Probabilistic PCA (Tipping & Bishop, 1999)

data $= \{y_1, \dots, y_m\}$ $\quad y_i \in \mathbb{R}^p$

Let $w_1, \dots, w_K$ be such that $\underset{p\times K}{W} = [w_1, \dots, w_K]$ is the matrix with the K eigenvectors of S:

$$S = \frac{1}{m} \sum_{i=1}^{m} (y_i - \bar{y})(y_i - \bar{y})^T$$

and

$$S w_j = \lambda_j w_j \quad \text{for eigenvalues } \lambda_1 \dots \lambda_K$$

The PC representation is

$$x_i = W^T (y_i - \bar{y})$$

where $\underbrace{\frac{1}{m} \sum_{i=1}^{m} x_i x_i^T}_{\text{Covariance}}$ is diagonal $\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_K \end{pmatrix}$

Of all orthogonal linear projections $x_i = W^T(y_i - \bar{y})$, the PC projection minimizes the squared reconstruction error $\sum_{i=1}^{m} \| y_i - \hat{y}_i \|$, where the optimal linear reconstruction of y_i is given by $\hat{y}_i = W x_i + \bar{y}$

<u>Question</u>: Where is the probabilistic model for the observed data?

Factor analysis

$$y_i = Wx_i + \mu + \epsilon_i \qquad x_i \sim N(0, I_k)$$
$$\epsilon \sim N(0, \Psi)$$
$$\Psi = \begin{pmatrix} \psi_1 & & 0 \\ & \ddots & \\ 0 & & \psi_k \end{pmatrix}$$

$$\Rightarrow \quad y_i \sim N(\mu, WW^T + \Psi)$$

PCA treats covariance & variance identically
FA treats covariance & variance separately

When $\Psi = \sigma^2 I_k$ is known $\Rightarrow$ ML=LS $\Rightarrow$ PC solution emerges

Young (1940), Whittle (1952), Basilevsky (1994, 361–363)

PROBABILISTIC PCA

$$y_i = Wx_i + \mu + \epsilon_i \qquad x_i \sim N(0, I_k)$$
$$\epsilon_i \sim N(0, \sigma^2 I_p)$$

$$\Rightarrow \quad y_i | x_i \sim N(Wx_i + \mu;\ \sigma^2 I_k)$$
$$y_i \sim N(\mu, C) \qquad C = WW^T + \sigma^2 I_p$$

$$\mathcal{L}(w, \sigma^2 | \text{data}) = -\frac{n}{2}\left\{ k \log 2\pi + \log|C| + \operatorname{tr}(C^{-1}S) \right\}$$

where $S \equiv S_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})(y_i - \hat{\mu})^T \qquad \hat{\mu} = \mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i$

EM for factor analysis (Rubin & Thayer, 1982)

PC distributions

(*) $x_i | y_i \sim N(M^{-1}W^T(y_i-\mu), \sigma^2 M^{-1})$

where $\underset{K\times K}{M} = W^T W + \sigma^2 I_K \Rightarrow W_{ML} = U_K(\Lambda_K - \sigma^2 I_K)^{1/2} R$

U_K: First K eigenvectors of S

$\Lambda_K = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_K \end{pmatrix}$: First K eigenvectors of S

R: Arbitrary KxK orthogonal (rotation) matrix

$\sigma^2_{ML} = \frac{1}{P-K} \sum_{j=K+1}^{P} \lambda_j$ is the variance lost in the projection across the lost dimensions.

PPCA has a couple of interesting features:

i) It is covariant under rotation of the original data axes as in PCA

ii) Neither factors in a 2-factor model is the same as the factor in a 1-factor model. That is not true for PPCA.

Back to (*)

$E(x_i|y_i) = M^{-1} W_{ML}^T (y_i - \mu_{ML})$; $M^{-1} \xrightarrow[\sigma\to 0]{} (W_{ML}^T W_{ML})^{-1}$

so $E(x_i|y_i)$ represents an orthogonal projection into the latent space an PCA is recovered.

In general, $W_{ML} E(x_i | y_i) + \mu_{ML}$ is not an orthogonal projection of y_i.

An orthogonal recoustruction is

$$W_{ML} (W_{ML}^T W_{ML})^{-1} M \, E(x_i | y_i) + \mu_{ML}$$

FA ≡ Generally applied to elucidate an explanation of data

PPCA ≡ Mechanism for probabilistic dimension reduction, or as a variable complexity predictive density model.

Ca' Foscari University of Venice

2024 ISBA World Meeting
July 1-7, 2024 - Venice

→ Spike-and-slab Bayesian sparse PCA
Ning & Ning (2024) STACO, 34, 112 (16 pages)

→ Bayesian inference on PCA using RJMCMC
Zhang, Chan, Kwok & Yeung (2004) AAAI

→ Bayesian PCA (1999)
Bishop, In Advances in Neural Information Processing Systems (NIPS) 11, volume 11, 382–388

→ Sparse PCA (2006)
Zou, Hastie & Tibshirani, JCGS, 15(2), 265–286.

→ Bayesian estimation of PC for Functional data
Suarez & Ghosal (2017), BA, 12(2), 311–333.