Principal Component Analysis and Whitening

Whitening means that we linearly transform a zero-mean random vector \( \mathbf{x} \) with \( n \) elements by multiplying it with some \( n \times n \)-matrix \( \mathbf{V} \) so that the \( n \)-vector \( \mathbf{z} = \mathbf{V} \mathbf{x} \) has a covariance matrix \( \mathbf{C}_z = \mathbf{I} \).

Let \( \mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n) \) be the matrix whose columns are the unit-norm eigenvectors of \( \mathbf{C}_x \), and let \( \mathbf{D} = \text{diag}(d_1, \ldots, d_n) \) be the diagonal matrix of (positive) eigenvalues of \( \mathbf{C}_x \). An important instance of \( \mathbf{V} \) is the matrix \( \mathbf{E} \mathbf{D}^{-1/2} \), which is called the inverse square root \( \mathbf{C}_x^{-1/2} \) of the covariance matrix \( \mathbf{C}_x \).

1. The whitening matrix \( \mathbf{V} \) is not unique. Show that \( \mathbf{U} \mathbf{V} \) is also a whitening matrix if \( \mathbf{U} \) is an orthogonal matrix \( (\mathbf{U}^T \mathbf{U} = \mathbf{U} \mathbf{U}^T = \mathbf{I}) \) and \( \mathbf{V} \) is an arbitrary whitening matrix.

2. The vector \( \hat{\mathbf{x}} = \sum_{i=1}^{m} (\mathbf{e}_i^T \mathbf{x}) \mathbf{e}_i \) is called the projection of \( \mathbf{x} \) onto the subspace spanned by the first \( m \) eigenvectors \( (m \leq n) \). Show that the mean-square (reconstruction) error is

\[
E\{|\mathbf{x} - \hat{\mathbf{x}}|^2\} = \sum_{i=m+1}^{n} d_i .
\]

Hint: \( \text{tr}\{\mathbf{C}_x\} = E\{\mathbf{x}^T \mathbf{x}\} = \sum_{i=1}^{n} d_i \).

3. An important practical problem in PCA is how to choose \( m \), i.e. to find a limit below which the eigenvalues are so small as to be insignificant. Sometimes an optimal \( m \) can be found from prior information about the vector \( \mathbf{x} \). For instance, assume that \( \mathbf{x} \) is of length \( n \) and obeys a signal-noise model

\[
\mathbf{x} = \sum_{i=1}^{m} \mathbf{a}_i \mathbf{s}_i + \mathbf{n}
\]

where \( m < n \). The \( \mathbf{a}_i \) are some fixed, pairwise orthogonal vectors that span an \( m \)-dimensional signal subspace, and the random vector \( \mathbf{s} = (s_1, \ldots, s_m)^T \) is white. The term \( \mathbf{n} \) is white noise, for which \( E\{\mathbf{n}^T \mathbf{n}\} = \sigma^2 \mathbf{I} \) holds. Show that

a) the covariance matrix of \( \mathbf{x} \) is \( \mathbf{C}_x = \sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^T + \sigma^2 \mathbf{I} \).

b) the eigenvalues are constant beyond index \( m \): \( d_1 \geq d_2 \geq \ldots \geq d_m > d_{m+1} = \ldots = d_n = \sigma^2 \).

Higher-Order Moments

4. Show that for two zero-mean, independent random variables \( s_1 \) and \( s_2 \),

a) \( \text{kurt}(s_1 + s_2) = \text{kurt}(s_1) + \text{kurt}(s_2) \) and

b) \( \text{kurt}(\alpha s_1) = \alpha^4 \text{kurt}(s_1) \), where \( \alpha \) is constant.
5. Nongaussianity can be measured by the kurtosis. To use kurtosis for ICA estimation of sources $s$, we need to prove that the maxima of the function

$$F_q(q) = \text{kurt}(q^T s) = |q_1^4 \text{kurt}(s_1) + q_2^4 \text{kurt}(s_2)|$$

for $|q| = 1$ are obtained when only one of the components of $q = (q_1, q_2)^T$ is nonzero.

a) Make the change of variables $t_i = q_i^2$. What is the geometrical form of the constraint set of $t = (t_1, t_2)^T$?

b) Assume that $\text{kurt}(s_1) > \text{kurt}(s_2) > 0$. What is the geometrical shape of sets $F_t(t) = \text{const.}$? By a geometrical argument, show that the maximum of $F_t(t)$ is obtained when one of the $t_i$ is one and the other one is zero.

c) Assume that $\text{kurt}(s_1) < \text{kurt}(s_2) < 0$. Use the same logic as in b).

d) Assume that the kurtoses have different signs. What is the shape of sets $F_t(t) = \text{const.}$ now?

Information Theory

The differential entropy $H$ of a continuous-valued random vector $x$ with density $p_x$ is defined as

$$H(x) = - \int p_x(\xi) \log p_x(\xi) \, d\xi .$$

The negentropy is defined as

$$J(x) = H(x_{\text{gauss}}) - H(x)$$

where $x_{\text{gauss}}$ is a gaussian random vector of the same covariance matrix $\Sigma$ as $x$.

6. Show that the entropy is not scale-invariant, i.e. it changes as we linearly transform to a different coordinate system $y = A x$, where $A$ is invertible. Hint: Use the density of a transformation, $p_y(\eta) = |\det A|^{-1} p_x(A^{-1} \eta)$, and prove $H(y) = H(x) + \log |\det A|$.

7. Prove $H(x_{\text{gauss}}) = \frac{1}{2} \log |\det \Sigma| + \frac{n}{2} [1 + \log 2\pi]$. Hint: use the definition of a gaussian pdf from Exercises 2.

8. Show that the negentropy is scale-invariant: $J(A x) = J(x)$.

Mutual information (“Transinformation”) is a measure of the information that members of a set of random variables have about the other random variables in the set. The mutual information $I$ between $n$ scalar random variables, $x_i, i = 1, \ldots, n$, is

$$I(x_1, x_2, \ldots, x_n) = \sum_{i=1}^n H(x_i) - H(x)$$

9. Show that the following relation holds:

$$I(x_1, x_2, \ldots, x_n) = \int p_x(\xi) \log \frac{p_x(\xi)}{p_1(\xi_1) p_2(\xi_2) \cdots p_n(\xi_n)} \, d\xi .$$

Dr. R. Kempter, phone 2093-8925, room 2315, r.kempter(AT)biologie.hu-berlin.de
Prof. Dr. A.V.M. Herz, phone 2093-9112, room 2325, a.herz(AT)biologie.hu-berlin.de

Problems handed out on Monday, 08.05.2006.
Solutions to be handed in by Monday, 15.05.2006, 1215 pm.
Discussion and presentation of the problems on Friday, 19.05.2006, 830 – 1000 am in front of room 2317 I-W (Zwischengeschoß).