UNIT 3

PROBABILISTIC PCA

The Probability Model

The use of the isotropic Gaussian noise model $N(\mathbf{0}, \sigma^2\mathbf{I})$ for $\boldsymbol{\epsilon}$ in conjunction with equation (1) implies that the $\mathbf{x}$-conditional probability distribution over $\mathbf{t}$-space is given by

$$\mathbf{t}|\mathbf{x} \sim N(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}). \tag{2}$$

With the marginal distribution over the latent variables also Gaussian and conventionally defined by $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$, the marginal distribution for the observed data $\mathbf{t}$ is readily obtained by integrating out the latent variables and is likewise Gaussian:

$$\mathbf{t} \sim N(\boldsymbol{\mu}, \mathbf{C}), \tag{3}$$

where the observation covariance model is specified by $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$. The corresponding log-likelihood is then

$$L = -\frac{N}{2}\left\{d\ln(2\pi) + \ln|\mathbf{C}| + \mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{S}\right)\right\}, \tag{4}$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}. \tag{5}$$

The maximum-likelihood estimator for $\boldsymbol{\mu}$ is given by the mean of the data, in which case $\mathbf{S}$ is the sample covariance matrix of the observations $\{\mathbf{t}_n\}$. Estimates for $\mathbf{W}$ and σ^2 may be obtained by iterative maximisation of L, for example using the EM algorithm given in Appendix B, which is based on the algorithm for standard factor analysis of Rubin and Thayer (1982). However, in contrast to factor analysis, M.L.E.s for $\mathbf{W}$ and σ^2 may be obtained explicitly, as we see shortly.

Later, we will make use of the conditional distribution of the latent variables $\mathbf{x}$ given the observed $\mathbf{t}$, which may be calculated using Bayes' rule and is again Gaussian:

$$\mathbf{x}|\mathbf{t} \sim N(\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}), \tag{6}$$

where we have defined $\mathbf{M} = \mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I}$. Note that $\mathbf{M}$ is of size $q \times q$ while $\mathbf{C}$ is $d \times d$.

Properties of the Maximum-Likelihood Estimators

In Appendix A it is shown that, with $\mathbf{C}$ given by $\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$, the likelihood (4) is maximised when:

$$\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{7}$$

where the q column vectors in the $d \times q$ matrix $\mathbf{U}_q$ are the principal eigenvectors of $\mathbf{S}$, with corresponding eigenvalues $\lambda_1, \ldots, \lambda_q$ in the $q \times q$ diagonal matrix $\mathbf{\Lambda}_q$, and $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal rotation matrix. Other combinations of eigenvectors (i.e. non-principal ones) correspond to saddle-points of the likelihood function. Thus, from (7), the latent variable model defined by equation (1) effects a mapping from the latent space into the *principal subspace* of the observed data.

It may also be shown that for $\mathbf{W} = \mathbf{W}_{\mathrm{ML}}$, the maximum-likelihood estimator for σ^2 is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j, \tag{8}$$

which has a clear interpretation as the variance 'lost' in the projection, averaged over the lost dimensions.

In practice, to find the most likely model given $\mathbf{S}$, we would first estimate σ^2_{ML} from (8), and then $\mathbf{W}_{\mathrm{ML}}$ from (7), where for simplicity we would effectively ignore $\mathbf{R}$ (i.e. choose $\mathbf{R} = \mathbf{I}$). Alternatively, we might employ the EM algorithm detailed in Appendix B, where $\mathbf{R}$ at convergence can be considered arbitrary.

Factor Analysis Revisited

Although the above estimators result from application of a simple constraint to the standard factor analysis model, we note that an important distinction resulting from the use of the isotropic noise covariance $\sigma^2\mathbf{I}$ is that PPCA is covariant under rotation of the original data axes, as is standard PCA, while factor analysis is covariant under component-wise rescaling. Another point of contrast is that in factor analysis, neither of the factors found by a two-factor model is necessarily the same as that found by a single-factor model. In probabilistic PCA, we see above that the principal axes may be found incrementally.

Dimensionality Reduction

The general motivation for PCA is to transform the data into some reduced-dimensionality representation, and with some minor algebraic manipulation of $\mathbf{W}_{\mathrm{ML}}$, we may indeed obtain the standard projection onto the principal axes if desired. However, it is more natural from a probabilistic perspective to consider the dimensionality-reduction process in terms of the distribution of the latent variables, conditioned on the observation. From (6), this distribution may be conveniently summarised by its *mean*:

$$\langle \mathbf{x}_n | \mathbf{t}_n \rangle = \mathbf{M}^{-1}\mathbf{W}_{\mathrm{ML}}{}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}). \tag{9}$$

(Note, also from (6), that the corresponding conditional covariance is given by $\sigma^2_{\mathrm{ML}}\mathbf{M}^{-1}$ and is thus independent of n.) It can be seen that when $\sigma^2 \to 0$, $\mathbf{M}^{-1} \to (\mathbf{W}_{\mathrm{ML}}{}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}})^{-1}$ and (9) then represents an orthogonal projection into latent space and so standard PCA is recovered. However, the density model then becomes singular, and thus undefined. In practice, with $\sigma^2 > 0$ as determined by (8), the latent projection becomes skewed towards the origin as a result of the Gaussian marginal distribution for $\mathbf{x}$. Because of this, the reconstruction $\mathbf{W}_{\mathrm{ML}}\langle \mathbf{x}_n | \mathbf{t}_n \rangle + \boldsymbol{\mu}$ is *not* an orthogonal projection of $\mathbf{t}_n$, and is therefore not optimal (in the squared reconstruction-error sense). Nevertheless, optimal reconstruction of the observed data from the conditional latent mean may still be obtained, in the case of $\sigma^2 > 0$, and is given by $\mathbf{W}_{\mathrm{ML}}(\mathbf{W}_{\mathrm{ML}}{}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}})^{-1}\mathbf{M}\langle \mathbf{x}_n | \mathbf{t}_n \rangle + \boldsymbol{\mu}$.

Reference:

Michael E. Tipping, Christopher M. Bishop 1999, Probabilistic PCA, *Journal of the Royal Statistical Society, Series B,* **61**, *Part* 3, *pp.* 611–622.