1 Introduction

Let us define a dataset

$$\mathcal{D} = \{\mathbf{x}_n\}, \quad n \in \{1, \dots, N\}$$

where $\mathbf{x}_i$ are m-vectors.

The q-principal axes $\mathbf{w}_j$ are orthonormal onto which the retained variance under the projection is maximized.

Sample covariance matrix

$$\mathbf{S} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T$$

We define η_i to be the projected vectors.

$$\eta_i = \mathbf{W}^T(\mathbf{x}_i - \bar{\mathbf{x}})$$

where $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_q\}$

The variable η are uncorrelated such that the covariance

$$\mathbf{S}_\eta = \frac{1}{N-1}\sum_{i=1}^{N}\eta_i\eta_i^T$$

is diagonal with elements λ_j, the eigenvectors of $\mathbf{S}$.

We set the optimal linear reconstruction

$$\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{x}_i + \bar{\mathbf{x}}$$

which will maximize the L2-error (or squared reconstruction error).

2 Latent Variable Model

We use the factor analysis model

$$\mathbf{x} = \mathbf{W}\eta + \mu + \epsilon$$

where we assume

$$\eta \sim \mathcal{N}(0, \mathbf{1}) \tag{2.1}$$
$$\epsilon \sim \mathcal{N}(0, \Sigma) \tag{2.2}$$
$$\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{W}\mathbf{W}^T + \Sigma) \tag{2.3}$$

Where we constrain Σ to be diagonal. The observed variables x_i are conditionally independent on the latent variables $\theta = (\eta, \epsilon, \mu, \Sigma)$.

3 Probabilistic PCA

Using an isotropic Gaussian noise model $\Sigma = \sigma^2\mathbf{1}$, we end up with the conditional distribution

$$\mathbf{x} \mid \eta \sim \mathcal{N}(\mathbf{W}\mathbf{x} + \mu, \sigma^2\mathbf{1})$$

The corresponding log-likelihood is

$$\mathcal{L} = -\frac{N}{2}(d\log(2\pi) + \log\det(C) + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S}))$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{1}$.

By Bayes rule, we can get the posterior

$$\eta \mid \mathbf{x} \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x} - \mu), \sigma^2\mathbf{M}^{-1})$$

where $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{1}$. Using results from matrix differention, we get

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}} = N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{W})$$

There are 3 possible solution (2 non-trivial). The first one is

$$\mathbf{C} = \mathbf{S} \implies \mathbf{W}\mathbf{W}^T = \mathbf{S} - \sigma^2\mathbf{1}$$

Therefore

$$\hat{\mathbf{W}}_{\text{ML}} = \mathbf{U}(\Lambda - \sigma^2\mathbf{1})^{1/2}\mathbf{R}$$

Where $\mathbf{U}$ is the left projector operator of the singular value decomposition of $\mathbf{S}$ and Λ is eigenvalue diagonal matrix. $\mathbf{R}$ is an arbitrary rotation matrix.

The operator may be ranked reduce to include only non-zero eigenvalues.

From this maximum likelihood estimator, we get

$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{m-q}\sum_{j=q+1}^{m}\lambda_j$$

which is the lost variance by the projection averaged over the lost dimension.

sectionAn Expectation Maximization Algorithm for PPCA

In this approach, the latent variables θ are the *missing data*. The complete data likelihood is

$$\mathcal{L}_C = \sum_{i=1}^{N}\log(p(\mathbf{x}_i \mid \eta_i))$$

From our previous definitions, this is

$$p(\mathbf{x}_i \mid \eta_i) = (2\pi\sigma^2)^{-m/2}\exp\left\{-\frac{||\mathbf{x}_i - \mathbf{W}\eta_i - \mu||^2}{2\sigma^2}\right\}(2\pi)^{-m/2}\exp\left\{-\frac{||\eta_i||^2}{2}\right\}$$

In the expectation step (E step), we take the expectation of $\mathcal{L}_C$ with respect to the distribution $p(\eta_i \mid \mathbf{x}_i, \mathbf{W}, \sigma^2)$

$$\langle\mathcal{L}_C\rangle = -\sum_{i=1}^{N}\frac{m}{2}\log(\sigma^2) + \frac{1}{2}\text{Tr}(\langle\eta_i\eta_i^T\rangle) + \frac{1}{2\sigma^2}(\mathbf{x}_i-\mu)^T(\mathbf{x}_i-\mu) - \frac{1}{\sigma^2}\langle\eta_i\rangle^T\mathbf{W}^T(\mathbf{x}_i-\mu) + \frac{1}{2\sigma^2}\text{Tr}(\mathbf{W}^T\mathbf{W}\langle\eta_i\eta_i^T\rangle)$$

where

$$\langle\eta_i\rangle = \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}_i - \mu) \tag{3.1}$$

$$\langle\eta_i\eta_i^T\rangle = \sigma^2\mathbf{M}^{-1} + \langle\eta_i\rangle\langle\eta_i\rangle^T \tag{3.2}$$

$$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{1} \tag{3.3}$$

These values are computed using the fixed statistics $\mathbf{W}$ and σ^2.

In the maximization step (M-step), $\langle\mathcal{L}_C\rangle$ is maximized with respect to $\mathbf{W}$ and σ^2 given

$$\tilde{\mathbf{W}}\left(\sum_{i=1}^{N}(\mathbf{x_i}-\mu)\langle\eta_i\rangle^T\right)\left(\sum_{i=1}^{N}\langle\eta_i\eta_i^T\rangle\right)^{-1}$$

and

$$\tilde{\sigma}^2 = \frac{1}{Nm}\sum_{i=1}^{N}||\mathbf{x}_i-\mu||^2 - 2\langle\eta_i\rangle^T\tilde{\mathbf{W}}^T(\mathbf{x}_i-\mu) + \text{Tr}(\langle\eta_i\eta_i^T\rangle)\tilde{\mathbf{W}}^T\tilde{\mathbf{W}}$$