0.1 Variables

$d = 1 \dots D$ - Dimension of data space
$q = 1 \dots Q$ - Dimension of latent/embedded space
$n = 1 \dots N$ - Number of data points

0.2 Centered Data

Single centered data point

$$\mathbf{x}_n \in \Re^{D \times 1} \tag{1}$$

Expected value:

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \mathbf{0} \tag{2}$$

Matrix of all data points

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} \in \Re^{N \times D} \tag{3}$$

Covariance Maxtrix

$$\mathbf{S} = \frac{1}{N} \mathbf{X}^T \mathbf{X} \in \Re^{D \times D} \tag{4}$$

Inner Product Maxtrix

$$\mathbf{X}\mathbf{X}^T = \begin{bmatrix} \mathbf{x}_1^T \mathbf{x}_1 & \cdots & \mathbf{x}_1^T \mathbf{x}_N \\ \vdots & \ddots & \vdots \\ \mathbf{x}_N^T \mathbf{x}_1 & \cdots & \mathbf{x}_N^T \mathbf{x}_N \end{bmatrix} \in \Re^{N \times N} \tag{5}$$

0.3 Latent Variables / Points in Embedded Space

Single point

$$\mathbf{y}_n \in \Re^{Q \times 1} \tag{6}$$

Matrix of all points

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \\ \vdots \\ \mathbf{y}_N^T \end{bmatrix} \in \Re^{N \times Q} \tag{7}$$

$\mathbf{X}$ and $\mathbf{Y}$ are design matrices

0.4 Linear Mapping

Linear Mapping Matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_Q \end{bmatrix} \in \Re^{D \times Q} \tag{8}$$

$$\mathbf{w}_q \in \Re^{D \times 1} \tag{9}$$

We can use the matrix to create a linear mapping between two spaces, input and embedded. Typically,

$$Q < D \tag{10}$$

Project the data to the embedded space

$$y_{n,q} = \mathbf{x}_n^T \mathbf{w}_q \tag{11}$$

$$\mathbf{y}_n = \mathbf{W}^T \mathbf{x}_n \tag{12}$$

$$\mathbf{Y} = \mathbf{X}\mathbf{W} \tag{13}$$

Recreate the data from the embedded space representation

$$\widetilde{\mathbf{x}}_n = \sum_{q=1}^{Q} y_{n,q} \mathbf{w}_q = \mathbf{W}\mathbf{y}_n \tag{14}$$

$$\widetilde{\mathbf{X}} = \mathbf{Y}\mathbf{W}^T \tag{15}$$

Vectors w_q can be seen as basis vectors for the new embedded space.

For convenience (and without loss of generality) the vectors $\mathbf{w}_q$ can be assumed to be orthonormal

$$\mathbf{w}_i^T \mathbf{w}_j = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \tag{16}$$

Show a drawing here.

1 PCA

Find such space in which either variance is maximized or error is minimized.

1.1 Maximizing Variance

Maximize variance of the projected data

$$\underset{\mathbf{W}}{\text{argmax}} \sum_{q=1}^{Q} \sigma_q^2(\mathbf{W}) \tag{17}$$

Let's calculate covariance in the embedded space:

$$\mathbf{T} = \frac{1}{N}\mathbf{Y}^T\mathbf{Y} \tag{18}$$

$$\mathbf{Y} = \mathbf{X}\mathbf{W} \tag{19}$$

$$\mathbf{T} = \frac{1}{N}\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W} = \mathbf{W}^T\mathbf{S}\mathbf{W} \tag{20}$$

For $Q = 1$ and $D = 2$:

$$\sigma_1^2 = \mathbf{w}_1^T\mathbf{S}\mathbf{w}_1 \tag{21}$$

As we will see, this minimization can be solved using spectral approaches.

Minimize with constraint $\mathbf{w}_1^T\mathbf{w}_1 = 1$ using Lagrange multipliers method, we get unconstrained minimization of:

$$\mathbf{w}_1^T S\mathbf{w}_1 + \lambda_1(1 - \mathbf{w}_1^T\mathbf{w}_1) \tag{22}$$

Setting the derivative to 0, we get:

$$S\mathbf{w}_1 = \lambda_1\mathbf{w}_1 \tag{23}$$

This makes $\mathbf{w}_1$ an eigenvector of $\mathbf{S}$ and λ_1 the corresponding eigenvalue. We can retrieve the original variance for each value of $\mathbf{w}_1$:

$$\mathbf{w}_1^T S\mathbf{w}_1 = \lambda_1\mathbf{w}_1^T\mathbf{w}_1 = \lambda_1 = \sigma_1^2 \tag{24}$$

$$\lambda_1 = \sigma_1^2 \tag{25}$$

As a result, it is best to choose the eignvector with the largest eigenvalue to maximize the variance. This operation can be repeated iteratively and the "remaining" variance (this has a term!) will be given by:

$$\sum_{i=Q+1}^{D} \lambda_i \tag{26}$$

1.2 Minimizing Error

Maximize the projection error

$$\underset{\mathbf{W}}{\text{argmin}} \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \widetilde{\mathbf{x}}_n(\mathbf{W})\|^2 \tag{27}$$

This leads to an identical solution:

$$S\mathbf{w}_i = \lambda_i \mathbf{w}_i \tag{28}$$

with a corresponding distortion measure:

$$\sum_{i=Q+1}^{D} \lambda_i \tag{29}$$

2 Continuous Latent Variable Model

Maximum likelihood estimation is often used to find parameters of a statistical model based on a set of data samples:

$$\underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\boldsymbol{\Theta}) \tag{30}$$

If we have a latent variable model, we first have to obtain the marginal likelihood. We integraate over the latent variables:

$$p(\mathbf{x}_n|\boldsymbol{\Theta}) = \int p(\mathbf{x}_n|\mathbf{y}_n, \boldsymbol{\Theta}) p(\mathbf{y}_n) d\mathbf{y}_n \tag{31}$$

[likelihood], [prior]

3 Probabilistic PCA

Represent X using a lower dimensional set of latent variables Y. Previously, we had:

$$\widetilde{\mathbf{x}}_n = \mathbf{W}\mathbf{y}_n \tag{32}$$

$$\mathbf{x}_n = \widetilde{\mathbf{x}}_n + \epsilon_n \tag{33}$$

Now, assume a linear relationship with noise added (to model the reconstruction error):

$$\mathbf{x}_n = \mathbf{W}\mathbf{y}_n + \boldsymbol{\eta}_n \tag{34}$$

Where the noise $\boldsymbol{\eta}_n \in \Re^{Dx1}$ is assumed to be an independent sample from a spherical Gaussian distribution:

$$p(\boldsymbol{\eta}_n) = \mathcal{N}(\boldsymbol{\eta}_n|\mathbf{0}, \sigma^2\mathbf{I}) \tag{35}$$

The likelihood of an input data point can then be written as (using independence of data points):

$$p(\mathbf{X}|\mathbf{Y}, \mathbf{W}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{W}\mathbf{y}_n, \sigma^2\mathbf{I}) \tag{36}$$

To obtain the marginal likelihood, we integrate over the latent variables:

$$p(\mathbf{X}|\mathbf{W}) = \int p(\mathbf{X}|\mathbf{Y}, \mathbf{W})p(\mathbf{Y})d\mathbf{Y} \tag{37}$$

which requires us to specify a prior over $\mathbf{Y}$. To obtain a probabilistic PCA, we have to use a zero mean, unit covariance Gaussian distribution:

$$p(\mathbf{Y}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{y}_n|\mathbf{0}, \mathbf{I}) \tag{38}$$

The final marginal likelihood can be found analytically:

$$p(\mathbf{X}|\mathbf{W}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) \tag{39}$$

Parameters $\mathbf{W}$ are found through maximization of that one (Tipping and Bishop '99).

$$\underset{\mathbf{W}}{\text{argmax}} = \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{x}_n|\mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) \tag{40}$$

The result can be found analytically using spectral methods.

$$\mathbf{W} = \mathbf{U}_Q\mathbf{L}\mathbf{V}^T \qquad \mathbf{L} = (\mathbf{\Lambda}_Q - \sigma^2\mathbf{I})^{\frac{1}{2}} \tag{41}$$

Where $\mathbf{V}$ is an arbitrary rotation matrix and $\mathbf{U}_Q$ is a matrix of Q eigenvectors with largest eigenvalues $\mathbf{\Lambda}_Q$ of $\mathbf{S} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$. Therefore W consists of scaled and rotated eigenvectors of the covariance matrix S for which the eigenvalues are largest. Therefore, the model has an interpretation as a probabilistic version of PCA.

4 Dual Probabilistic PCA

A dual representation of PPCA can be achieved by marginalizing over the parameters W rather than the latent variables Y and optimizing Y rather than W.

To obtain the marginal likelihood, we integrate over the parameters:

$$p(\mathbf{X}|\mathbf{Y}) = \int p(\mathbf{X}|\mathbf{Y}, \mathbf{W})p(\mathbf{W})d\mathbf{W} \tag{42}$$

which requires us to specify a prior over $\mathbf{W}$. To obtain a probabilistic PCA, we have to use a zero mean, unit covariance Gaussian distribution:

$$p(\mathbf{W}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_d|\mathbf{0}, \mathbf{I}) \tag{43}$$

The final marginal likelihood can be found analytically:

$$p(\mathbf{X}|\mathbf{Y}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{x}_n|\mathbf{0}, \mathbf{Y}\mathbf{Y}^T + \sigma^2\mathbf{I}) \tag{44}$$

Latent variables $\mathbf{Y}$ are found through maximization of that one (Neil Lawrence IJML'05).

$$\underset{\mathbf{Y}}{\operatorname{argmax}} = \sum_{d=1}^{D} \ln \mathcal{N}(\mathbf{x}_n|\mathbf{0}, \mathbf{Y}\mathbf{Y}^T + \sigma^2\mathbf{I}) \tag{45}$$

The result can be found analytically using spectral methods.

$$\mathbf{X} = \mathbf{U'}_Q\mathbf{L}\mathbf{V}^T \qquad \mathbf{L} = (\mathbf{\Lambda}_Q - \sigma^2\mathbf{I})^{\frac{1}{2}} \tag{46}$$

Where $\mathbf{V}$ is an arbitrary rotation matrix and $\mathbf{U'}_Q$ is a matrix of Q eigenvectors with largest eigenvalues $\mathbf{\Lambda}_Q$ of $\frac{1}{D}\mathbf{X}\mathbf{X}^T$. This can be shown to be equivalent to probabilistic PCA.

5 Gaussian Processes

$$f : \Re^Q \rightarrow \Re \tag{47}$$

$$\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\} \subset \Re^Q \tag{48}$$

$$p(f(\mathbf{y}_1), f(\mathbf{y}_2), \ldots, f(\mathbf{y}_N)) \tag{49}$$

$$k(\mathbf{y}_i, \mathbf{y}_j) \tag{50}$$

$$p(f(\mathbf{y}_1), f(\mathbf{y}_2), \ldots, f(\mathbf{y}_N)) = \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{51}$$

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{y}_1, \mathbf{y}_1) & \cdots & k(\mathbf{y}_1, \mathbf{y}_N)) \\ \ldots & \ddots & \ldots \\ k(\mathbf{y}_N, \mathbf{y}_1) & \cdots & k(\mathbf{y}_N, \mathbf{y}_N)) \end{bmatrix} \tag{52}$$

$$\sigma^2\mathbf{I} \tag{53}$$

$$k(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{y}_i^T\mathbf{y}_j + \sigma^2\delta_{ij} \tag{54}$$

$$\mathbf{K} = \mathbf{Y}\mathbf{Y}^T + \sigma^2\mathbf{I} \tag{55}$$