthogonal $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_m$. If $\mathbf{u}_j$ and $\mathbf{v}_j$ are the jth column vectors of $\mathbf{U}$ and $\mathbf{V}$, respectively, and s_j is the jth diagonal element of $\mathbf{S}$, Eq. (1) is equivalent to a summation of m rank-one matrices

$$\mathbf{X} = \sum_{j=1}^{m} \mathbf{u}_j s_j \mathbf{v}_j^T \tag{2}$$

The diagonalizing aspect of SVD can be better appreciated by pre- and post-multiplying Eq. (1) by $\mathbf{U}^T$ and $\mathbf{V}$, respectively

$$\mathbf{S} = \mathbf{U}^T\mathbf{X}\mathbf{V} \tag{3}$$

which is equivalent to

$$s_j = \mathbf{u}_j^T \mathbf{X} \mathbf{v}_j, \quad j = 1, \ldots, m \tag{4}$$

If $\mathbf{X}$ is of rank r, $m-r$ of the singular values will be zero. In fact, a threshold depending on the computer word size is usually chosen, and all singular values less than this threshold are set to zero, so that often even more than $m-r$ of the singular values will be zero. Even singular values larger than the threshold may be set to zero, of course, and it will be shown in Section 6 that a statistical test may be used to determine which singular values may be granted that distinction.

$\mathbf{V}$ is a transformation matrix that rotates/reflects the matrix $\mathbf{X}$ into a matrix whose columns are orthogonal. This can be seen by post-multiplying Eq. (1) by $\mathbf{V}$

$$\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{S} \tag{5}$$

This equation is equivalent to

$$\mathbf{X}\mathbf{v}_j = s_j \mathbf{u}_j, \quad j = 1, \ldots, m \tag{6}$$

Since the $s_j\mathbf{u}_j$ are orthogonal, the columns of $\mathbf{XV}$ are orthogonal. This will be used in Section 5, where the matrix $\mathbf{XV}$ will be referred to as $\mathbf{W}$.

The columns of $\mathbf{V}$ are in fact unit normal vectors in the transformed coordinate system. This will be shown at the end of Section 5 in connection with Eqs. (36) and (37).

The Fortran SVD programs used in this analysis are SVDCMP, SVDVAR, and SVBKSB, taken from Numerical Recipes (Press et al., 1992). However, there was a bug in our version of the Numerical Recipes subroutine SVDCMP, and a corrected version was downloaded from http://info.ifpan.edu.pl/~kisiel/struct/rgdfit/rgdfit.for.

The condition number κ of a matrix can be defined as the ratio of the largest singular value to the smallest. When κ^{-1} approaches the computer's floating point precision, the accuracy of floating point calculations will disappear. Double precision on a 32-bit computer (64-bit word size with an implicit 53-bit mantissa) is approximately 10^{-16}. When using double precision, the threshold mentioned above will be taken as something close to 10^{-16}. Press et al. (1992, Section 2.6, p. 56) state that if the small singular values (those below a given threshold) have not been zeroed, then the SVD technique is “just as ill-conditioned as any direct method, and you are misusing SVD.” They also suggest that choosing the threshold is somewhat subjective: “SVD cannot be applied blindly, then. You have to exercise some discretion in deciding at what threshold to zero the small [singular values].” These statements will be considered further in Sections 7.2 and 7.3.

Alternative factorizations and decompositions of $\mathbf{X}$ and of the least squares sums of squares and products matrix $\mathbf{X}^T\mathbf{X}$ (see Eq. (8) in the next section), for purposes of orthogonal least squares reduction procedures, have been described by Miller (2002, Section 2.1, 2.2).

3. Least squares

A least squares problem may be represented by the matrix equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{7}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations (or data), $\boldsymbol{\beta}$ is an $m \times 1$ vector of unknown coefficients or parameters to be estimated, $\mathbf{X}$ is an $n \times m$ design matrix, i.e. matrix of the m (non-stochastic) coordinates or basis functions in the model evaluated at the n observation points, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of unknown random observational errors. The problem is then to estimate the model coefficients $\boldsymbol{\beta}$ by minimizing $\boldsymbol{\epsilon}^2$ with respect to $\boldsymbol{\beta}$. It is assumed here that there are no errors in the elements of $\mathbf{X}$ and that the errors in $\mathbf{Y}$ are independently and normally distributed with zero mean and unknown but constant variance. It is assumed also that $n > m$, i.e. the least squares problem is overdetermined. If $n < m$ the problem would be underdetermined, and it would be