SVD and its Application to Generalized Eigenvalue Problems

Thomas Melzer

May 28, 2003

Contents

0.1 Singular Value Decomposition

Singular value decomposition (SVD) is an extremely powerful and useful tool in Linear Algebra. In this appendix, we will only give the formal definition of SVD and discuss some of its more important properties. For a more comprehensive numerical discussion see, for example, [3] and [4]; [4] gives also a complete implementation of the SVD-algorithm in the C-programming language.

SVD decomposes a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ into the product of two orthonormal matrices, $\mathbf{U} \in \mathbb{R}^{m \times m}$, $\mathbf{V} \in \mathbb{R}^{n \times n}$, and a pseudo-diagonal matrix $\mathbf{D} = diag(\sigma_1, .., \sigma_\rho) \in \mathbb{R}^{m \times n}$, with $\rho = min(m, n)$ (i.e., all components except the first ρ diagonal components are zero), such that

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{1}$$

Any matrix can be decomposed in this fashion. If we denote by $\mathbf{u}_i$ and $\mathbf{v}_i$ the columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, we can rewrite Eq. 1 as weighted sum of corresponding outer products

$$\mathbf{A} = \sum_{i=1}^{\rho} \sigma_i \mathbf{u}_i \mathbf{v}_i^T. \tag{2}$$

By convention, the diagonal elements σ_i of $\mathbf{D}$ - which are referred to as **singular values** - are non-negative[1] and sorted in decreasing order, i.e. $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_\rho \geq 0$; this convention makes the SVD unique except when one or more singular values occur with multiplicity greater one (in which case the corresponding columns of $\mathbf{U}$ and $\mathbf{V}$ can be replaced by linear combinations of themselves, see [4]). Let us further denote by ρ_0 the index of the last non-zero singular value, i.e., $\sigma_i = 0$ for $i > \rho_0$ and $\sigma_i > 0$ for $i \leq \rho_0$.

Depending on the values of m and n, the decomposition depicted by Eq. 1 is often given in a more compact manner; in particular, if the number of rows of $\mathbf{A}$ is greater than or equal to its number of columns, i.e., $m > n$, the last $m - n$ columns of $\mathbf{U}$ and the last (all zero!) $m - n$ rows of $\mathbf{D}$ can be omitted. Hence, in this "economy size" decomposition,$\mathbf{U} \in \mathbb{R}^{m \times n}$ will be a column-orthonormal matrix (i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ will still hold), and $\mathbf{D} \in \mathbb{R}^{n \times n}$ a diagonal matrix, respectively.

[1] If σ_i is negative, we can make it positive by simply multiplying either the ith column of $\mathbf{U}$ or the ith column of $\mathbf{V}$ by -1.

0.1.1 Range and Nullspace

An important property of SVD is that it explicitly constructs orthonormal bases for the range and the nullspace of $\mathbf{A}$. This follows directly from the decomposition given in Eq. 2 and the orthonormality of $\mathbf{U}$ and $\mathbf{V}$:

- All vectors $\mathbf{b}$ in the **range** of $\mathbf{A}$ ($range(\mathbf{A})$), i.e., all vectors $\mathbf{b} = \mathbf{Ax}$ that can be obtained as a linear combination of the columns of $\mathbf{A}$, lie in the span of the first ρ_0 columns of $\mathbf{U}$ (the columns with corresponding non-zero singular values). On the other hand, each of the first ρ_0 columns of $\mathbf{U}$ lies in the range of $\mathbf{A}$ (since $\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{Av}_i$ for $1 \leq i \leq \rho_0$). Thus, $\mathbf{u}_1..\mathbf{u}_{\rho_0}$ form an orthonormal basis of $range(\mathbf{A})$.
- All vectors $\mathbf{w}$ in the **nullspace** of $\mathbf{A}$, i.e., all vectors satisfying $\mathbf{Ax} = \mathbf{0}$, must be orthogonal to the first ρ_0 columns of $\mathbf{V}$ (these are the rows of $\mathbf{V}^T$ with corresponding non-zero singular values); otherwise, for some $1 \leq i \leq \rho_0$, we have $\sigma_i\mathbf{u}_i(\mathbf{v}_i^T\mathbf{w}) \neq \mathbf{0}$, which, together with the orthogonality of the $\mathbf{u}_i$, implies $\mathbf{Ax} \neq \mathbf{0}$. It follows that $\mathbf{w}$ must lie in the subspace spanned by the last $n - \rho_0$ columns of $\mathbf{V}$, which themselves lie in the nullspace of $\mathbf{A}$. Thus, $\mathbf{v}_{\rho_0+1}..\mathbf{v}_n$ form an orthonormal basis of the nullspace of $\mathbf{A}$.

It is clear from the discussion above that SVD can be used to find an orthonormal basis spanning the linear subspace induced by any n linearly independent vectors $\mathbf{a}_i \in \mathbb{R}^m, n \leq m$: if one computes the SVD of the matrix whose columns are the $\mathbf{a}_i$, the solution is given by the first n column vectors of $\mathbf{U}$. If, however, $\mathbf{D}$ contains less than n non-zero singular values, i.e, $\rho_0 < \rho = n$, the $\mathbf{a}_i$ are linearly dependent and only the first ρ_0 column vectors of $\mathbf{U}$ should be retained.

Although the task of constructing an orthonormal basis has historically been addressed by *Gram-Schmidt-orthonormalization*, SVD should be preferred because it is numerically far more stable.

0.1.2 Rank, Condition Number and Matrix Inversion

The **rank** of a matrix $\mathbf{A}$ is defined as the number of its linearly independent rows (columns), or, equivalently, as the dimensionality of the linear subspace spanned by its rows (columns). In particular, we have

$$rank(\mathbf{A}) = dim(range(\mathbf{A})) = \rho_0. \tag{3}$$

The dimensionality of the the nullspace of $\mathbf{A}$ is also referred to as its **nullity** ($nullity(\mathbf{A})$); for quadratic matrices $\mathbf{A} \in \mathbb{R}^{m \times m}$, it holds that

$$rank(\mathbf{A}) + nullity(\mathbf{A}) = m. \tag{4}$$

A quadratic matrix $\mathbf{A}$ with $rank(\mathbf{A}) = m$, i.e., $\rho = \rho_0$ is said to be non-singular or to have **full rank**; in particular, a quadratic matrix is invertible *iff* it is non-singular. This is clearly a very desirable property, since it implies that all linear equations

$$\mathbf{b} = \mathbf{A}\mathbf{w} \tag{5}$$

yield a unique solution, namely $\mathbf{w} = \mathbf{A}^{-1}\mathbf{b}$. The inversion itself can easily be computed using SVD; since both $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices, it holds that $\mathbf{U}^{-1} = \mathbf{U}^T$ and $\mathbf{V}^{-1} = \mathbf{V}^T$, and thus we have

$$\mathbf{A}^{-1} = (\mathbf{V}^T)^{-1}\mathbf{D}^{-1}\mathbf{U}^{-1} = \mathbf{V}diag(1/\sigma_1, .., 1/\sigma_\rho)\mathbf{U}^T. \tag{6}$$

While the inversion of $\mathbf{U}$ and $\mathbf{V}^T$ is trivial, the matrix $\mathbf{D}$ might contain zero singular values (i.e., $\rho > \rho_0$), in which case the matrix $\mathbf{A}$ itself is singular and the corresponding diagonal entries in $\mathbf{D}^{-1}$ become infinite. In practice, even very small (but non-zero) singular values will cause $\mathbf{A}$ to become numerically singular; in fact, the relevant quantity here is the ratio of the largest to the smallest singular value, σ_1/σ_p, which is referred to as the **condition number** of $\mathbf{A}$. If the condition number becomes too large, i.e., when its reciprocal approaches the machine precision ϵ, $\mathbf{A}$ is said to be **ill-conditioned**. In particular, when we have one or more zero singular values, the condition number becomes infinite. In practice, the singular values can be used to compute the effective rank ρ_e ($\leq \rho_0$) of $\mathbf{A}$ w.r.t. a given threshold ϵ_e as the index of the smallest singular value for which $\sigma_i/\sigma_1 > \epsilon_e$ still holds, i.e., $\sigma_i/\sigma_1 > \epsilon_e$ for $i \leq \rho_e$ and $\sigma_i/\sigma_1 \leq \epsilon_e$ for $i > \rho_e$. ϵ_e is typically chosen several orders of magnitude larger than ϵ, but the concrete value is problem dependent and will normally have to be determined empirically.

0.1.3 Equation Solving and Linear Least Squares

As shown in the previous section, SVD can be used to solve quadratic linear systems, provided the coefficient matrix $\mathbf{A}$ is non-singular. If, however, $\mathbf{A}$ is singular, the system will no longer yield a unique solution (since we can add any vector from the - non-empty - nullspace of $\mathbf{A}$ to a particular solution $\mathbf{w}_0$) or, worse, no solution at all (if $\mathbf{b} \notin range(\mathbf{A})$).

Both issues can be addressed by discarding too small singular values, i.e., by replacing the diagonal entries $1/\sigma_i$ in $\mathbf{D}^{-1}$ (cf. Eq. 6) with 0 for $i > \rho_e$, resulting in a modified matrix inverse

$$\hat{\mathbf{A}}^{-1} = \mathbf{V} diag(1/\sigma_1, .., 1/\sigma_{\rho_e}, 0, ..0)\mathbf{U}^T, \tag{7}$$

solution vector

$$\mathbf{w} = \hat{\mathbf{A}}^{-1}\mathbf{b} \tag{8}$$

and reconstruction

$$\hat{\mathbf{b}} = \mathbf{A}\mathbf{w}. \tag{9}$$

This operation has several effects:

- By discarding the singular values with index $i > \rho_0$, we restrict the inversion to the range of $\mathbf{A}$. Let c_i denote the projection of $\mathbf{b}$ onto the ith column of $\mathbf{U}$, i.e, $c_i = \mathbf{u}_i^T\mathbf{b}$. Then we have (cf. Eqs. 7, 8)

  $$\mathbf{w} = \sum_{i=1}^{\rho_0} \mathbf{v}_i c_i/\sigma_i$$

  and, due to the orthonormality of the $\mathbf{v}_i$,

  $$\hat{\mathbf{b}} = \sum_{i=1}^{\rho_0} \mathbf{u}_i c_i$$

  (cf. Eq. 9); thus, the reconstruction $\hat{\mathbf{b}}$ is just the projection of $\mathbf{b}$ onto the linear subspace spanned by the first ρ_0 columns of $\mathbf{U}$ (i.e., the range of $\mathbf{A}$). Analytically, this approach yields the minimum residual error in the least squares sense. In practice, additionally discarding small, but non-zero singular values (with index $i > \rho_e$) will normally yield a numerically more robust solution with smaller effective residual error $\|\mathbf{b} - \hat{\mathbf{b}}\|$ (see below).

- By zeroing the inverse of very small singular values, we discard quantities that are dominated by roundoff-error. This is based solely on a numerical argument: analytically, the contribution of small singular values should not be neglected, as long as they are non-zero. From a numerical point of view, however, when the relative magnitude of a singular value approaches machine precision, even the first few significant digits are likely to consist only of roundoff-error. Thus, zeroing out these values will normally lead to a better solution $\mathbf{w}$ with smaller residual error [4].

- If $\mathbf{A}$ is singular, the solution will not be unique: the set of all solutions is given by the hyperplane spanned by the last $\rho - \rho_0$ columns of $\mathbf{V}$ and centered at $\mathbf{w}_0$, whereby $\mathbf{w}_0$ is a particular solution. However, Eq. 8 will always yield the solution with the smallest norm (for a proof, see [4]).

The discussion above carries over without modification to the general case of rectangular matrices. In particular, if $\mathbf{A}$ corresponds to the coefficient matrix of an overdetermined system of linear equations, we can use SVD to obtain the linear least squares solution of this system in a numerically stable fashion.

Consider the following quadratic (least squares) optimization problem: minimize

$$\|\mathbf{A}\mathbf{w} - \mathbf{b}\| \tag{10}$$

for given $\mathbf{A} \in \mathbb{R}^{m \times n}, m > n$, and $\mathbf{b} \in \mathbb{R}^m$. The gradient of Eq. 10 w.r.t. $\mathbf{w}$ is obtained as

$$2\mathbf{A}^T(\mathbf{A}\mathbf{w} - \mathbf{b}). \tag{11}$$

A necessary (and, due to the convexity of Eq. 10, also sufficient) condition for optimality is obtained by setting the gradient to $\mathbf{0}$:

$$\mathbf{A}^T\mathbf{A}\mathbf{w} = \mathbf{A}^T\mathbf{b}. \tag{12}$$

Eq. 12 can also be regarded as a set of n coupled linear equations in $\mathbf{w}$, known as the *normal equations* of the least squares problem Eq. 10. Assuming that $\mathbf{A}^T\mathbf{A}$ is non-singular, we obtain the solution $\mathbf{w}^*$ as:

$$\mathbf{w}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}, \tag{13}$$

whereby the quantity

$$\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T \in \mathbb{R}^{n \times m} \tag{14}$$

is known as the *pseudo-inverse* of $\mathbf{A}$. The pseudo-inverse can be regarded as a generalization of matrix inversion to non-square matrices; indeed, it holds that $\mathbf{A}^\dagger\mathbf{A} = \mathbf{I}$ (however, $\mathbf{A}\mathbf{A}^\dagger = \mathbf{I}$ is not true in general). In practice, direct computation of the pseudo-inverse is prone to numerical instability; furthermore, it can not be computed at all when - e.g., due to column degeneracies - the rank of $\mathbf{A}$ becomes less than n (in which case $\mathbf{A}^T\mathbf{A}$ will be singular). If, on the other hand, we use SVD to "invert" the coefficient matrix $\mathbf{A}$ (cf.

Eq. 7), we do not only get a clear diagnosis of the numerical situation, but we are also able to deal with numerical roundoff-errors (by zeroing small singular values) and with column degeneracies which would otherwise cause $\mathbf{A}^T\mathbf{A}$ to become singular.

The formal equivalence between SVD-inversion and pseudo-inversion, provided that $rank(\mathbf{A}) = n$, can easily be seen by replacing $\mathbf{A}$ with its SVD-decomposition in Eq. 14. We have

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

$$\mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^2\mathbf{V}^T \tag{15}$$

and, finally

$$\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T\mathbf{V}\mathbf{D}\mathbf{U}^T = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T. \tag{16}$$

In the derivation of Eq. 15 and Eq. 16, we made again use of the orthonormality of $\mathbf{U}$ and $\mathbf{V}$.

Note that if $rank(\mathbf{A}) < n$ (i.e., $\rho_0 < \rho$), we can no longer compute the pseudo-inverse according to Eq. 13. In this degenerate case, SVD-inversion is still applicable, but the solution is no longer unique. As in the case of singular quadratic matrices, the solution space is spanned by the last $\rho - \rho_0$ columns of $\mathbf{V}$, and SVD will return the representative with the smallest norm.

0.2 Eigenvalue Decomposition and Symmetric Matrices

Let $\mathbf{A} \in \mathbb{R}^{m\times m}$ be a square matrix. A vector $\mathbf{e} \in \mathbf{C}^m, \mathbf{e} \neq \mathbf{0}$ and a scalar $\lambda \in \mathbf{C}$ fulfilling

$$\mathbf{A}\mathbf{e} = \lambda\mathbf{e} \tag{17}$$

are called an **eigenvector** resp. **eigenvalue** of $\mathbf{A}$. We say also that $\mathbf{e}$ is an eigenvector belonging to the eigenvalue λ. It is easily seen that if $\mathbf{e}$ is an eigenvector belonging to λ, this is also true for all vectors $\{\alpha\mathbf{e} : \alpha \in \mathbb{R}, \alpha \neq 0\}$; thus, an eigenvector is effectively a 1-dimensional subspace of $\mathbf{C}^m$. By reformulating Eq. 17 as

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{e} = \mathbf{0}, \tag{18}$$

it becomes clear that the eigenvectors can also be characterized as the non-trivial (i.e., non-zero) solutions of the homogeneous system of linear equations

given by $\mathbf{A} - \lambda\mathbf{I}$. In order for such non-trivial solutions to exist, however, $\mathbf{A} - \lambda\mathbf{I}$ must be singular, that is, its determinant

$$p_{\mathbf{A}}(\lambda) = det(\mathbf{A} - \lambda\mathbf{I})$$

must be zero. $p_{\mathbf{A}}(\lambda)$ is a mth-order polynomial in λ, which is referred to as the *characteristic polynomial* of $\mathbf{A}$. Thus, analytically, the eigenvalues of $\mathbf{A}$ can be found as the roots of its characteristic polynomial, while the corresponding eigenvectors are obtained by substituting the eigenvalues into Eq. 18. Since the characteristic polynomial of a $m \times m$ matrix is of degree m and thus has exactly m solutions in $\mathbf{C}$, every $m \times m$ matrix has exactly m eigenvalues. In general, we will obtain not only real, but also complex solutions (both for the eigenvectors and the eigenvalues). Also some of the eigenvalues may occur with multiplicity greater one or be zero. Zero-eigenvalues are special in the sense that they can occur only for rank-deficient and thus non-invertible matrices: as can be easily be seen from Eq. 17, their associated (non-zero) eigenvectors must lie in the nullspace of the matrix $\mathbf{A}$. In the case of eigenvalues λ_m occurring with multiplicity greater one, we can replace any eigenvectors $\mathbf{e}_{m_k}$ belonging to λ_m with linear combinations of themselves: if $\mathbf{e}_{m_i}, \mathbf{e}_{m_j}$ have the same associated eigenvalue λ_m, then $\mu_i\mathbf{e}_{m_i} + \mu_j\mathbf{e}_{m_j}$ is also an eigenvector belonging to λ_m for all $\mu_i, \mu_j \in \mathbb{R}$ (this follows directly from the linearity of the "matrix operator" $\mathbf{A}$). The number of linearly independent eigenvectors belonging to an eigenvalue λ_m is upper-bounded by the multiplicity of λ_m, $mul(\lambda_m)$; in general, however, there is no guarantee that $mul(\lambda_m)$ linearly independent eigenvectors do exist.

0.2.1 Eigenvalue Decomposition of a Square Matrix

Let $\mathbf{E} \in \mathbb{R}^{m \times m}$ be the matrix whose columns are the eigenvectors of $\mathbf{A}$, i.e. $\mathbf{E} = (\mathbf{e}_1, .., \mathbf{e}_m)$, and $\Lambda = diag(\lambda_1, .., \lambda_m)$ a diagonal matrix holding the corresponding eigenvalues. Then, according to Eq. 17, it holds that

$$\mathbf{AE} = \mathbf{E}\Lambda. \tag{19}$$

Let us further assume that $\mathbf{A}$ has m linearly independent eigenvectors. In this case, $\mathbf{E}$ is invertible, and, by multiplying Eq.19 by $\mathbf{E}^{-1}$ from the right, we have

$$\mathbf{A} = \mathbf{E}\Lambda\mathbf{E}^{-1}. \tag{20}$$

Eq.20 is called the **eigenvalue decomposition** (EVD) or also *spectral factorization* of $\mathbf{A}$. We point out again that not every square matrix has a spectral factorization.

0.2.2 Eigenvalue Decomposition of a Symmetric Matrix

In this thesis, we are mainly concerned with real, symmetric matrices. A quadratic matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ is called symmetric if it equals its transpose, that is, if $\mathbf{A} = \mathbf{A}^T$. Real, symmetric matrices have some important (and, from the practitioners point of view, quite pleasing) analytical properties w.r.t. their eigenvalue decomposition:

1. The eigenvalues of a real, symmetric matrix are all real.
2. Every real, symmetric matrix has a spectral factorization.
3. The eigenvectors of a real, symmetric matrix belonging to different eigenvalues are orthogonal. But even in the case of eigenvalues occurring with multiplicity greater one, a complete set of m orthogonal eigenvectors can always be found.
4. The inverse of a real, symmetric matrix $\mathbf{A}$ is again a real, symmetric matrix; $\mathbf{A}^{-1}$ has the same eigenvectors as $\mathbf{A}$, but with reciprocal eigenvalues.
5. For real, symmetric matrices, EVD and SVD become equivalent.

The most important point here is the last one, for everything else can be deduced from the equivalence and the properties of the SVD; thus, we will first proof point 5).

Let the SVD of $\mathbf{A}$ be

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{21}$$

As a consequence of $\mathbf{A}$'s symmetry, $\mathbf{U}$ and $\mathbf{V}$ are identical. Thus, we have

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{22}$$

or, expressed as weighted sum of outer products,

$$\mathbf{A} = \sum_{i=1}^{m} \sigma_i \mathbf{u}_i \mathbf{u}_i^T, \tag{23}$$

whereby $\mathbf{u}_i$ denotes the ith column of $\mathbf{U}$ and and σ_i the corresponding singular value[2]. Note that Eq. 22 is effectively a reformulation of Eq. 20 which replaces inversion by transposition; it is a well known fact in linear algebra that these operations are equivalent for orthonormal matrices. From Eq. 23 and the orthonormality of the $\mathbf{u}_i$ it follows that $\mathbf{A}\mathbf{u}_i = \sigma_i \mathbf{u}_i$ for all $1 \leq i \leq m$, i.e., the $\mathbf{u}_i/\sigma_i$ are indeed eigenvector-eigenvalue pairs of $\mathbf{A}$. Since the $\mathbf{u}_i$ are mutually orthonormal (and thus linearly independent) and a $m \times m$ matrix can have at most m linearly independent eigenvectors, this proofs the equivalence[3] of EVD and SVD for the real, symmetric case. Points 1)-4) follow now directly from the properties of SVD discussed in section 0.1.

0.2.3 Autocorrelation and Gram Matrices

Given N observations $\mathbf{x}_i \in \mathbb{R}^p$ arranged in the columns of the sample matrix $\mathbf{X} \in \mathbb{R}^{p \times N} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$. We are interested in the quantities

$$\hat{\mathbf{S}} = \frac{1}{N} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = \frac{1}{N} \mathbf{X}\mathbf{X}^T \in \mathbb{R}^{p \times p} \tag{24}$$

$$\mathbf{G} = (\mathbf{x}_i^T \mathbf{x}_j)_{1 \leq i,j \leq p} = \mathbf{X}^T \mathbf{X} \in \mathbb{R}^{N \times N}. \tag{25}$$

The **sample autocorrelation matrix** is an unbiased estimate of the population autocorrelation matrix $\mathbf{S} = \mathcal{E}[\mathbf{x}\mathbf{x}^T]$, whereas the **Gram matrix G** is obtained by forming all possible inner products between the samples $\mathbf{x}_i \in \mathbf{X}$. There exists a strong relationship between these two matrices, which can be exploited to reformulate algorithms based on second order statistics in terms of Gram - and, subsequently - kernel matrices (for a more elaborate discussion, see [5]). In the following, we will focus on the relation between $\mathbf{G}$ and $N\hat{\mathbf{S}}$ (i.e., we drop the scalar $\frac{1}{N}$ in Eq.24) for the sake of clarity of presentation; in practice, this omission rarely matters, but can easily be compensated for if required.

The matrices $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ contain essentially the same information; in particular

[2]In the special case of symmetric matrices, SVD may actually give negative singular values λ_i, because we have no longer the freedom to multiply either $\mathbf{u}_i$ or $\mathbf{v}_i$ by -1 (since, for symmetric matrices, they must be identical).

[3]The equivalence is not to be confused with identity. Even different implementations of EVD may give different eigenvectors. As discussed above, such ambiguities arise when one or more eigenvalues occur with multiplicity greater one.

- they are both positive semi-definite (i.e., they are symmetric and have non-negative eigenvalues),
- they have identical rank $\rho_0 \leq min(p, N)$
- they share the same non-zero eigenvalues, and
- their respective eigenvectors can be obtained from the eigenvectors of the other matrix by a simple linear mapping.

To see this, let us consider the SVD of the sample matrix $\mathbf{X}$

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{26}$$

Making again use of the orthonormality of $\mathbf{U}$ and $\mathbf{V}$, we have

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}\mathbf{U}^T = \mathbf{U}\mathbf{D}^2\mathbf{U}^T \tag{27}$$

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^2\mathbf{V}^T, \tag{28}$$

i.e., the eigenvectors of $\mathbf{X}\mathbf{X}^T$ are obtained as the left singular vectors $\mathbf{U}$, while the eigenvectors of $\mathbf{X}^T\mathbf{X}$ are obtained as the right singular vectors $\mathbf{V}$. Furthermore, we see that the non-zero eigenvalues for both matrices are given by the squared singular values of $\mathbf{X}$, i.e., $\lambda_i = \sigma_i^2$ (which also shows that their eigenvalues must be ≥ 0).

Another important (and very useful) property of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ is that the sample matrix $\mathbf{X}$ can be used to map the eigenvectors of $\mathbf{X}^T\mathbf{X}$ ($\mathbf{V}$) onto the eigenvectors of $\mathbf{X}\mathbf{X}^T$ ($\mathbf{U}$)

$$\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V} = \mathbf{U}\mathbf{D}. \tag{29}$$

Similarly, $\mathbf{X}^T$ can be used to obtain the eigenvectors of $\mathbf{X}^T\mathbf{X}$ from those of $\mathbf{X}\mathbf{X}^T$. In both cases, the mapped eigenvectors do not have unit length, but will be scaled by the square roots of their corresponding eigenvalues.

Centering

In most cases, we are interested in the covariance of the data rather than in its autocorrelation. Normally, centering (mean normalization) of the data is applied before any higher order moments are computed. However, this is not always possible (in particular, in the context of kernel methods) . In

this subsection, we will discuss how to compensate for the effect of the mean **after** the autocorrelation and the Gram matrices have been computed.

Let $\mathbf{1}_N$ be a $N \times 1$ column vector with all components equal to $\frac{1}{N}$ and $\mathbf{1}_{N\times N} = N\mathbf{1}_N\mathbf{1}_N^T$ a $N \times N$ matrix with all components equal to $\frac{1}{N}$. Then the estimated mean $\hat{\mathbf{m}}$ and covariance matrix $\hat{\mathbf{C}}$ can be written as

$$\hat{\mathbf{m}} = \mathbf{X}\mathbf{1}_N \tag{30}$$

and

$$\begin{aligned}(N-1)\hat{\mathbf{C}} &= (\mathbf{X} - \mathbf{X}\mathbf{1}_{N\times N})(\mathbf{X} - \mathbf{X}\mathbf{1}_{N\times N})^T \\ &= \mathbf{X}\mathbf{X}^T - 2\mathbf{X}\mathbf{1}_{N\times N}\mathbf{X}^T + \mathbf{X}\mathbf{1}_{N\times N}\mathbf{X}^T \\ &= \mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{1}_{N\times N}\mathbf{X}^T \\ &= N\hat{\mathbf{S}} - \mathbf{X}\mathbf{1}_{N\times N}\mathbf{X}^T, \end{aligned} \tag{31}$$

whereby in the derivation of Eq. 31 we made use of the symmetry and idempotency[4] of $\mathbf{1}_{N\times N}$. Eq. 31 expresses the the estimated covariance solely in terms of the sample matrix; in most textbooks on statistics, it is given in a different (though equivalent) form, which makes the dependency on the estimated mean $\hat{\mathbf{m}}$ explicit

$$\hat{\mathbf{C}} = \frac{1}{N-1}(\mathbf{X}\mathbf{X}^T - N\hat{\mathbf{m}}\hat{\mathbf{m}}^T). \tag{32}$$

Both formulae are, however, somewhat susceptible to roundoff-error; for practical calculations, it is better to compute $\hat{\mathbf{C}}$ as the autocorrelation of the mean normalized data, possibly followed by an additional error correction step (e.g., the *corrected two-pass algorithm*); see [4] for details.

Similarly, we obtain an expression for the centered Gram matrix $\mathbf{G}_C$

$$\begin{aligned}\mathbf{G}_C &= (\mathbf{X} - \mathbf{X}\mathbf{1}_{N\times N})^T(\mathbf{X} - \mathbf{X}\mathbf{1}_{N\times N}) \\ &= \mathbf{X}^T\mathbf{X} - \mathbf{1}_{N\times N}\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{X}\mathbf{1}_{N\times N} + \mathbf{1}_{N\times N}\mathbf{X}^T\mathbf{X}\mathbf{1}_{N\times N} \\ &= \mathbf{G} - \mathbf{1}_{N\times N}\mathbf{G} - \mathbf{G}\mathbf{1}_{N\times N} + \mathbf{1}_{N\times N}\mathbf{G}\mathbf{1}_{N\times N}. \end{aligned} \tag{33}$$

Eq. 33 is of some practical importance in the context of kernel methods: when we replace the Gram matrix by a kernel matrix, we perform an implicit mapping ϕ of the input samples into a non-linear, intermediate feature space (the kernel matrix is the Gram matrix of the mapped samples). In general,

[4] A matrix (operator) $\mathbf{A}$ is called *idempotent* if $\mathbf{A}\mathbf{A} = \mathbf{A}$.

we cannot compute the mapping ϕ directly and, as a consequence, are not able to compute the mean of the mapped samples in feature space. However, by virtue of Eq. 33, we can compute the Gram matrix of the centered data as function of the original Gram matrix (without having to know the mean).

0.3 Generalized Eigenproblem

The **generalized eigenproblem** can be stated as follows: given matrices $\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$, find $\mathbf{e} \in \mathbb{R}^n, \lambda \in \mathbb{R}$, so that

$$\mathbf{Ae} = \lambda \mathbf{Be}. \tag{34}$$

If $\mathbf{B}$ is non-singular, the solutions can be obtained by solving the equivalent (standard) eigenvalue problem

$$\mathbf{B}^{-1}\mathbf{Ae} = \lambda \mathbf{e}. \tag{35}$$

In particular, if $\mathbf{B} = \mathbf{I}$, we obtain the ordinary eigenvalue problem Eq. 17 as a special case of Eq. 34. It is easily seen that, as in the case of the ordinary eigenvalue problem, any linear combination of two eigenvectors $\mathbf{e}_{m_i}, \mathbf{e}_{m_j}$ belonging to the same eigenvalue λ_m yields another eigenvector $\mu_i \mathbf{e}_{m_i} + \mu_j \mathbf{e}_{m_j}$ belonging to λ_m (for all $\mu_i, \mu_j \in \mathbb{R}$).

0.3.1 Rayleigh Quotient

Now let as assume that both $\mathbf{A}$ and $\mathbf{B}$ are symmetric and, in addition, $\mathbf{B}$ is also positive definite. The ratio

$$r(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{Aw}}{\mathbf{w}^T\mathbf{Bw}}, \tag{36}$$

which is known as **Rayleigh quotient**, is closely related to the generalized eigenproblem stated above. To see this, let us determine the extremum (stationary) points of $r(\mathbf{w})$, i.e., the points $\mathbf{w}^*$ satisfying $\nabla r(\mathbf{w}^*) = \mathbf{0}$. The gradient $\nabla r(\mathbf{w})$ is calculated as

$$\nabla r(\mathbf{w}) = \frac{2\mathbf{Aw} - 2\mathbf{w}^T\mathbf{AwBw}}{(\mathbf{w}^T\mathbf{Bw})^2} = \frac{2\mathbf{Aw} - 2r(\mathbf{w})\mathbf{Bw}}{\mathbf{w}^T\mathbf{Bw}}. \tag{37}$$

Setting $\nabla r(\mathbf{w})$ to $\mathbf{0}$, we obtain

$$\mathbf{Aw} = r(\mathbf{w})\mathbf{Bw}, \tag{38}$$

which is recognized as Eq. 34. Thus, the extremum points $\mathbf{w}^*$ (extremum values $r(\mathbf{w}^*)$) of the Rayleigh Quotient $r(\mathbf{w})$ are obtained as the eigenvectors $\mathbf{e}$ (eigenvalues $\lambda(\mathbf{e})$) of the corresponding generalized eigenproblem.

An important consequence of the symmetry of $\mathbf{A}$ and $\mathbf{B}$ is that generalized eigenvectors $\mathbf{e}_i$, $\mathbf{e}_j$ belonging to different eigenvalues λ_i and λ_j, respectively, are orthogonal w.r.t. the inner products induced by $\mathbf{A}$ and $\mathbf{B}$, i.e.,

$$\mathbf{e}_i\mathbf{A}\mathbf{e}_j = \mathbf{e}_i\mathbf{B}\mathbf{e}_j = \mathbf{0}, \tag{39}$$

if $\mathbf{e}_j$ and $\mathbf{e}_i$ belong to different eigenvalues; the condition Eq. 39 also implies that $\mathbf{e}_j$ and $\mathbf{e}_i$ are linearly independent (for a proof, see [1]).

0.3.2 Simultaneous Diagonalization

Given two symmetric matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, *simultaneous diagonalization* (see, for example, [2]) tries to find a nonsingular transformation matrix $\mathbf{T}$, such that

$$\mathbf{T}^T\mathbf{A}\mathbf{T} = \mathbf{\Phi} \tag{40}$$
$$\mathbf{T}^T\mathbf{B}\mathbf{T} = \mathbf{I}, \tag{41}$$

whereby $\mathbf{\Phi}$ is a diagonal matrix and $\mathbf{I}$ is the unit matrix.

Simultaneous diagonalization starts by finding an intermediate transformation $\mathbf{T}'$ that transforms $\mathbf{B}$ into the unit matrix. This step is also referred to as *whitening*; if the EVD of $\mathbf{B}$ is given by $\mathbf{F}^T\mathbf{B}\mathbf{F} = \mathbf{\Lambda}_B$ ($\mathbf{\Lambda}_B$ being the diagonal eigenvalue matrix), then the whitening transform $\mathbf{T}'$ is obtained as $\mathbf{F}\mathbf{\Lambda}_B^{-\frac{1}{2}}$. During the second step, the simultaneous diagonalization algorithm determines an orthonormal transformation $\mathbf{T}''$, that diagonalizes $(\mathbf{T}')^T\mathbf{A}\mathbf{T}'$, and, due to its orthonormality, has no effect on the unit matrix. The final transformation is then obtained as $\mathbf{T} = \mathbf{T}'\mathbf{T}''$.

As can easily be verified, this implies that

$$\mathbf{A}\mathbf{T} = \mathbf{B}\mathbf{T}\mathbf{\Phi} \tag{42}$$
$$\mathbf{B}^{-1}\mathbf{A}\mathbf{T} = \mathbf{T}\mathbf{\Phi}, \tag{43}$$

i.e., $\mathbf{\Phi}$ and $\mathbf{T}$ are the eigenvalues and eigenvectors, respectively, of the generalized eigenproblem Eq. 34, and, consequently, the extremum values (points) of the corresponding Rayleigh quotient 36.

In practice, the simultaneous diagonalization of two symmetric matrices can be computed by two consecutive applications of EVD or SVD, thus making it an attractive tool for solving the more complex symmetric generalized eigenproblem (or obtaining the extrema of a Rayleigh quotient). It is also applicable when $\mathbf{B}$ is singular: assuming $rank(\mathbf{B}) = k < n$, only the first k eigenvectors (belonging to non-zero eigenvalues) of $\mathbf{B}$ are used in the whitening of $\mathbf{B}$

$$(\mathbf{T}'_{[k]})^T \mathbf{B} \mathbf{T}'_{[k]} = \mathbf{I}_{[k]}, \tag{44}$$

whereby $\mathbf{T}' \in \mathbb{R}^{n \times k}$, and the second transformation $\mathbf{T}''$ is determined by diagonalizing $(\mathbf{T}'_{[k]})^T \mathbf{A} \mathbf{T}'_{[k]} \in \mathbb{R}^{k \times k}$. The final transformation is then given by $\mathbf{T} = \mathbf{T}'\mathbf{T}'' \in \mathbb{R}^{n \times k}$.

This approach, which could be referred to as *reduced rank simultaneous diagonalization*, combines diagonalization with dimensionality reduction and effectivly constrains the diagonalization to the range of $\mathbf{B}$. A good discussion (in the context of LDA) can be found in [6].

Bibliography

[1] Magnus Borga. *Learning Multidimensional Signal Processing*. Linköping Studies in Science and Technology, Dissertations, No. 531. Department of Electrical Engineering, Linköping University, Linköping, Sweden, 1998.

[2] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, second edition, 1990.

[3] Gene H. Golub and Charles F. van Loan. *Matrix Computations*. Johns Hopkins University Press, second edition, 1989.

[4] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 1992.

[5] Alberto Ruiz and Pedro Lopez-de Teruel. Nonlinear kernel-based statistical pattern analysis. *IEEE Trans. Neural Networks*, 12(1):16–32, 2001.

[6] Hua Yu and Jie Yang. A direct LDA algorithm for high-dimensional data – with application to face recognition. *Pattern Recognition*, 34(10):2067–2070, 2001.