Supplement 3: Compression

Mladen Victor Wickerhauser

PMF 2022: Dimensionality Reduction and Manifold Estimation

July 22, 2022

1 Optimality of K-L.

Let $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be a collection of points in $\mathbf{R}^d$, treated as samples from a d-variate normal distribution.

The mean of the distribution is estimated by the average

$$\mathrm{E}(\mathbf{x}) \approx \bar{\mathbf{x}} \stackrel{\text{def}}{=} \sum_{n=1}^{N} \mathbf{x}_n,$$

and the covariance is likewise estimated by the *sample covariance matrix*

$$\mathrm{cov}\,(\mathbf{x}) = \mathrm{E}([\mathbf{x}-\bar{\mathbf{x}}][\mathbf{x}-\bar{\mathbf{x}}]^T) \approx M \in \mathbf{R}^{d\times d},$$

where

$$M(i,j) = \frac{1}{N-1}\sum_{n=1}^{N}[\mathbf{x}_n(i) - \bar{x}(i)][\mathbf{x}_n(j) - \bar{x}(j)].$$

This M is a positive semidefinite symmetric matrix, in fact positive definite if there are at least d distinct values in $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, and it may therefore be diagonalized by an orthogonal similarity transformation $M \mapsto U^T M U$, as will be discussed below. That diagonalizing transformation is called the *empirical Karhunen-Loève transform*, and it is an optimum for various functions on orthogonal matrices as will be seen.

1.1 Matrices and Eigenvalues

Let M be a $d \times d$ matrix. Write $M(i,j)$ for the element in row i and column j, for $i, j \in \{1, \dots, d\}$. This is the notation used by Octave, a software system for linear algebra computations.

An *eigenvalue* λ of M is a number for which there exists a nonzero *eigenvector*, say $\mathbf{v} \in \mathbf{R}^d$, such that $M\mathbf{v} = \lambda v$, or equivalently,

$$(\lambda I - M)\mathbf{v} = \mathbf{0},$$

where I is the $d \times d$ identity matrix and $\mathbf{0}$ is the zero vector:

$$I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}; \qquad \mathbf{0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Matrix I has $I(i,i) = 1$, $i = 1, \ldots, d$ and $I(i,j) = 0$ at all i, j with $i \neq j$. The ones are said to lie on the *main diagonal*.

Eigenvalues are roots of the *characteristic polynomial* of M:

$$q(\lambda) \stackrel{\text{def}}{=} \det(\lambda I - M).$$

This q is a polynomial of degree d, and its coefficients are themselves homogenous polynomials of degree d in the coefficients $\{M(i,j) : 1 \leq i, j \leq d\}$, of which there are d^2.

Note that λ is an eigenvalue iff $q(\lambda) = \det(\lambda I - M) = 0$, iff $\lambda I - M$ is singular, iff there exists a nonzero solution $\mathbf{v}$ to $(\lambda I - M)\mathbf{v} = \mathbf{0}$.

By the Fundamental Theorem of Algebra, polynomial q may be factored into linear terms in its d complex-number roots $\lambda 1, \ldots, \lambda_d$:

$$q(\lambda) = \prod_{k=1}^{d} (\lambda - \lambda_k).$$

The number of appearances in this product of a particular root, namely a particular eigenvalue, is called its *multiplicity*.

1.2 Positive Definite Symmetric Matrices

Suppose that the real-valued $d \times d$ square matrix M is *symmetric*, namely that $M^T = M$. By the Spectral Theorem, there is an *orthogonal* $d \times d$ matrix V, namely one satisfying $VV^T = V^TV = I$, whose columns form an orthonormal basis for $\mathbf{R}^d$, such that

$$M = VDV^T, \qquad D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix},$$

where D is a diagonal matrix containing the eigenvalues of M, with repetitions according to their multiplicity.

Say that such a matrix M is *positive definite*, and write $M > 0$, iff $(\forall i)\lambda_i > 0$.

Theorem 1. *Symmetric matrix M is positive definite if and only if $\mathbf{x}^T M\mathbf{x} > 0$ for every nonzero vector $\mathbf{x} \in \mathbf{R}^d$.*

Proof. ($\implies$) Suppose that $M = VDV^T$ is positive definite, and $\mathbf{x}$ is nonzero. Let $\mathbf{y} = V^T\mathbf{x}$ so that $\mathbf{x} = V\mathbf{y}$ and $\mathbf{x}^T = (V\mathbf{y})^T = \mathbf{y}^T V^T$. Then $\mathbf{y} \neq \mathbf{0}$, so

$$\mathbf{x}^T M\mathbf{x} = \mathbf{y}^T V^T VDV^T V\mathbf{y} = \mathbf{y}^T D\mathbf{y} = \sum_{i=1}^{d} \lambda_i y_i^2 > 0, \quad \text{for } \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} \in \mathbf{R}^d,$$

since $(\forall i)\lambda_i > 0$ and $(\exists i) y_i \neq 0$ so that $y_i^2 > 0$.

($\Longleftarrow$) Let $\mathbf{v}$ be an eigenvector of M with eigenvalue λ. Then

$$\lambda \|\mathbf{v}\|^2 = \mathbf{v}^T(\lambda \mathbf{v}) = \mathbf{v}^T M \mathbf{v} > 0.$$

Since $\|\mathbf{v}\|^2 > 0$, conclude that $\lambda > 0$ □

Now consider *principal submatrices*, which are obtained from M by deleting rows and columns simultaneously.

Theorem 2. *Any principal submatrix of a positive definite symmetric matrix is positive definite symmetric.*

Proof. The principal submatrix inherits symmetry since rows and columns are removed simultaneously.

Now let M be a $d \times d$ matrix and let N be a principal $k \times k$ submatrix of M. Suppose that nonzero $\mathbf{y} \in \mathbf{R}^d$ satisfies

$$\mathbf{y}^T N \mathbf{y} \leq 0.$$

Then the vector $\mathbf{x} \in \mathbf{R}^d$ obtained by injecting $\mathbf{y}$ into $\mathbf{R}^d$ at the retained row and column coordinates, with zeros at the deleted coordinates, will also be nonzero and will satisfy

$$\mathbf{x}^T M \mathbf{x} \leq 0.$$

Conclude that if any principal submatrix is not positive definite, then M is not positive definite. □

Some immediate consequences for a positive definite M are:

- All diagonal elements are positive: $(\forall i) M(i,i) > 0$.
- The eigenvalues of any principal submatrix are all positve.
- Any principal submatrix will be invertible.

Remark. In Octave notation, the principal submatrix with kept rows and columns $(i_1, \ldots, i_k)$ is obtained from matrix M as follows:

```
kept=[i1,...,ik]; N=M(kept,kept);
```

The kept rows and columns are, of course, the complement in $(1, \ldots, d)$ of the deleted rows and columns.

1.3 Parametrization by Orthogonals

Let M be a $d \times d$ positive definite symmetric matrix, so that $M = UDU^T$ with diagonal matrix D of its positive eigenvalues $D(k,k) = \lambda_k$ and its diagonalizing orthogonal matrix U.

The main diagonal elements of M are convex combinations of the eigenvalues:

$$M(i,i) = \sum_{j,k} U(i,j)D(j,k)U^T(k,i) = \sum_{k=1}^{d} U(i,k)^2 \lambda_k$$

Since U is orthogonal, its rows have unit norm, so for each i the sequence $U(i,1)^2, \ldots, U(i,d)^2$ sums to 1. More is actually true: the columns of U also have unit norm, so for each j the sequence $U(1,j)^2, \ldots, U(d,j)^2$ also sums to 1. The matrix of squared elements of U is therefore *doubly stochastic.*

Recall that a function $f : \mathbf{R} \to \mathbf{R}$ is *concave* iff, for any $x, y \in \mathbf{R}$ and any $t \in [0,1]$,

$$f(tx + [1-t]y) \geq tf(x) + [1-t]f(y). \tag{1}$$

Remark. This definition applies more generally to a function with a convex domain $K \subset \mathbf{R}^d$, namely a set K for which $\mathbf{x}, \mathbf{y} \in K \implies t\mathbf{x} + [1-t]\mathbf{y} \in K$ for all $0 \leq t \leq 1$.

Theorem 3. *Suppose that $f : \mathbf{R} \to \mathbf{R}$ is a concave function, $x_1, \ldots, x_d$ are real numbers, and A is a doubly stochastic $d \times d$ matrix. Then*

$$\sum_{k=1}^{d} f(y_k) \geq \sum_{k=1}^{d} f(x_k), \qquad \textit{where} \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} = A \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

Proof. For each k, write

$$y_k = \sum_{j=1}^{d} A(k,j)x_j.$$

By Eq.1, since f is concave,

$$f(y_k) \geq \sum_{j=1}^{d} A(k,j)f(x_j).$$

Now sum over $k = 1, \ldots, d$ to get

$$\begin{aligned} \sum_{k=1}^{d} f(y_k) &\geq \sum_{k=1}^{d}\sum_{j=1}^{d} A(k,j)f(x_j) \\ &= \sum_{j=1}^{d} \left(\sum_{k=1}^{d} A(k,j) \right) f(x_j) = \sum_{j=1}^{d} f(x_j), \end{aligned}$$

since A is doubly stochastic, so $\sum_{k=1}^{d} A(k,j) = 1$ for all j. □

1.4 Maximizing Coding Gain

Suppose that M is a positive definite symmetric $d \times d$ matrix. Let U be any $d \times d$ orthogonal matrix and define

$$M_U \stackrel{\text{def}}{=} U^T M U.$$

Such a *similarity transformation* preserves symmetry:

$$M_U^T = (U^T M U)^T = U^T M^T (U^T)^T = U^T M U = M_U.$$

It also preserves eigenvalues. Suppose that λ is an eigenvalue of M. Let $\mathbf{x}$ be a nonzero vector with $M\mathbf{x} = \lambda\mathbf{x}$, let $\mathbf{y} = U^T\mathbf{x} \neq \mathbf{0}$, and compute

$$M_U\mathbf{y} = U^T M U U^T \mathbf{x} = U^T M \mathbf{x} = \lambda U^T \mathbf{x} = \lambda \mathbf{y}.$$

Thus λ is an eigenvalue of M_U. A similar argument shows that every eigenvalue of M_U is also an eigenvalue of M. Hence M_U has all positive eigenvalues just like M and is likewise positive definite.

Transform coding gain from U measures the concentration of variance onto the main diagonal elements of M_U:

$$G(U) \stackrel{\text{def}}{=} \sum_{k=1}^{d} \log \frac{1}{M_U(k,k)} = \log \prod_{k=1}^{d} \frac{1}{M_U(k,k)},$$

This requires $M_U(k,k) > 0$, all k, which is assured by Theorem 2 and its consequences.

Theorem 4. *For any orthogonal U,*

$$G(U) \leq -\log\det M.$$

Equality holds if and only if U diagonalizes M.

Proof. Observe that log is a concave function, and that

$$-G(U) = \sum_{k=1}^{d} \log M_U(k,k)$$

But the diagonal elements of $M_U(k,k)$ are the output of a doubly stochastic matrix applied to the vector of eigenvalues $\lambda_1, \ldots, \lambda_d$ of M. By Theorem 3,

$$-G(U) \geq \sum_{k=1}^{d} \log \lambda_k = \log\det M,$$

from which the inequality follows.

Equality holds if and only if U diagonalizes M, in which case M_U is a diagonal matrix with some permutation of the eigenvalues on its main diagonal. $\square$

1.5 Minimizing Entropy

Recall that the *trace* of a matrix M, denoted $\operatorname{tr} M$, is the sum of it main diagonal elements, and also the sum of its eigenvalues:

$$\operatorname{tr} M \stackrel{\text{def}}{=} \sum_{k=1}^{d} M(k,k) = \sum_{k=1}^{d} \lambda_k.$$

Thus trace is invariant under similarity transformations:

$$(\forall U)\operatorname{tr} M_U = \operatorname{tr} M.$$

Dividing M_U by $\operatorname{tr} M$ normalizes the main diagonal to be nonnegative with sum 1. It may then be considered a discrete pdf, and its concentration measured by *entropy*:

$$\sum_{k=1}^{d} p_U(k) \log \frac{1}{p_U(k)},$$

where $p_U(k) \stackrel{\text{def}}{=} M_U(k,k)/\operatorname{tr} M_U = M_U(k,k)/\operatorname{tr} M$.

However, the normalization is unnecessary since the function $x \to x\log(1/x = -x\log x$ is concave on all of $\mathbf{R}^+$, as may be easily checked by differentiation. It may be also extended to $x = 0$ by continuity as $0\log 0 = 0\log(1/0) = 0$. Instead, consider the unnormalized entropy

$$H(U) \stackrel{\text{def}}{=} \sum_{k=1}^{d} M_U(k,k) \log \frac{1}{M_U(k,k)}.$$

This function has a feature in common with H, proved by a similar application of Theorem 3:

Theorem 5. *The minimum value of $H(U)$, which is*

$$\sum_{k=1}^{d} \lambda_k \log \frac{1}{\lambda_k}$$

is attained at any orthogonal matrix U that diagonalizes M. □

1.6 Information Cost Functions

Any concave function $f : \mathbf{R} \to \mathbf{R}$ defines an *information cost function*:

$$I(U) \stackrel{\text{def}}{=} \sum_{k=1}^{d} f(M_U(k,k)).$$

Again, Theorem 3 implies that I behaves like transform coding gain:

Theorem 6. *The minimum value of $I(U)$ is attained at any orthogonal matrix U that diagonalizes M.* □

Since the Karhonen-Loève transform is the diagonalizing orthogonal matrix for the empirical covariance matrix, these results may be summarized as follows:

Theorem 7. *Suppose that M is the empirical covariance matrix for a set of samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbf{R}^d$. Let $I(U)$ be any information cost function on the main diagonal of M_U. Then the Karhonen-Loève transform U, which makes M_U diagonal, attains the minimum value for $I(U)$.* □