ON THE STATISTICAL MEANING OF TRUNCATED SINGULAR VALUE DECOMPOSITION

MOODY T. CHU*

Abstract. Empirical data collected in practice usually are not exact. For various reasons it is often suggested in many applications to replace the original data matrix by some lower dimensional representation obtained via subspace approximation or truncation. The truncated singular value decomposition, for example, is one of the most commonly used representations. This note attempts to shed some light on the statistical meaning of this lower dimensional representation.

Key words. data matrix, minimum-variance estimate, truncated singular value decomposition, subspace method, low rank approximation.

1. Introduction.

An indispensable task in almost every discipline of sciences is to analyze a certain data to search for relationships between a set of exogenous and endogenous variables. The exigencies of such a task becomes especially strong in this era of information and digital technologies as massive amounts of data are being generated at almost every level of applications.

It is generally acknowledged that most of the information gathering devices or methods at present have only finite bandwidth. One thus cannot avoid the fact that the data collected often are not exact. For example, signals received by antenna arrays often are contaminated by instrumental noises; astronomical images acquired by telescopes often are blurred by atmospheric turbulence; database prepared by document indexing often are biased by subjective judgment; and even empirical data obtained in laboratories often do not satisfy intrinsic physical constraints. Before any deductive sciences can further be applied, it is important to first reconstruct the data matrices so that the inexactness is reduced while certain feasibility conditions are satisfied.

Furthermore, in many situations the data observed from complex phenomena represent the integrated result of several interrelated variables acting together. When these variables are less precisely defined, the actual information contained in the original data matrix might be overlapping, fuzzy, and no longer that clearly cut. A reduced system might provide as well the same level of fidelity as the original system.

One common ground in the various approaches for noise removal, model reduction, feasibility reconstruction, and so on, is to replace the original data matrix by a lower dimensional representation obtained somehow via subspace approximation or truncation. The truncated singular value decomposition, for example, is

*Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205. This research was supported in part by the National Science Foundation under grant DMS-9803759.

one commonly used candidate for replacement. Despite the many reported successes in application and the many seemingly intuitive arguments to support this approach, there appears to be a lack of rigorous mathematics to justify exactly what is really going on behind this low rank approximation. This short note is an attempt to fill that gap from a statistical point of view.

2. From a Random Variable Point of View. We first consider a general random (column vector) variable $\mathcal{X}$ in R^n with a certain unspecified distribution. Let $\mathcal{E}[\mathcal{X}]$ denote the expected value of $\mathcal{X}$. Typically, $\text{cov}(\mathcal{X}) := \mathcal{E}[(\mathcal{X} - \mathcal{E}[\mathcal{X}])(\mathcal{X} - \mathcal{E}[\mathcal{X}])^T] \in R^{n \times n}$ is defined as the *covariance matrix* of $\mathcal{X}$. Being symmetric and positive semi-definite, the deterministic matrix $\text{cov}(\mathcal{X})$ enjoys a spectral decomposition

$$\text{cov}(\mathcal{X}) = \sum_{j=1}^{n} \lambda_j \boldsymbol{p}_j \boldsymbol{p}_j^T \tag{2.1}$$

where we also assume that eigenvalues are arranged in the descending order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Observe that $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ form an orthonormal basis for R^n. Express the random column variable $\mathcal{X}$ as

$$\mathcal{X} = \sum_{j=1}^{n} (\boldsymbol{p}_j^T \mathcal{X}) \boldsymbol{p}_j. \tag{2.2}$$

Note that the columns in the matrix $P := [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n]$ are deterministic vectors themselves. The randomness of $\mathcal{X}$ therefore must come solely from the randomness of each coefficient in (2.2). The following observation sheds important insight on the portion of randomness of $\mathcal{X}$ in each of the eigenvector direction.

THEOREM 2.1. *Let $\boldsymbol{\alpha} := P^T \mathcal{X}$. Then $\boldsymbol{\alpha}$ is a random variable whose components are mutually stochastically independent. Indeed,*

$$\mathcal{E}[\boldsymbol{\alpha}] = P^T \mathcal{E}[\mathcal{X}], \tag{2.3}$$

$$\mathit{var}(\boldsymbol{\alpha}) = \mathit{diag}\{\lambda_1, \ldots, \lambda_n\}. \tag{2.4}$$

Proof. The expected value (2.3) of $\boldsymbol{\alpha}$ is obvious. The covariance matrix of $\boldsymbol{\alpha}$ is given by

$$\begin{aligned} \text{cov}(\boldsymbol{\alpha}) &= \mathcal{E}[(\boldsymbol{\alpha} - \mathcal{E}[\boldsymbol{\alpha}])(\boldsymbol{\alpha} - \mathcal{E}[\boldsymbol{\alpha}])^T] \\ &= \mathcal{E}[(P^T(\mathcal{X} - \mathcal{E}[\mathcal{X}])(\mathcal{X} - \mathcal{E}[\mathcal{X}])^T P)] = \text{diag}\{\lambda_1, \ldots, \lambda_n\} \end{aligned}$$

where the last equality follows from the definition of (2.1). □

Recall that smaller variance means the values of the random variable are more clustered around the mean. It is fair to say that a random variable with larger variance is harder to *predict* than a random variable with smaller variance.

From Theorem 2.1, we make one important observation. That is, the larger the eigenvalue λ_j of $\mathrm{cov}(\mathcal{X})$ is, the larger the variance of the random (scalar) variable α_j is. Consider the fact from (2.2) that the random variable $\mathcal{X}$ is made of random contributions from each of the n directions $\boldsymbol{p}_j$, $j = 1, \ldots, n$. Consider also that the contribution from each direction is governed independently by the distribution of the corresponding random variable α_j. Thus, the less the variance of α_j is, the less unpredictability of the contribution from the direction $\boldsymbol{p}_j$. As far as the random nature of $\mathcal{X}$ is concerned, it is intuitively correct from a statistical point of view that those coefficients α_j with larger variances should represent a more integral part in the stochastic nature of $\mathcal{X}$. It is in this context that we may *rank* the importance of corresponding eigenvectors $\boldsymbol{p}_j$ as *essential* components for the variable $\mathcal{X}$ according to the magnitude of λ_j.

If it becomes desirable to approximate the random variable $\mathcal{X}$ by another unbiased yet *simpler* variable $\hat{\mathcal{X}}$, we see from Theorem 2.1 that $\hat{\mathcal{X}}$ had better capture those components corresponding to larger λ_j in the expression (2.2). We quantify this notion below that provides the basic idea of truncation.

So that this note is self-contained, we first reprove a useful result that is classical in estimation theory [1, 2].

THEOREM 2.2. *Let $\boldsymbol{x} \in R^n$ and $\boldsymbol{y} \in R^r$ denote two random variables with mean zero, respectively. Then the coefficient matrix $K \in R^{n\times r}$ that gives rise to the best unbiased linear estimation $\hat{\boldsymbol{x}} = K\boldsymbol{y}$ of $\boldsymbol{x}$ in the sense of minimizing $\mathcal{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2]$ is*

$$K = \mathcal{E}[\boldsymbol{x}\boldsymbol{y}^T](\mathcal{E}[\boldsymbol{y}\boldsymbol{y}^T])^{-1}. \tag{2.5}$$

In this case, each $\hat{x}_i$ is the minimum-variance estimate of the corresponding x_i, respectively, for $i = 1, \ldots, n$.

Proof. Let the matrix K be written in rows, i.e., $K = [\boldsymbol{k}_1^T, \ldots \boldsymbol{k}_n^T]^T$. Observe that

$$\mathcal{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] = \sum_{i=1}^{n} \mathcal{E}[(x_i - \hat{x}_i)^2] = \sum_{i=1}^{n} \mathcal{E}[(\boldsymbol{k}_i^T\boldsymbol{y} - x_i)^2].$$

Thus it suffices to consider minimizing each individual term $g(\boldsymbol{k}_i) := \mathcal{E}[(\boldsymbol{y}^T\boldsymbol{k}_i - x_i)^2]$, $i = 1, \ldots n$, in the above summation. (It is in this sense that the term "minimum-variance" unbiased estimate for each component is used.) The first order optimality condition for g to be minimized at $\boldsymbol{k}_i$ is

$$\nabla g(\boldsymbol{k}_i) = 2\mathcal{E}[(\boldsymbol{y}^T\boldsymbol{k}_i - x_i)\boldsymbol{y}] = 0. \tag{2.6}$$

We may rewrite the necessary condition for $i = 1, \ldots n$ collectively as

$$\mathcal{E}[\boldsymbol{y}\boldsymbol{y}^T]K^T = \mathcal{E}[\boldsymbol{y}\boldsymbol{x}^T].$$

It follows that K is given by (2.5). □

Corollary 2.3. *Assume that $\hat{\boldsymbol{x}}$ is the minimum-variance estimate of $\boldsymbol{x}$ under the same setting as in the above theorem. Then*

$$cov(\boldsymbol{x} - \hat{\boldsymbol{x}}) = cov(\boldsymbol{x}) - cov(\hat{\boldsymbol{x}}), \tag{2.7}$$

$$\mathcal{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] = trace(cov(\boldsymbol{x} - \hat{\boldsymbol{x}})) = \mathcal{E}[\boldsymbol{x}^T\boldsymbol{x}] - \mathcal{E}[\boldsymbol{x}^T K\boldsymbol{y}]. \tag{2.8}$$

Proof. By definition,

$$\mathrm{cov}(\boldsymbol{x} - \hat{\boldsymbol{x}}) = \mathrm{cov}(\boldsymbol{x}) - \mathcal{E}[\boldsymbol{x}\hat{\boldsymbol{x}}^T] - \mathcal{E}[\hat{\boldsymbol{x}}\boldsymbol{x}^T] + \mathrm{cov}(\hat{\boldsymbol{x}}).$$

Observer then by substitution that

$$\mathcal{E}[\boldsymbol{x}\hat{\boldsymbol{x}}^T] = \mathcal{E}[\hat{\boldsymbol{x}}\boldsymbol{x}^T] = \mathcal{E}[\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^T] = \mathcal{E}[\boldsymbol{x}\boldsymbol{y}^T](\mathcal{E}[\boldsymbol{y}\boldsymbol{y}^T])^{-1}\mathcal{E}[\boldsymbol{y}\boldsymbol{x}^T].$$

The equation (2.7) is proved. The residual (2.8) can be calculated from

$$\mathcal{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] = \sum_{i=1}^{n} g(\boldsymbol{k}_i) = -\sum_{i=1}^{n} \mathcal{E}[(\boldsymbol{k}_i^T\boldsymbol{y} - x_i)x_i]$$

by using (2.6). □

Returning to the problem of approximating the random variable $\mathcal{X}$ by an unbiased yet simpler variable $\hat{\mathcal{X}}$, consider the case that by a simpler variable $\tilde{\mathcal{X}}$ we mean a random variable limited to a *lower dimensional* subspace. Our goal then is to find a proper subspace $\mathcal{S}$ and a particular random variable $\tilde{\mathcal{X}}$ on $\mathcal{S}$ such that $\mathcal{E}[\|\mathcal{X} - \tilde{\mathcal{X}}\|^2]$ is minimized.

Observe first that, given any r-dimensional subspace $\mathcal{S}$, there exists a matrix $K \in R^{n\times r}$ such that columns of the matrix product PK, with P given by (2.1), form a basis for $\mathcal{S}$. Any unbiased random variable $\tilde{\mathcal{X}}$ restricted to $\mathcal{S}$ can then be expressed in the form

$$\tilde{\mathcal{X}} = PK\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ stands for a certain (column) random variable in R^r. We may further assume that components in $\boldsymbol{\beta}$ are mutually independent because, if otherwise, we may simply do a spectral decomposition of $\boldsymbol{\beta}$ similarly to (2.1) and Theorem 2.1. It follows that $\mathcal{E}[\|\mathcal{X} - \tilde{\mathcal{X}}\|^2] = \mathcal{E}[\|\boldsymbol{\alpha} - K\boldsymbol{\beta}\|^2]$. The minimum-variance problem is now reduced to the problem of finding K and $\boldsymbol{\beta}$ so that $\mathcal{E}[\|\boldsymbol{\alpha}-K\boldsymbol{\beta}\|^2]$ is minimized.

From Theorem 2.2, however, we know that at an optimizer the coefficient matrix K and the variable $\boldsymbol{\beta}$ are not totally unrelated. Indeed, given $\boldsymbol{\beta}$, the optimal matrix K is completely determined and is given by

$$K = \mathcal{E}[\boldsymbol{\alpha}\boldsymbol{\beta}^T](\mathcal{E}[\boldsymbol{\beta}\boldsymbol{\beta}^T])^{-1}. \tag{2.9}$$

From (2.8), we further know that to obtain the minimum-variance approximation of $\mathcal{X}$, it only remains to choose $\boldsymbol{\beta}$ so that

$$\mathcal{E}[\boldsymbol{\alpha}^T K\boldsymbol{\beta}] = \left\langle \mathcal{E}[\boldsymbol{\alpha}\boldsymbol{\beta}^T](\mathcal{E}[\boldsymbol{\beta}\boldsymbol{\beta}^T])^{-1}, \mathcal{E}[\boldsymbol{\alpha}\boldsymbol{\beta}^T]\right\rangle \tag{2.10}$$

is maximized. This nonlinear optimization problem turns out to have a simple solution as we shall see from the proof of the following theorem.

Theorem 2.4. *Suppose that $\mathcal{X}$ is a random variable in R^n with mean zero and that its covariance matrix has a spectral decomposition given by (2.1). Then among all unbiased variables restricted to any r-dimensional subspaces in R^n, the random variable*

$$\hat{\mathcal{X}} := \sum_{j=1}^{r}(\boldsymbol{p}_j^T\mathcal{X})\boldsymbol{p}_j \tag{2.11}$$

is the best linear minimum-variance estimate of $\mathcal{X}$ in the sense that $\mathcal{E}[\|\mathcal{X} - \hat{\mathcal{X}}\|^2]$ is minimized.

Proof. We already know that $\mathcal{E}[\|\mathcal{X} - \tilde{\mathcal{X}}\|^2] = \mathcal{E}[\|\boldsymbol{\alpha} - K\boldsymbol{\beta}\|^2]$. From Corollary 2.3, we also know that

$$\mathcal{E}[\|\boldsymbol{\alpha} - K\boldsymbol{\beta}\|^2] = \operatorname{trace}(\operatorname{cov}(\boldsymbol{\alpha})) - \operatorname{trace}(\operatorname{cov}(K\boldsymbol{\beta}))$$

is minimized. In fact, in the proof of Theorem 2.2, we have pointed out that the estimation $\hat{\boldsymbol{\alpha}} = K\boldsymbol{\beta}$ is a component-wise minimum-variance estimation. That is, for each $i = 1, \ldots n$, the coefficient matrix K has the effect that

$$\mathcal{E}[(\alpha_i - \hat{\alpha}_i)^2] = \mathcal{E}[\alpha_i^2] - \mathcal{E}[\hat{\alpha}_i^2]$$

is minimized. Recall that $\mathcal{E}[\alpha_i^2] = \lambda_i$. Thus, we should somehow select $\boldsymbol{\beta}$ in such a way so that the corresponding $\hat{\boldsymbol{\alpha}} = K\boldsymbol{\beta}$ will have $\mathcal{E}[\hat{\alpha}_i^2] = \lambda_i$ for as many i's as possible. More specifically, since the trace is to sum over the differences $\lambda_i - \mathcal{E}[\hat{\alpha}_i^2]$ whereas $\lambda_1 \geq \ldots \geq \lambda_n$ and since we only have r degrees of freedom to determine $\boldsymbol{\beta}$, the best we can hope is to choose $\boldsymbol{\beta}$ so that the first k eigenvalues $\lambda_1, \ldots, \lambda_r$ are matched.

It turns out that if we choose the special case $\boldsymbol{\beta} = [\alpha_1, \ldots, \alpha_r]^T$, then

$$\hat{\boldsymbol{\alpha}} = [\alpha_1, \ldots, \alpha_r, 0, \ldots, 0]^T$$

and the corresponding $PK\hat{\boldsymbol{\beta}}$ is precisely given by (2.11). □

It is important to note that in the above linear minimum-variance estimation, the variable $\mathcal{X}$ is *centered* at zero. If $\mathcal{X}$ is not centered at zero, the expression for truncation would be much more complicated. Somehow this centering has been ignored in many practices where low rank approximation is used. Without the centering, we really would like to raise the flag that the resulting truncated data should suffer from the loss of some significant statistical meanings. We shall comment more on this in the next section.

On the other hand, the following theorem shows that the choice of $\hat{\mathcal{X}}$ not only makes the diagonal entries of $\operatorname{cov}(\hat{\mathcal{X}})$ best approximate those of $\operatorname{cov}(\mathcal{X})$, but that the entire matrix $\operatorname{cov}(\hat{\mathcal{X}})$ be reasonably close to $\operatorname{cov}(\mathcal{X})$ as well.

THEOREM 2.5. *Suppose that $\mathcal{X}$ is a random variable in R^n with mean zero and that its covariance matrix has a spectral decomposition given by (2.1). Then among all unbiased variables restricted to any r-dimensional subspaces in R^n, the random variable $\hat{\mathcal{X}}$ defined in (2.11) also minimizes $\|cov(\hat{\mathcal{X}}) - cov(\mathcal{X})\|$.*

Proof. The proof is quite straightforward. It is well known that the best rank r approximation to the matrix $\operatorname{cov}(\mathcal{X})$ is given by the truncated summation $\sum_{i=1}^{r} \lambda_i \boldsymbol{p}_i \boldsymbol{p}_i^T$, which clearly is also the covariance matrix of $\hat{\mathcal{X}}$. □

3. From a Random Sample Point of View. The discussion thus far is based on the fact that the random variable $\mathcal{X}$ is completely known. Such an assumption is not realistic in practice since often the probability distribution function of the underlying random variable $\mathcal{X}$ is not a priori known. One common practice in application then is to simulate the random variable $\mathcal{X}$ by a collection of ℓ random samples. These samples are recorded in a $n \times \ell$ matrix X. Each column of X represents one random sample of the underlying random (column vector) variable $\mathcal{X} \in R^n$. It is known that when ℓ is large enough, many of the stochastic properties of $\mathcal{X}$ can be recouped from X.

The question now is how to retrieve a sample data matrix from X to represent the minimum-variance approximation $\hat{\mathcal{X}}$ of $\mathcal{X}$. To begin with, we shall assume that $\mathcal{E}[\mathcal{X}] = 0$ and that the samples X has been centered, i.e., the mean value of each row is zero. The connection lies in the observations that the covariance matrix of the samples

$$R = \frac{XX^T}{\ell}$$

converges to $\operatorname{cov}(\mathcal{X})$ by the law of large numbers. Analogous to (2.1), let

$$R = \sum_{i=1}^{n} \mu_i \boldsymbol{u}_i \boldsymbol{u}_i^T \tag{3.1}$$

be the spectral decomposition of R with eigenvalues $\mu_1 \geq \ldots \geq \mu_n$ and orthonormal eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$. Then it follows from the observation in Theorem 2.4 that the best low dimensional minimum-variance estimate $\hat{\mathcal{X}}$ to $\mathcal{X}$ should be represented by the matrix

$$\hat{X} := \sum_{j=1}^{r} \boldsymbol{u}_j (\boldsymbol{u}_j^T X). \tag{3.2}$$

The *low dimension* estimate $\hat{\mathcal{X}}$ to the (continuous) random variable $\mathcal{X}$ is now comfortably translated into a *low rank* approximation $\hat{X}$ to the (discrete) random sample matrix X.

Indeed, the singular value decomposition of X

$$X = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{3.3}$$

shares the same eigenvectors of R as its left singular vectors, i.e., $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ with singular values given by $\sigma_i = \sqrt{\ell \mu_i}$, $i = 1, \ldots n$, respectively. The notion of the truncated singular value decomposition of X is simply the partial sum $\sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$, which we now see is precisely $\hat{X}$ defined in (3.2).

In this sense, the truncated singular value decomposition of a give data matrix X representing random samples of an unknown random variable $\mathcal{X}$ now has a statistical meaning. That is, the truncated rank r singular value decomposition represents random samples of the best minimum-variance linear estimate $\hat{\mathcal{X}}$ to $\mathcal{X}$ among all possible r-dimensional subspaces.

4. Conclusion. In many applications, the truncated singular value decomposition of the observed data matrix is used to *filter* out the *noises*. In this note, we try to convey the notion that the truncation singular value decomposition is in fact the best minimum-variance estimation of the underlying *unknown* random variable, be it contaminated by noises or not. Note that in Theorem 2.2 no relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ is assumed. Likewise, no relationship between $\hat{\mathcal{X}}$ and $\mathcal{X}$ is assumed prior to the conclusion proved in Theorem 2.4. The notion of *truncation* now is manifested through the notion of best minimum-variance estimation.

Although it is obvious in the context of linear algebra that the truncated rank r singular value decomposition $\hat{X}$ of a given matrix X minimizes the 2-norm or Frobenius norm of the difference $X - Y$ among all possible rank r matrices Y, one must wonder what this low rank approximation means if X is a random matrix (from any kind of unknown distribution). If each column of X represents an *unpredictable* sample of a certain unknown distribution, one must wonder how much fidelity the corresponding column in the truncated matrix $\hat{X}$ really represents and how to measure it. The statistical interpretation of truncated singular value decomposition discussed in this note should fill that gap. The truncated singular value decomposition $\hat{X}$ not only is the best approximation to X in the sense of norm, but also is the closest approximation to X in the sense of statistics. It maintains the most significant stochastic portion of the original data matrix X.

Finally, other than the truncated singular value decomposition, many other types of low rank approximation to the given data matrix X have been proposed. For example, one of the most contentious issues in the latent semantic indexing (LSI) for data mining is to find a suitable low rank representation of the original term-document indexing matrix. For this issue alone, many parties are in fierce competition to patent their special techniques. In this note, we have shown the significance of those larger singular values and the corresponding left singular vectors. Generally speaking, any lower rank approximation to an empirical data matrix X should carry properties similar, if not identical, to the truncated singular value decomposition, i.e., should contain substantial stochastic information about the original random variable $\mathcal{X}$. It perhaps is not too judgemental to say that any low rank approximation (for data mining) without this notion in mind, regardless how efficient the computation could be, is equivalent to an attempt to *see things through the glass of darkness.*

REFERENCES

[1] D. G. Luenberger, Optimization by Vector Space Methods, John Wiley & Sons, New York, 1968.

[2] J. L. Melsa and D. L. Cohn, Decision and Estimation Theory, McGrow-Hill, New York, 1978.