ACM/ESE 118

Singular value decomposition

Definition. For any real $n \times p$ matrix X, there exist orthogonal matrices (i.e., $\mathsf{U}^T\mathsf{U} = \mathsf{I}$ and $\mathsf{V}^T\mathsf{V} = \mathsf{I}$)

$$\mathsf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n) \in \Re^{n\times n} \quad \text{and} \quad \mathsf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p) \in \Re^{p\times p}$$

and a diagonal matrix

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_q) \in \Re^{n\times p}, \qquad q = \min(n, p),$$

with

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q \geq 0,$$

such that

$$\mathsf{X} = \mathsf{U}\boldsymbol{\Sigma}\mathsf{V}^T.$$

The columns $\mathbf{u}_i$ of the matrix U are the *left singular vectors*; the columns $\mathbf{v}_i$ of the matrix V are the *right singular vectors*; and the diagonal elements σ_i of the matrix $\boldsymbol{\Sigma}$ are the *singular values*.

Properties. The *rank* r of the matrix X is the number of nonzero singular values,

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_q = 0.$$

The right singular vectors $\mathbf{v}_i$ associated with zero singular values ($i > r$) span the *nullspace* of the matrix X. The left singular vectors $\mathbf{u}_i$ associated with nonzero singular values ($i \leq r$) span the *range* of X.

Least squares estimates (full rank). If the matrix X of predictors is of full rank (i.e., $n \geq p$ and $r = p$), the least squares estimate $\hat{\boldsymbol{\beta}}_{LS}$ of the regression coefficients in the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} = \mathcal{N}(0, s^2\,\mathsf{I}), \tag{1}$$

is unique and unbiased and can be written as

$$\hat{\boldsymbol{\beta}}_{LS} = \sum_{i=1}^{p} \frac{\mathbf{u}_i^T\mathbf{y}}{\sigma_i}\,\mathbf{v}_i. \tag{2}$$

That is, the least squares estimate $\hat{\boldsymbol{\beta}}_{LS}$ is a linear combination of the right singular vectors $\mathbf{v}_i$ of the matrix X of predictors.

The covariance matrix $\mathrm{cov}(\hat{\boldsymbol{\beta}}_{LS}) = s^2(\mathsf{X}^T\mathsf{X})^{-1}$ of the least squares estimate (2) can be written as

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}_{LS}) = s^2 \sum_{i=1}^{p} \frac{\mathbf{v}_i \mathbf{v}_i^T}{\sigma_i^2}. \tag{3}$$

[Note that s^2 is the variance of the error term in the regression model (1) and σ_i^2 is a squared singular value.] This expression shows that an error of order $O(s)$ in the data will typically lead to an error in the estimate $\hat{\boldsymbol{\beta}}_{LS}$ of order $O(s/\sigma_p)$, where σ_p is the smallest singular value of X. If the matrix X has singular values σ_i that are significantly smaller than the standard deviation s of the error, the least squares estimates are poorly constrained and can have large variance.

Least squares estimates (rank-deficient). If the matrix X of predictors is rank-deficient (i.e., $r < p$), the least squares estimate $\hat{\boldsymbol{\beta}}_{LS}$ of the regression coefficients is not unique; adding any vector $\alpha\mathbf{v}_i$ $(i > r, \alpha \in \Re)$ from the nullspace of X to a least squares estimate $\hat{\boldsymbol{\beta}}_{LS}$ gives a new least-squares estimate. Among the estimates $\hat{\boldsymbol{\beta}}_{LS}$ that minimize the residual sum of squares, the estimate

$$\hat{\boldsymbol{\beta}}_{LS} = \sum_{i=1}^{r} \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i} \mathbf{v}_i \tag{4}$$

has minimum norm $\|\hat{\boldsymbol{\beta}}_{LS}\|$. The minimum-norm property is often desirable because it implies that vectors $\mathbf{v}_i$ $(i > r)$ in the nullspace of X, which often represent high-frequency components of the estimate, are filtered out. The minimum-norm estimate (4) is often the "smoothest" estimate of the regression coefficients.

Ill-posed problems. In ill-posed problems, the singular values σ_i gradually decay toward zero and become so small that the least-squares estimate (2) becomes unstable (small changes in data have large effect on estimate). The variance (3) of the least-squares estimate becomes large. In this case, it is necessary to *regularize* the estimate of the regression coefficients by imposing additional constraints. The additional constraints imply that the estimates will no longer be unbiased; however, their variance can be greatly reduced.

Regularization by truncated SVD. In regularization by truncated SVD (TSVD), one filters out not only the right singular vectors $\mathbf{v}_i$ that are associated with zero singular values σ_i, as in the rank-deficient case, but also those associated with "small" singular values σ_i ($i > \hat{r}$). The estimate of the regression coefficients regularized by TSVD is given by

$$\hat{\boldsymbol{\beta}}_{\hat{r}} = \sum_{i=1}^{\hat{r}} \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i} \mathbf{v}_i, \tag{5}$$

where $\hat{r}$ is an estimated effective rank of the matrix X of predictors. The effective rank $\hat{r}$ is the number of singular values that are significantly different from zero; it is the effective number of predictors. The TSVD estimate (5) is the rank-$\hat{r}$ estimate of the regression coefficients with minimum norm. (We will discuss later how one can obtain an estimate of the effective rank $\hat{r}$.)

The TSVD estimate is biased toward zero; however, its variance is reduced compared with that of the least-squares estimate.

References

Golub, G. H., and C. F. van Loan, 1993. *Matrix Computations*. 2nd ed. Johns Hopkins University Press. [See chapters 2.5 and 5.5 for a proof of SVD theorem and for a more detailed mathematical discussion of properties of SVD.]

Hansen, P. C., 1998: *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. Society for Industrial and Applied Mathematics. [See chapters 2.1 and 3.1–3.2 for a more detailed discussion of the use of SVD in rank-deficient and ill-posed problems.]