Latent Linear Models

Tiannuo Liang Shunxing Yan

School of Mathematical Sciences
University of Science and Technology of China

Oct 8, 2020

Table of Contents

Motivation

- The correlation of variables may be caused by some latent factors. We can reduce the number of variables by attributing variables with the same essence to one latent factor.
- Sometimes we need to study some latent variables which cannot be observed directly.
  We can only observe their manifestations.

Factor Analysis

A simple example of FA model:

$$p(\mathbf{z}_i|\theta) = \mathcal{N}(\mathbf{z}_i|\mu_0, \mathbf{\Sigma}_0), \tag{1}$$
$$p(\mathbf{x}_i|\mathbf{z}_i, \theta) = \mathcal{N}(\mathbf{W}\mathbf{z}_i + \mu, \mathbf{\Psi}). \tag{2}$$

- $\mathbf{x}_i \in \mathbb{R}^D$: observation;
- $\mathbf{z}_i \in \mathbb{R}^L (L < D)$: vector of latent varibles;
- $\mathbf{W} \in \mathcal{M}_{DL}(\mathbb{R})$: factor loading matrix;
- $\mathbf{\Psi} \in \mathcal{M}_{DD}(\mathbb{R})$.
- Take $\mathbf{\Psi}$ to be diagonal, and let $\mathbf{z}_i$ explain the correlation.

Factor Analysis

$$p(\mathbf{x}_i|\theta) = \int \mathcal{N}(\mathbf{W}\mathbf{z}_i + \mu, \mathbf{\Psi})\mathcal{N}(\mathbf{z}_i|\mu_0, \mathbf{\Sigma}_0)d\mathbf{z}_i \tag{3}$$

$$= \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mu_0 + \mu, \mathbf{\Psi} + \mathbf{W}\mathbf{\Sigma}_0\mathbf{W}^T). \tag{4}$$

$$\tilde{\mathbf{W}} = \mathbf{W}\mathbf{\Sigma}_0^{-\frac{1}{2}}, \tilde{\mu} = \tilde{\mathbf{W}}\mu_0 + \mu, \tag{5}$$

$$p(\mathbf{x}_i|\theta) = \mathcal{N}(\mathbf{x}_i|\tilde{\mu}, \mathbf{\Psi} + \tilde{\mathbf{W}}\tilde{\mathbf{W}}^T). \tag{6}$$

Set $\mu_0 = \mathbf{0}, \mathbf{\Sigma}_0 = \mathbf{I}$ W.L.O.G.

Factor Analysis

FA approximates the covariance matrix of the visible vector by:

$$\mathbf{C} \triangleq cov(\mathbf{x}) = \mathbf{W}\mathbf{W}^T + \mathbf{\Psi}. \tag{7}$$

- Using $O(LD)$ parameters;
- A compromise between a full covariance Gaussian, with $O(D^2)$ parameters, and a diagonal covariance, with $O(D)$ parameters.

$$var(x_{ij}) = h_j^2 + \psi_j, \tag{8}$$

- $h_j^2 \triangleq \sum_k w_{jk}^2$: communality;
- ψ_j: uniqueness.

Posterior

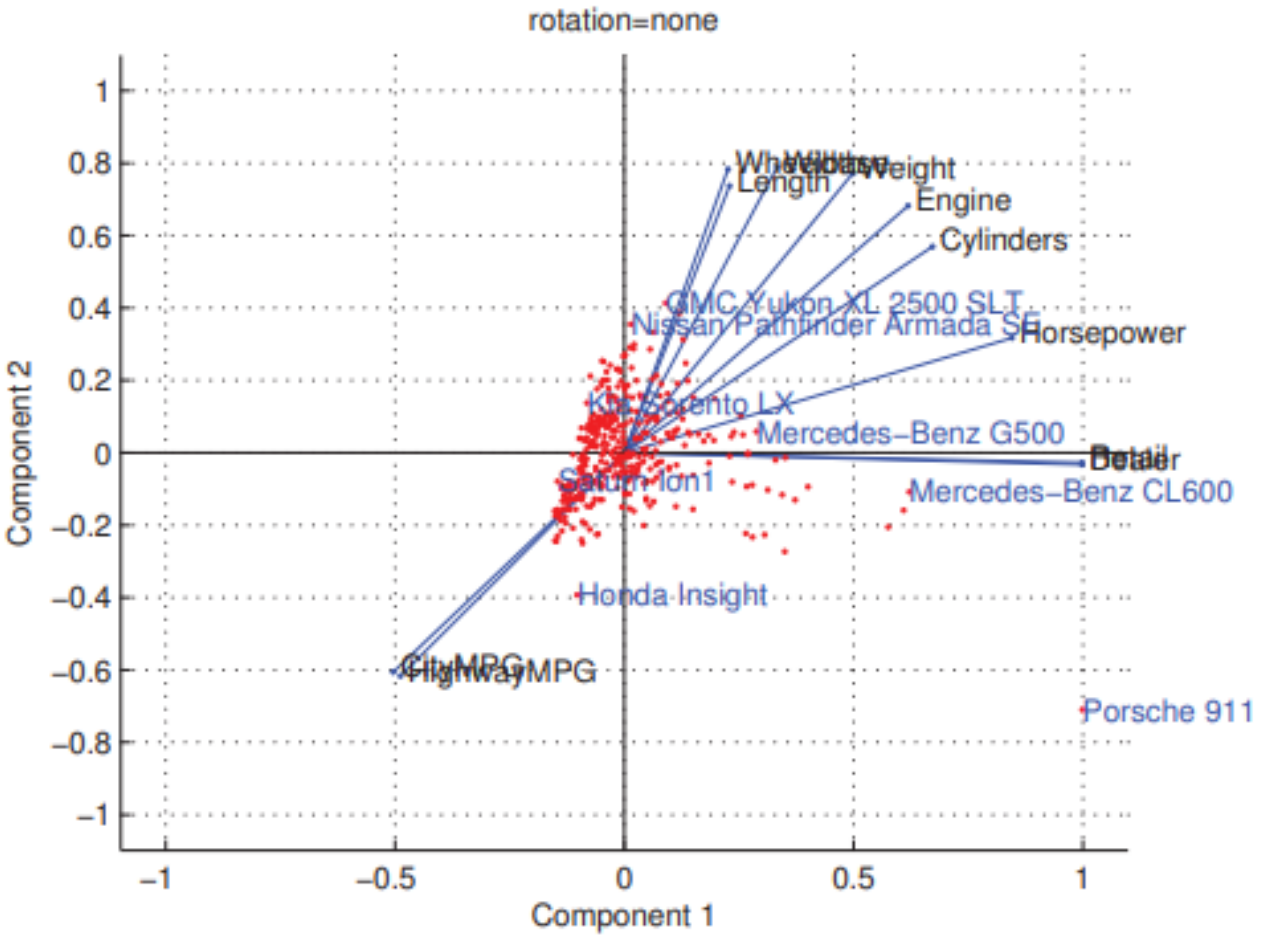
$$p(\mathbf{z}_i|\theta) = \mathcal{N}(\mathbf{z}_i|\mathbf{0}, \mathbf{I}), \tag{9}$$
$$p(\mathbf{x}_i|\mathbf{z}_i, \theta) = \mathcal{N}(\mathbf{W}\mathbf{z}_i + \mu, \mathbf{\Psi}). \tag{10}$$

If we know the true parameters:

$$p(\mathbf{z}_i|\mathbf{x}_i, \theta) = \mathcal{N}(\mathbf{z}_i|\mathbf{m}_i, \mathbf{\Sigma}_i), \tag{11}$$
$$\mathbf{\Sigma}_i = (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1} \triangleq \mathbf{\Sigma}, \tag{12}$$
$$\mathbf{m}_i = \mathbf{\Sigma}\mathbf{W}^T\mathbf{\Psi}^{-1}(\mathbf{x}_i - \mu). \tag{13}$$

$\mathbf{m}_i$: latent factors.

Biplot



$$\mathbf{m}_i = \mathbf{\Sigma W}^T \mathbf{\Psi}^{-1}(\mathbf{x}_i - \mu). \tag{14}$$

Unidentifiability

Let $\mathbf{R} \in \mathcal{M}_{LL}(\mathbb{R})$ be an orthogonal matrix, and $\tilde{\mathbf{W}} = \mathbf{WR}$, then

$$\tilde{\mathbf{C}} = \tilde{\mathbf{W}}\tilde{\mathbf{W}}^T + \mathbf{\Psi} = \mathbf{W}\mathbf{W}^T + \mathbf{\Psi} = \mathbf{C}. \tag{15}$$

- To ensure a unique solution, we need to remove $L(L-1)/2$ degrees of freedom.

$$D + LD - L(L-1)/2 \leq D(D+1)/2, \tag{16}$$

$$L \leq (2D + 1 - \sqrt{8D+1})/2. \tag{17}$$

Solutions

- Force **W** to be orthonormal, and order the columns by decreasing variance of the corresponding latent factors.
  (Principal component method)
- Force **W** to be lower triangular, ensuring that the k'th visible feature is only generated by the first k latent factors.
  (The first L visible features must be chosen carefully)
- ICA;
- Sparse factor analysis;
- Factor rotation.
  Find a rotation matrix **R** which encourages **WR** to be sparse, increasing the interpretability.

Factor Rotation

Variable	Estimated factor loadings F_1	F_2	Communalities $\hat{h}_i^2$
1. Gaelic	.553	.429	.490
2. English	.568	.288	.406
3. History	.392	.450	.356
4. Arithmetic	.740	−.273	.623
5. Algebra	.724	−.211	.569
6. Geometry	.595	−.132	.372

Factor Rotation

Variable	Estimated rotated factor loadings F_1^*	F_2^*	Communalities $\hat{h}_i^{*2} = \hat{h}_i^2$
1. Gaelic	.369	.594	.490
2. English	.433	.467	.406
3. History	.211	.558	.356
4. Arithmetic	.789	.001	.623
5. Algebra	.752	.054	.568
6. Geometry	.604	.083	.372

Factor Rotation

- Varimax

$$\mathbf{W}^* = \mathbf{WR} = (w_{jk}^*),\ \tilde{w}_{jk}^* = \frac{w_{jk}^*}{h_j^*}, \tag{18}$$

$$V = \sum_k [\frac{1}{D} \sum_j (\tilde{w}_{jk}^{*2} - \frac{1}{D} \sum_i \tilde{w}_{ik}^{*2})^2], \tag{19}$$

$$\mathbf{R}^* = argmax_{\mathbf{R}} V. \tag{20}$$

Overall factors?

Factor Rotation

- Quartimax

$$V = \sum_j [\frac{1}{L} \sum_k (\tilde{w}_{jk}^{*2} - \frac{1}{L} \sum_i \tilde{w}_{ji}^{*2})^2], \tag{21}$$

$$\mathbf{R}^* = argmax_{\mathbf{R}} V. \tag{22}$$

Parameter Estimation

$$p(\mathbf{x}_i|\theta) = \mathcal{N}(\mathbf{x}_i|\mu, \mathbf{\Psi} + \mathbf{W}\mathbf{W}^T). \tag{23}$$

- MLE

$$\log p(\mathbf{X}|\mu, \mathbf{W}, \mathbf{\Psi}) = -\frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_i (\mathbf{x}_i - \mu)^T \mathbf{C}^{-1}(\mathbf{x}_i - \mu) \tag{24}$$

$$= -\frac{N}{2}(\ln|\mathbf{C}| + tr(\mathbf{C}^{-1}\hat{\mathbf{\Sigma}})), \tag{25}$$

$$\hat{\mathbf{\Sigma}} \triangleq \frac{1}{N}\sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T. \tag{26}$$

Parameter Estimation

$$p(\mathbf{x}_i|\theta) = \mathcal{N}(\mathbf{x}_i|\mu, \mathbf{\Psi} + \mathbf{W}\mathbf{W}^T). \tag{27}$$

- MLE

$$\hat{\mu} = \bar{\mathbf{x}}, \tag{28}$$

$$\hat{\mathbf{\Sigma}}\hat{\mathbf{\psi}}^{-1}\hat{\mathbf{W}} = \hat{\mathbf{W}}(\mathbf{I} + \hat{\mathbf{W}}^T\hat{\mathbf{\psi}}^{-1}\hat{\mathbf{W}}), \tag{29}$$

$$\hat{\mathbf{\Psi}} = diag(\hat{\mathbf{\Sigma}} - \hat{\mathbf{W}}\hat{\mathbf{W}}^T). \tag{30}$$

Constrain $\hat{\mathbf{W}}^T\hat{\mathbf{\psi}}^{-1}\hat{\mathbf{W}}$ to be diagonal.

Parameter Estimation

$$p(\mathbf{x}_i|\theta) = \mathcal{N}(\mathbf{x}_i|\mu, \mathbf{\Psi} + \mathbf{W}\mathbf{W}^T). \tag{31}$$

- Principal Component Method

$$\hat{\mathbf{W}} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}, \tag{32}$$

$$\hat{\mathbf{\Psi}} = diag(\hat{\mathbf{\Sigma}} - \hat{\mathbf{W}}\hat{\mathbf{W}}^T). \tag{33}$$

- $\mathbf{\Lambda} \in \mathcal{M}_{LL}(\mathbb{R})$: the diagonal matrix of the first L eigenvalues of $\hat{\mathbf{\Sigma}}$;
- $\mathbf{V} \in \mathcal{M}_{DL}(\mathbb{R})$: the first L eigenvectors of $\hat{\mathbf{\Sigma}}$.

Parameter Estimation

$$p(\mathbf{x}_i|\theta) = \mathcal{N}(\mathbf{x}_i|\mu, \mathbf{\Psi} + \mathbf{W}\mathbf{W}^T). \tag{34}$$

- Iterated Principal Factor Method
  (Based on the empirical correlation matrix $\hat{\mathbf{R}}$)
- 1.$\hat{\mathbf{R}}^{(k)} = \hat{\mathbf{R}} - \hat{\mathbf{\Psi}}^{(k-1)}$;
- 2.$\hat{\mathbf{W}}^{(k)} = \mathbf{V}^{(k)}\mathbf{\Lambda}^{(k)\frac{1}{2}}$;
  $\mathbf{\Lambda}^{(k)}$: the diagonal matrix of the first L eigenvalues of $\hat{\mathbf{R}}^{(k)}$;
  $\mathbf{V}^{(k)}$: the first L eigenvectors of $\hat{\mathbf{R}}^{(k)}$;
- 3.$\hat{\mathbf{\Psi}}^{(k)} = diag(\hat{\mathbf{R}} - \hat{\mathbf{W}}^{(k)}\hat{\mathbf{W}}^{(k)T})$;
- Repeat 1-3 until convergence.

Parameter Estimation

- Iterated Principal Factor Method

The initial estimation $\hat{\boldsymbol{\Psi}}^{(0)}$:

- $\hat{\boldsymbol{\Psi}}^{(0)} = diag(\hat{\mathbf{R}}^{-1})^{-1}$;
- $\hat{\psi}_j^{(0)} = 1 - \max_{k \neq j} r_{jk}^2$;
- $\hat{\boldsymbol{\Psi}}^{(0)} = \mathbf{O}$.

Mixture of Factor Analysers (MFA)

$$p(\mathbf{x}_i|\mathbf{z}_i, q_i = c, \theta) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_c + \mathbf{W}_c\mathbf{z}_i, \boldsymbol{\Psi}), \tag{35}$$

$$p(\mathbf{z}_i|\theta) = \mathcal{N}(\mathbf{z}_i|\mathbf{0}, \mathbf{I}), \tag{36}$$

$$p(q_i|\theta) = \mathrm{Cat}(q_i|\pi). \tag{37}$$

Mixture of Factor Analysers (MFA)

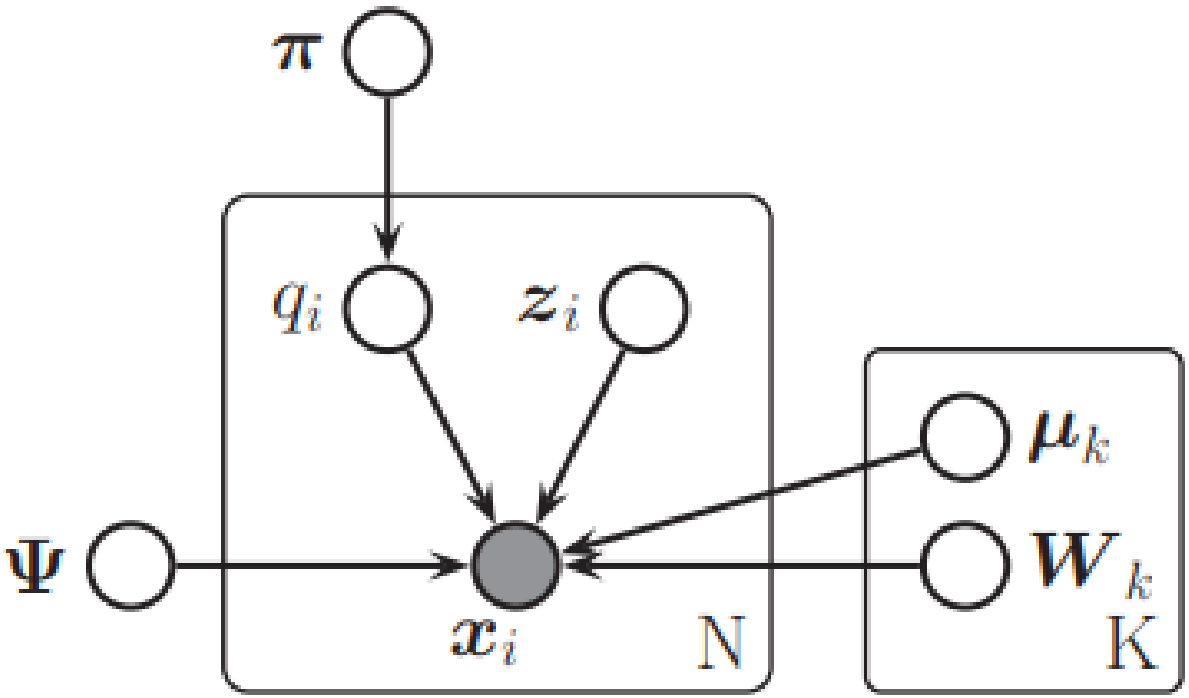


Figure 12.3 Mixture of factor analysers as a DGM.

EM

- E step

$$r_{ic} \triangleq p(q_i = c|\mathbf{x}_i, \boldsymbol{\theta}) \propto \pi_c \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_c, \mathbf{W}_c\mathbf{W}_c^T + \boldsymbol{\Psi}), \tag{38}$$

$$p(\mathbf{z}_i|\mathbf{x}_i, q_i = c, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}_i|\mathbf{m}_{ic}, \boldsymbol{\Sigma}_{ic}), \tag{39}$$

$$\boldsymbol{\Sigma}_{ic} = (\mathbf{I} + \mathbf{W}_c^T\boldsymbol{\Psi}^{-1}\mathbf{W}_c)^{-1} \triangleq \boldsymbol{\Sigma}_c, \tag{40}$$

$$\mathbf{m}_{ic} = \boldsymbol{\Sigma}_c\mathbf{W}_c^T\boldsymbol{\Psi}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c). \tag{41}$$

EM

- M step

Define

$$\tilde{\mathbf{W}}_c \triangleq (\mathbf{W}_c, \boldsymbol{\mu}_c), \tag{42}$$

$$\tilde{\mathbf{z}}_i \triangleq \begin{pmatrix} \mathbf{z}_i \\ 1 \end{pmatrix}, \tag{43}$$

$$\mathbf{b}_{ic} \triangleq \mathbb{E}(\tilde{\mathbf{z}}_i | \mathbf{x}_i, q_i = c) = \begin{pmatrix} \mathbf{m}_{ic} \\ 1 \end{pmatrix}, \tag{44}$$

$$\mathbf{C}_{ic} \triangleq \mathbb{E}(\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T | \mathbf{x}_i, q_i = c) = \begin{pmatrix} \boldsymbol{\Sigma}_{ic} + \mathbf{m}_{ic}\mathbf{m}_{ic}^T & \mathbf{m}_{ic} \\ \mathbf{m}_{ic}^T & 1 \end{pmatrix}. \tag{45}$$

EM

- M step

$$\hat{\tilde{\mathbf{W}}}_c = (\sum_i r_{ic}\mathbf{x}_i\mathbf{b}_{ic}^T)(\sum_i r_{ic}\mathbf{C}_{ic})^{-1}, \tag{46}$$

$$\hat{\boldsymbol{\Psi}} = \frac{1}{N} diag[\sum_{i,c} r_{ic}(\mathbf{x}_i - \hat{\tilde{\mathbf{W}}}_c\mathbf{b}_{ic})\mathbf{x}_i^T], \tag{47}$$

$$\hat{\pi}_c = \frac{1}{N}\sum_i r_{ic}. \tag{48}$$

Motivation

$\mathbf{x}_i$ lies in a D-dimensional linear space.
Consider a rank L linear approximation(Assume that $\bar{\mathbf{x}} = \mathbf{0}$):

$$\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{z}_i. \tag{49}$$

- $\mathbf{W} \in \mathcal{M}_{DL}(\mathbb{R})$: an orthonormal set of L linear basis vectors $\mathbf{w}_j \in \mathbb{R}^D$;
- $\mathbf{z}_i \in \mathbb{R}^L$: the corresponding score.

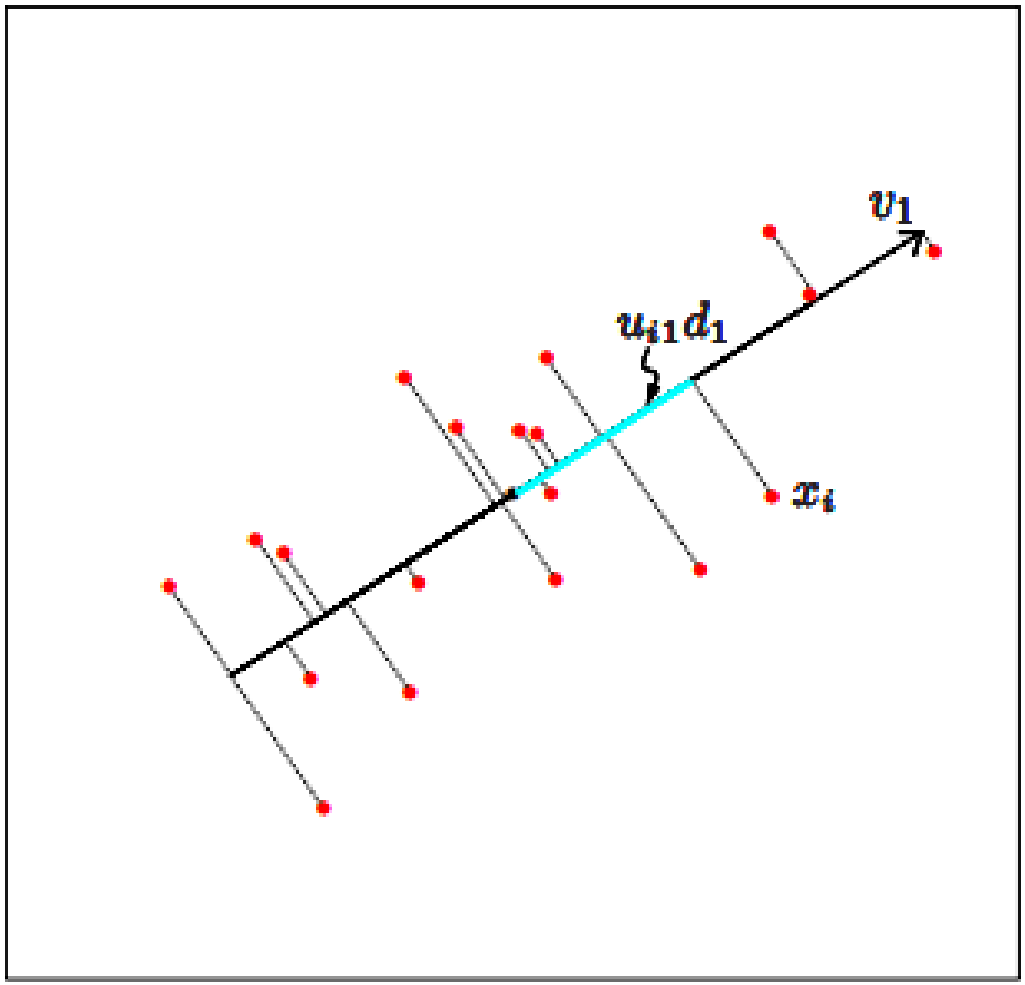
Principle Components Analysis

Define the average reconstruction error:

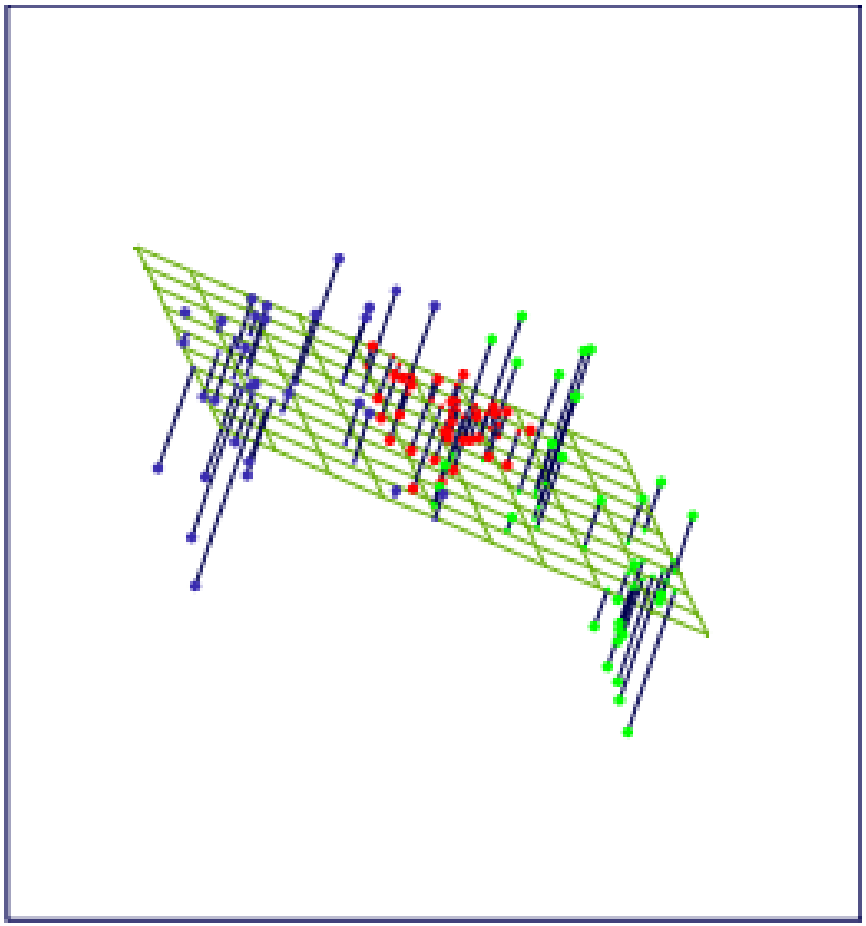
$$J(\mathbf{W}, \mathbf{Z}) \triangleq \frac{1}{N} \sum_i ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2 = \frac{1}{N} ||\mathbf{X}^T - \mathbf{W}\mathbf{Z}^T||_F^2. \tag{50}$$

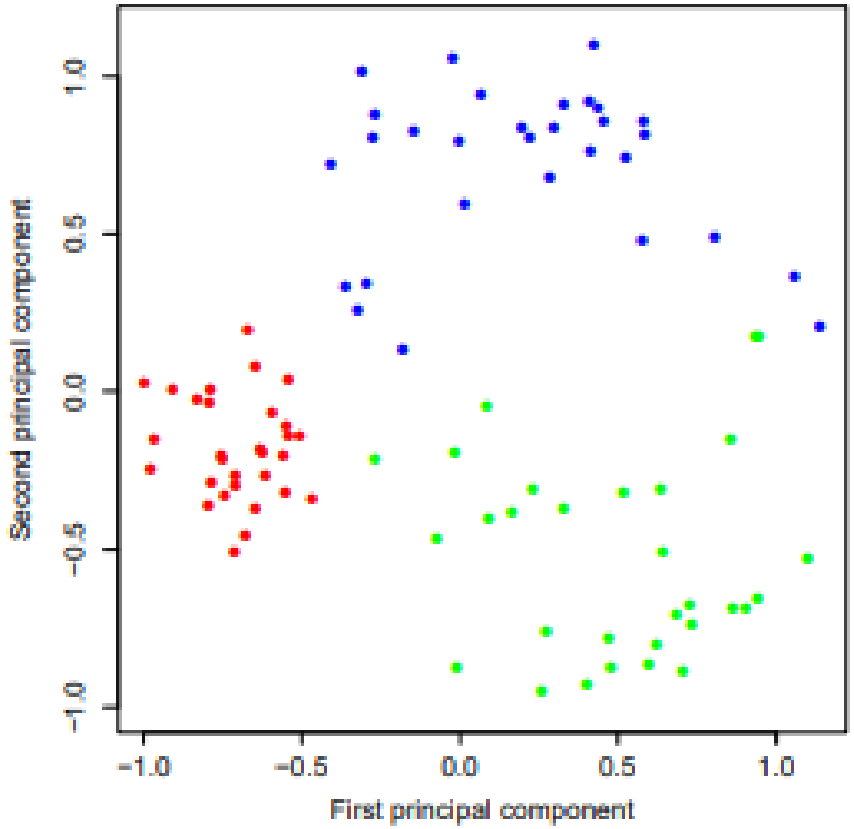
- $\mathbf{X} \in \mathcal{M}_{ND}(\mathbb{R})$: a matrix with $\mathbf{x}_i$ in its rows;
- $\mathbf{Z} \in \mathcal{M}_{NL}(\mathbb{R})$: a matrix with $\mathbf{z}_i$ in its rows;
- $||\mathbf{A}||_F^2 \triangleq \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{tr(\mathbf{A}^T\mathbf{A})}$.

Principle Components Analysis



Principle Components Analysis





Principle Components Analysis

Theorem

The optimal solution of $J(\mathbf{W}, \mathbf{Z})$ *is* $\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{W}}$*, and* $\hat{\mathbf{W}}$ *contains the first* L *eigenvectors of* $\hat{\mathbf{\Sigma}}$*.*

Proof

Let $\tilde{\mathbf{z}}_j = (z_{1j}, \ldots, z_{Nj})^T$, the j'th component of all the scores.

- $L = 1$

$$J(\mathbf{w}_1, \tilde{\mathbf{z}}_1) = \frac{1}{N}\sum_i ||\mathbf{x}_i - z_{i1}\mathbf{w}_1||^2 = C + \frac{1}{N}\sum_i(-2z_{i1}\mathbf{w}_1^T\mathbf{x}_i + z_{i1}^2) \quad (51)$$

since $||\mathbf{w}_1|| = 1$. Take derivatives wrt z_{i1} and equate to 0:

$$\frac{\partial J}{\partial z_{i1}} = \frac{1}{N}(-2\mathbf{w}_1^T\mathbf{x}_i + 2z_{i1}) = 0 \Longrightarrow z_{i1} = \mathbf{w}_1^T\mathbf{x}_i, \quad (52)$$

$$J(\mathbf{w}_1) = C - \frac{1}{N}\sum_i \mathbf{w}_1^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w}_1 = C - \mathbf{w}_1^T\hat{\mathbf{\Sigma}}\mathbf{w}_1. \quad (53)$$

Proof

$$\tilde{J}(\mathbf{w}_1) = -\mathbf{w}_1^T \hat{\mathbf{\Sigma}} \mathbf{w}_1 + \lambda_1 (\mathbf{w}_1^T \mathbf{w}_1 - 1), \tag{54}$$

$$\frac{\partial \tilde{J}}{\partial \mathbf{w}_1} = -2\hat{\mathbf{\Sigma}} \mathbf{w}_1 + 2\lambda_1 \mathbf{w}_1 = \mathbf{0}, \tag{55}$$

$$\mathbf{w}_1^T \hat{\mathbf{\Sigma}} \mathbf{w}_1 = \lambda_1 . \tag{56}$$

- The optimal direction is the eigenvector corresponding to the largest eigenvalue of $\hat{\mathbf{\Sigma}}$.

Proof

- $L = 2$

$$J(\mathbf{w}_1, \tilde{\mathbf{z}}_1, \mathbf{w}_2, \tilde{\mathbf{z}}_2) = \frac{1}{N} \sum_i ||\mathbf{x}_i - z_{i1}\mathbf{w}_1 - z_{i2}\mathbf{w}_2||^2 \quad (57)$$

$$= C + \frac{1}{N} \sum_i (-2z_{i1}\mathbf{w}_1^T \mathbf{x}_i + z_{i1}^2 - 2z_{i2}\mathbf{w}_2^T \mathbf{x}_i + z_{i2}^2) \quad (58)$$

since $\mathbf{w}_1^T \mathbf{w}_2 = 0$.

- Optimizing wrt $\mathbf{w}_1, \mathbf{z}_1$ gives the same solution.
- $\frac{\partial J}{\partial z_{i2}} = 0 \Longrightarrow z_{i2} = \mathbf{w}_2^T \mathbf{x}_i$.

Proof

$$J(\mathbf{w}_2) = C - \mathbf{w}_2^T \hat{\mathbf{\Sigma}} \mathbf{w}_2, \tag{59}$$

$$\tilde{J}(\mathbf{w}_2) = -\mathbf{w}_2^T \hat{\mathbf{\Sigma}} \mathbf{w}_2 + \lambda_2(\mathbf{w}_2^T \mathbf{w}_2 - 1) + \lambda_{12} \mathbf{w}_1^T \mathbf{w}_2, \tag{60}$$

$$\frac{\partial \tilde{J}}{\partial \mathbf{w}_2} = -2\hat{\mathbf{\Sigma}} \mathbf{w}_2 + 2\lambda_2 \mathbf{w}_2 + \lambda_{12} \mathbf{w}_1 = \mathbf{0}, \tag{61}$$

$$-2\mathbf{w}_1^T \hat{\mathbf{\Sigma}} \mathbf{w}_2 + 2\lambda_2 \mathbf{w}_1^T \mathbf{w}_2 + \lambda_{12} \mathbf{w}_1^T \mathbf{w}_1 = 0 \Longrightarrow \lambda_{12} = 0, \tag{62}$$

$$\hat{\mathbf{\Sigma}} \mathbf{w}_2 = \lambda_2 \mathbf{w}_2. \tag{63}$$

The proof continues in this way.

Remarks

- $$z_{i1} = \mathbf{w}_1^T \mathbf{x}_i, \tag{64}$$

  The optimal reconstruction weights are obtained by orthogonally projecting the data onto $\mathbf{w}_j$.

- $$J(\mathbf{w}_1) = C - \mathbf{w}_1^T \hat{\mathbf{\Sigma}} \mathbf{w}_1, \tag{65}$$

  The variance of z_{i1} is

$$\frac{1}{N} \sum_i z_{i1}^2 - (\frac{1}{N} \sum_i z_{i1})^2 \tag{66}$$

$$= \frac{1}{N} \sum_i \mathbf{w}_1^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{w}_1 - (\frac{1}{N} \mathbf{w}_1^T \sum_i \mathbf{x}_i)^2 = \mathbf{w}_1^T \hat{\mathbf{\Sigma}} \mathbf{w}_1. \tag{67}$$

Remarks

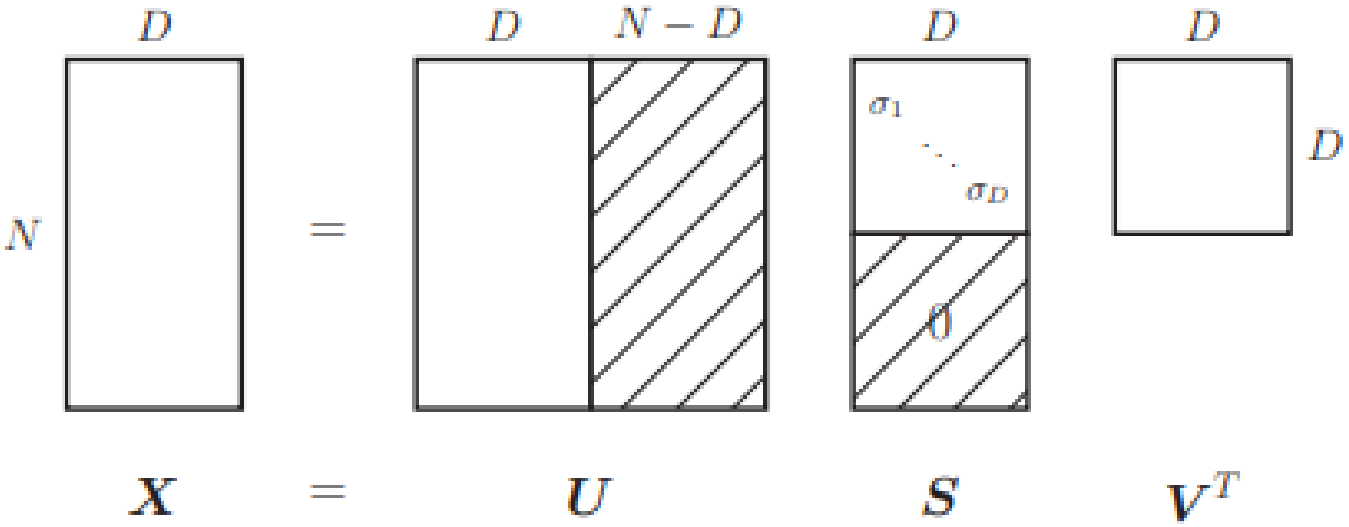
- Minimizing the reconstruction error is equivalent to maximizing the variance of the projected data.
- It is standard practice to standardize the data first, or equivalently, to work with correlation matrices instead of covariance matrices.

Singular Value Decomposition (SVD)

$$\mathbf{X} = \mathbf{USV}^T. \tag{68}$$

- $\mathbf{X} \in \mathcal{M}_{ND}(\mathbb{R})$;
- $\mathbf{U} \in \mathcal{M}_{NN}(\mathbb{R})$: orthonormal matrices with left singular vectors in columns;
- $\mathbf{V} \in \mathcal{M}_{DD}(\mathbb{R})$: orthonormal matrices with right singular vectors in columns;
- $\mathbf{S} \in \mathcal{M}_{ND}(\mathbb{R})$: $r = min(N, D)$ singular values $\sigma_i \geq 0$ on the main diagonal and 0s filling the rest of the matrix.

Economy Sized SVD



$$\mathbf{X} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^T. \tag{69}$$

- $N > D$:
  $\hat{\mathbf{U}} \in \mathcal{M}_{ND}(\mathbb{R}), \hat{\mathbf{S}}, \hat{\mathbf{V}} \in \mathcal{M}_{DD}(\mathbb{R})$.
- $N < D$:
  $\hat{\mathbf{U}}, \hat{\mathbf{S}} \in \mathcal{M}_{NN}(\mathbb{R}), \hat{\mathbf{V}} \in \mathcal{M}_{ND}(\mathbb{R})$.

Connection with Eigenvectors

Let $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ be an economy sized SVD of $\mathbf{X}$.

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T\mathbf{U}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T, \tag{70}$$

$$(\mathbf{X}\mathbf{X}^T)\mathbf{U} = \mathbf{U}\mathbf{D}. \tag{71}$$

- $\mathbf{D} = \mathbf{S}^2$;
- The eigenvalues of $\mathbf{X}\mathbf{X}^T$ are equal to the squared singular values;
- The eigenvectors of $\mathbf{X}\mathbf{X}^T$ are equal to $\mathbf{U}$.

Connection with Eigenvectors

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{V}\mathbf{D}\mathbf{V}^T, \tag{72}$$

$$(\mathbf{X}^T\mathbf{X})\mathbf{V} = \mathbf{V}\mathbf{D}. \tag{73}$$

- The eigenvalues of $\mathbf{X}^T\mathbf{X}$ are equal to the squared singular values;
- The eigenvectors of $\mathbf{X}^T\mathbf{X}$ are equal to $\mathbf{V}$.

Truncated SVD

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T. \tag{74}$$

- $r = rank(\mathbf{X})$;
- If σ_i's die off quickly, we have a rank L approximation using $(N + D + 1)L$ parameters:

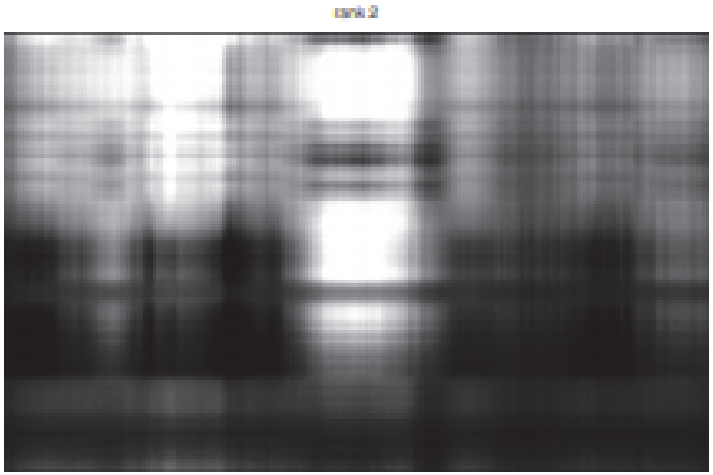
$$\mathbf{X} \approx \mathbf{X}_L = \mathbf{U}_{:,1:L}\mathbf{S}_{1:L,1:L}\mathbf{V}_{:,1:L}^T = \sum_{i=1}^{L} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{75}$$

$$\mathbf{X} - \mathbf{X}_L = \mathbf{U}_{:,(L+1):r}\mathbf{S}_{(L+1):r,(L+1):r}\mathbf{V}_{:,(L+1):r}^T, \tag{76}$$

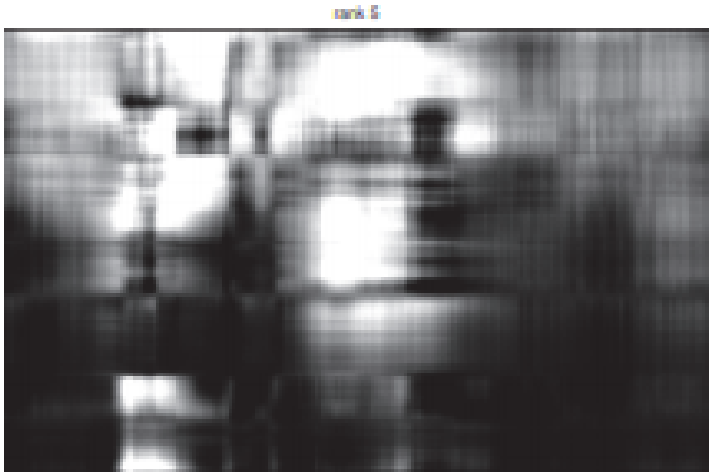
$$||\mathbf{X} - \mathbf{X}_L||_F^2 = tr((\mathbf{X} - \mathbf{X}_L)^T(\mathbf{X} - \mathbf{X}_L)) = \sum_{i=L+1}^{r} \sigma_i^2. \tag{77}$$

Truncated SVD

(a)



(b)



(c)

(d)

PCA and SVD

Let $\mathbf{X}_L = \mathbf{U}\mathbf{S}\mathbf{V}^T$ be a truncated SVD of $\mathbf{X}$.
Notice that $\hat{\mathbf{\Sigma}} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$.

- $\mathbf{\Lambda} = \frac{1}{N}\mathbf{D}$;
- $\hat{\mathbf{W}} = \mathbf{V}$;
- $\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{W}} = \mathbf{U}\mathbf{S}$;
- $\hat{\mathbf{X}} = \hat{\mathbf{Z}}\hat{\mathbf{W}}^T = \mathbf{X}_L$.

Probabilistic PCA(PPCA)

FA model:

$$p(\mathbf{z}_i|\theta) = \mathcal{N}(\mathbf{z}_i|\mathbf{0}, \mathbf{I}), \tag{78}$$

$$p(\mathbf{x}_i|\mathbf{z}_i, \theta) = \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}). \tag{79}$$

- Constrain $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$, and $\mathbf{W}$ to be orthonormal: probabilistic PCA.
- Assume that $\bar{\mathbf{x}} = \mathbf{0}$.

PPCA

- The observed data log-likelihood:

$$\log p(\mathbf{X}|\mathbf{W}, \sigma^2) = -\frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_i \mathbf{x}_i^T \mathbf{C}^{-1}\mathbf{x}_i \tag{80}$$

$$= -\frac{N}{2}(\ln|\mathbf{C}| + tr(\mathbf{C}^{-1}\hat{\mathbf{\Sigma}})). \tag{81}$$

- The maxima of the log-likelihood:

$$\hat{\mathbf{W}} = \mathbf{V}(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R}, \tag{82}$$

- $\mathbf{R} \in \mathcal{M}_{LL}(\mathbb{R})$: an arbitrary orthogonal matrix.
  Set $\mathbf{R} = \mathbf{I}$ W.L.O.G.

PPCA

- The MLE of σ^2:

$$\hat{\sigma}^2 = \frac{1}{D-L}\sum_{j=L+1}^{D}\lambda_j. \tag{83}$$

- The posterior over the latent factors:

$$p(\mathbf{z}_i|\mathbf{x}_i,\hat{\theta}) = \mathcal{N}(\mathbf{z}_i|\hat{\mathbf{F}}^{-1}\hat{\mathbf{W}}^T\mathbf{x}_i,\sigma^2\hat{\mathbf{F}}^{-1}), \tag{84}$$

$$\hat{\mathbf{F}} = \hat{\mathbf{W}}^T\hat{\mathbf{W}} + \hat{\sigma}^2\mathbf{I}. \tag{85}$$

- $\sigma^2 \to 0$: classical PCA.

EM

- $\tilde{\mathbf{Z}} \in \mathcal{M}_{LN}(\mathbb{R})$: storing the posterior means along columns;
- $\tilde{\mathbf{X}} = \mathbf{X}^T$.
- E step: $\tilde{\mathbf{Z}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\tilde{\mathbf{X}}$;
- M step: $\mathbf{W} = \tilde{\mathbf{X}}\tilde{\mathbf{Z}}^T(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T)^{-1}$.

Categorical PCA

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{86}$$

$$p(\mathbf{y}_i|\mathbf{z}_i, \theta) = \prod_r \mathrm{Cat}(y_{ir}|\mathcal{S}(\mathbf{W}_r\mathbf{z}_i + \mathbf{w}_{0r})). \tag{87}$$

- $y_{ir} \in \{1, \cdots, C\}(r = 1 : R)$;
- $\mathbf{w}_{0r}$: the offset term.

FA/PPCA

$$L^* = argmax_L \; p(L|D). \tag{88}$$

- Evaluating the marginal likelihood is difficult.
  (Use approximations or CV)
- Large number of models.
  (Set the model to the maximal size, and use automatic relevancy determination/EM)

PCA

A proxy for the likelihood is the reconstruction error:

$$E(\mathcal{D}, L) = \sum_i ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2. \tag{89}$$

- $\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \mathbf{z}_i = \mathbf{W}^T(\mathbf{x}_i - \boldsymbol{\mu})$;
- $\mathbf{W}, \boldsymbol{\mu}$ are estimated from $\mathcal{D}_{train}$.

$$E(\mathcal{D}_{train}, L) = \sum_{j=L+1}^{D} \lambda_j. \tag{90}$$

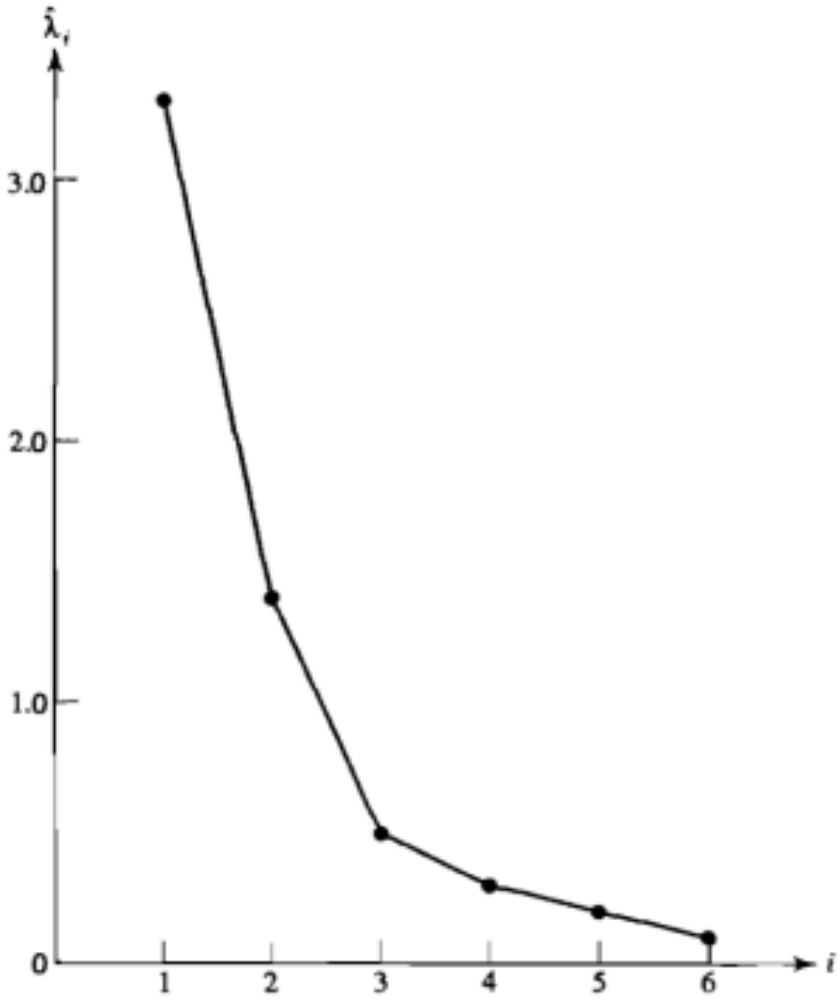
PCA

The fraction of variance explained:

$$F(\mathcal{D}_{train}, L) = \frac{\sum_{j=1}^{L} \lambda_j}{\sum_{j'=1}^{L_{max}} \lambda_{j'}}. \tag{91}$$

- Plot $E(\mathcal{D}_{train}, L) \sim L$.
- Scree plot: $\lambda_j \sim j$;
  Look for an elbow in the plot.
- $L^* = \min\{L | F(\mathcal{D}_{train}, L) > c\}$.

PCA



Profile Likelihood

Let λ_k be some measure of the error incurred by a model of size k, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{L_{max}}$.

$$\lambda_k \sim \mathcal{N}(\mu_1, \sigma^2) \; i.i.d. (k \leq L), \tag{92}$$

$$\lambda_k \sim \mathcal{N}(\mu_2, \sigma^2) \; i.i.d. (k > L). \tag{93}$$

Profile Likelihood

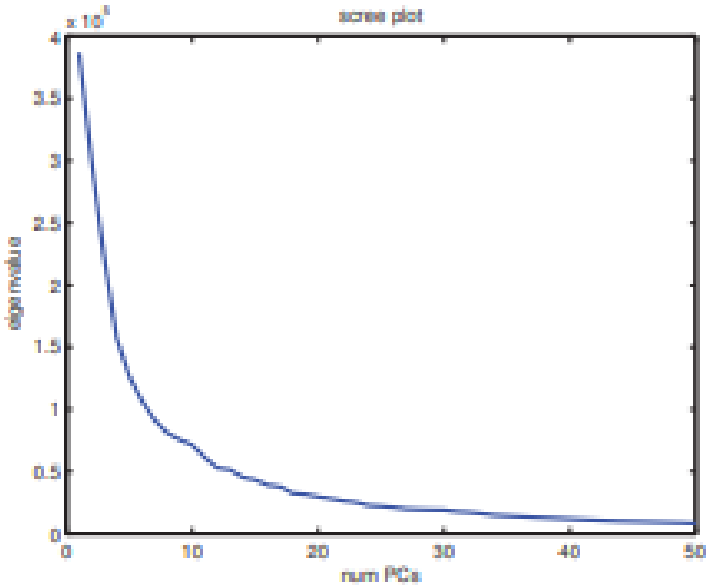
$$\mu_1(L) = \frac{1}{L}\sum_{k\leq L}\lambda_k, \mu_2(L) = \frac{1}{L_{max}-L}\sum_{k>L}\lambda_k, \tag{94}$$

$$\sigma^2(L) = \frac{1}{L_{max}}[\sum_{k\leq L}(\lambda_k - \mu_1(L))^2 + \sum_{k>L}(\lambda_k - \mu_2(L))^2], \tag{95}$$

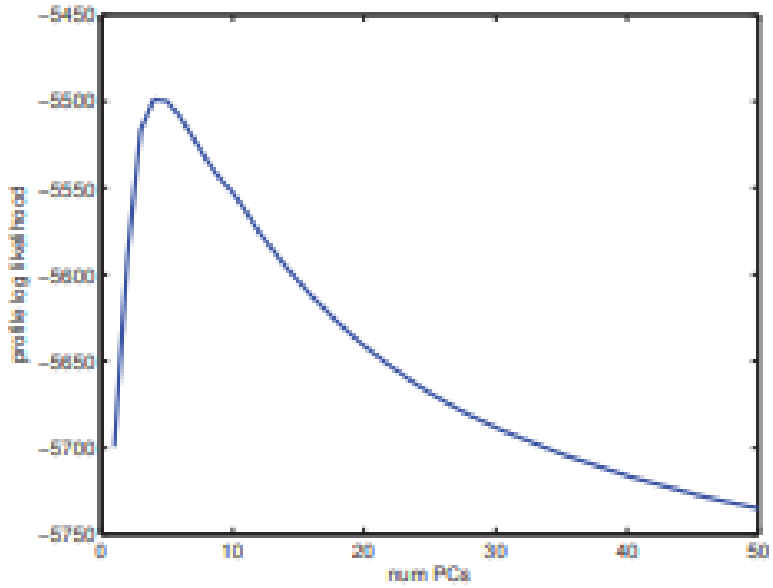
$$l(L) = \sum_{k\leq L}\log\mathcal{N}(\lambda_k|\mu_1(L),\sigma^2(L)) + \sum_{k>L}\log\mathcal{N}(\lambda_k|\mu_2(L),\sigma^2(L)), \tag{96}$$

$$L^* = argmax_L\ l(L). \tag{97}$$

Profile Likelihood



(a)



(b)

Outline

Canonical correlation analysis

In statistics, canonical correlation analysis (CCA), also called canonical variates analysis, is a way of inferring information from covariance matrices.
If we have two vectors (WLOG to be mean 0 and var 1) $X = (X_1, \ldots, X_{D_x})$ and $Y = (Y_1, \ldots, Y_{D_y})$ of random variables, and there are correlations among the variables, then canonical correlation analysis will find linear combinations of X and Y which have maximum correlation with each other.

Canonical correlation analysis

CCA seeks vectors W_x and W_y such that the random variables $W_x'X$ and $W_y'Y$ maximize the correlation

$$\rho_1 = \text{corr}\left(W_x'X, W_y'Y\right).$$

The random variables $U = W_x'X$ and $V = W_y'Y$ are the first pair of canonical variables.
Then one seeks vectors maximizing the same correlation subject to the constraint that they are to be uncorrelated with the first pair of canonical variables; this gives the second pair of canonical variables. This procedure may be continued up to $\min\{dim(X), dim(Y)\}$ times.

Computation

Seek for

$$(W_x, W_y) = \arg \max_{W_x, W_y \neq 0} \text{corr}\left(W_x' X, W_y' Y\right)$$

which is equivalent to

$$max : cov(W_x' X, W_y' Y) = W_x' S_{XY} W_y$$

$$s.t.\ var(W_x' X) = var(W_y' Y) = 1$$

Computation

By Lagrange multiplier method, we define $G(W_x, W_y)$

$$W_x' S_{XY} W_y - \frac{1}{2}\lambda_1 \left(W_x' S_{XX} W_x - 1\right) - \frac{1}{2}\lambda_2 \left(W_y' \Sigma_{YY} W_y - 1\right)$$

Let the partial derivatives for W_x and W_y to be 0 separately, we get

$$S_{XY} W_y - \lambda_1 S_{XX} W_x = 0$$

$$S_{YX} W_x - \lambda_2 S_{YY} W_y = 0$$

therefore:

$$\lambda_1 = \lambda_1 W_x' S_{XX} W_x = W_x' S_{XX} W_y = \lambda_2$$

Computation

Let $\lambda = \lambda_1 = \lambda_2$, we get

$$S_{XY} S_{YY}^{-1} S_{YX} W_x - \lambda^2 S_{XX} W_x = 0$$

$$S_{YX} S_{XX}^{-1} S_{XY} W_y - \lambda^2 S_{YY} W_y = 0$$

then W_x and W_y are eigenvectors of

$$M_1 = S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX}$$

$$M_2 = S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY}$$

about eigenvalue λ^2

Computation

Let

$$K = S_{XX}^{-1/2} S_{XY} S_{YY}^{-1/2}$$

$$\alpha = S_{XX}^{1/2} W_x$$

$$\beta = S_{YY}^{1/2} W_y$$

then

$$KK'\alpha = \lambda^2 \alpha$$

$$K'K\beta = \lambda^2 \beta$$

and

$$W_x = S_{XX}^{-1/2} \alpha$$

$$W_y = S_{XX}^{-1/2} \beta$$

Computation

More generally, if we'd k-th canonical variables

$$max : W_x' S_{XY} W_y$$

$$s.t.\ W_{xk}' S_{XX} W_{xk} = W_{yk}' S_{YY} W_{yk} = 1$$

$$W_{xk}' S_{XX} W_{xi} = W_{yk}' S_{YY} W_{yi} = 0, i \leq k-1$$

It to get by mathematical induction

$$W_{xk} = S_{XX}^{-1/2} \alpha_k$$

$$W_{yk} = S_{XX}^{-1/2} \beta_k$$

where α_k is the k-th eigenvector of KK' and β_k is the k-th eigenvector of $K'K$.

Large Sample Inferences

Consider hypothesis testing problem

$$H_0 : \rho_{t+1} = \rho_{t+2} = ... = 0$$

Under the null hypothesis,

$$-[n - \frac{1}{2}(p + q + 3)] \sum_{i \geq t+1} log(1 - \hat{\rho}_i)$$

is asymptotically $\chi^2_{(D_x - t)(D_y - t)}$

Regularized CCA

In the previous formula derivation process, we used S_{XX}^{-1}.
But we know it causes problems when it's singular.
One way to solve this problem is Regularized CCA, where we replace S_{XX} and S_{YY} with $S_{XX} + aI$ and $S_{YY} + bI$

We can use cross-validation to estimate (a, b).
Let

$$CV\,(a,b) = \text{corr}\left(\left\{W_x(a,b)^{(-i)\prime}x_i\right\}_{i=1}^{n}, \left\{W_y(a,b)^{(-i)\prime}y_i\right\}_{i=1}^{n}\right)$$

then we choose

$$(\hat{a}, \hat{b}) = argmax\ CV(a,b)$$

CCA in probabilistic perspective

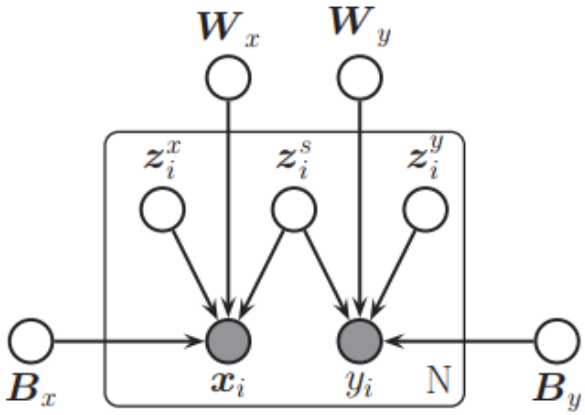
Consider the following model:

$$p(Z) = N(0, I_L) = N(Z^s|0, I_{Ls})N(Z^x|0, I_{Lx})N(Z^y|0, I_{Ly})$$

$$p(X|Z) = N(\hat{W}_x z^s + B_x z^x + \mu_x, \sigma^2 I_{Dx})$$

$$p(Y|Z) = N(\hat{W}_y z^s + B_y z^x + \mu_y, \sigma^2 I_{Dy})$$

just like the figure:



CCA in probabilistic perspective

CCA allows each view to have its own private subspace, but there is also a shared subspace.

Since the model is jointly Gaussian, we have:

$$(Y, X) \sim N(\mu, \hat{W}\hat{W}^T + \sigma^2 I)$$

where $\mu = (\mu_Y, \mu_X)$ and $\hat{W} = \begin{pmatrix} \hat{W}_x & B_x & 0 \\ \hat{W}_y & 0 & B_y \end{pmatrix}$

A simple application

The olive oil data in R package "classifly" consists of the percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) found in the lipid fraction of 572 Italian olive oils. There are 9 collection areas, 4 from southern Italy (North and South Apulia, Calabria, Sicily), two from Sardinia (Inland and Coastal) and 3 from northern Italy (Umbria, East and West Liguria).

Region	Area	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
1	North-Apulia	1075	75	226	7823	672	36	60	29
1	North-Apulia	1088	73	224	7709	781	31	61	29
1	North-Apulia	911	54	246	8113	549	31	63	29
1	North-Apulia	966	57	240	7952	619	50	78	35
1	North-Apulia	1051	67	259	7771	672	50	80	46
1	North-Apulia	911	49	268	7924	678	51	70	44
1	North-Apulia	922	66	264	7990	618	49	56	29

A simple application

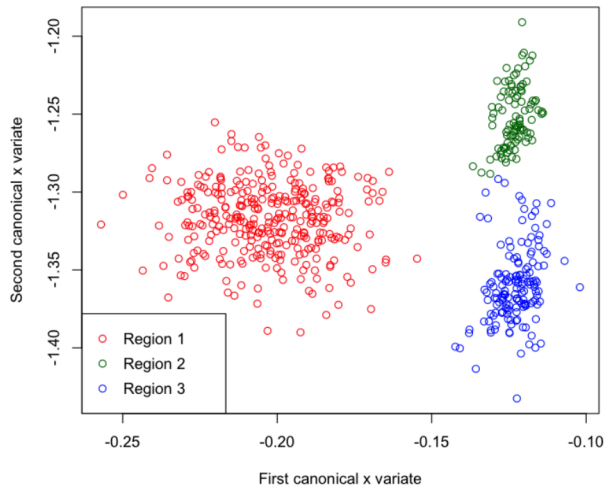
We are interested in the correlation between the three regions and fatty acid measurements. Define

$$Y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ & \dots & \end{pmatrix}$$

where $y_{ij} = 1$ iff the i-th sample is from region j.

A simple application

Make canonical correlation analysis of X and Y. Then we visualize data based on that.



That shows the classification effect is good.

Outline

Principal components regression

As we learned in Chapter 7 linear regression, PCR is a regression analysis technique based on PCA.
In PCR, instead of regressing the dependent variable directly, some principal components Z_i of the input variables X are used as regressors. And often the principal components with higher variances are selected.

Principal components regression

One major use of PCR lies in overcoming the multicollinearity problem which arises when two or more of the explanatory variables are close to being collinear. PCR can aptly deal with such situations by excluding some of the low-variance principal components in the regression step. However, for the purpose of predicting the outcome, the principal components with low variances may also be important, in some cases even more important.

Principal components regression

That exposes one shortcoming of PCR: the selection of the principal component Z_i has nothing to do with the output variables Y, only depends on input variables X.

Partial least squares

Partial least squares regression solves this problem. The PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space.

We call the model PLS1 when $D_y = 1$

PLS1

The mth PLS direction $\hat{\phi}_m$ solves:

$$max : cov_\alpha(X\alpha, Y)$$

$$s.t. \|\alpha\| = 1, \alpha' S_{XX} \hat{\varphi}_\ell = 0, \ell = 0, 1 \leq j < i$$

We can easily get the result:

PLS1

Algorithm 3.3 *Partial Least Squares.*

1. Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$, and $\mathbf{x}_j^{(0)} = \mathbf{x}_j,\ j = 1, \ldots, p$.

2. For $m = 1, 2, \ldots, p$

   (a) $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj}\mathbf{x}_j^{(m-1)}$, where $\hat{\varphi}_{mj} = \langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \rangle$.

   (b) $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$.

   (c) $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$.

   (d) Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$: $\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle]\mathbf{z}_m,\ j = 1, 2, \ldots, p$.

3. Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_1^p$. Since the $\{\mathbf{z}_\ell\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

PLS1 in probabilistic perspective

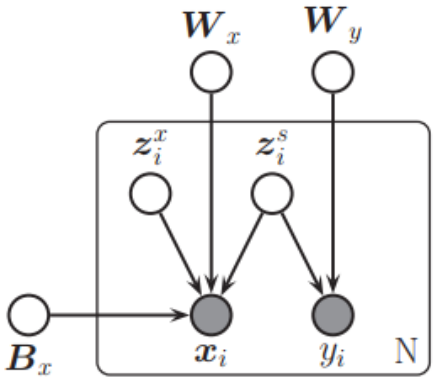
Consider the following model:

$$p(z) = N(0, I_L) = N(z^s|0, I_{Ls})N(z^x|0, I_{Lx})$$

$$p(y|z) = N(\hat{W}_y^T z^s + \mu_y, \sigma I_{Dy}^2)$$

$$p(x|z) = N(\hat{W}_x z + B_x z^x + \mu_x, \sigma^2 I_{Dx})$$

just like the figure:



PLS1 in probabilistic perspective

That is called partial least squares (PLS) . It is an asymmetric or more discriminative form of supervised PCA. The key idea is to allow some of the (co)variance in the input features to be explained by its own subspace, z^x , and to let the rest of the subspace, z^s , be shared between input and output.

Since the model is jointly Gaussian, we have:

$$(Y, X) \sim N(\mu, \hat{W}\hat{W}^T + \sigma^2 I)$$

where $\mu = (\mu_y, \mu_x)$ and $\hat{W} = \begin{pmatrix} \hat{W}_y & 0 \\ \hat{W}_x & B_x \end{pmatrix}$

PCA-CCA-PLS

Table 1. Cost functions optimized by the different methods

PCA	Maximize variance	$\frac{\mathbf{w}'\mathbf{S}_{\mathbf{XX}}\mathbf{w}}{\mathbf{w}'\mathbf{w}}$
		$\mathbf{w}'\mathbf{S}_{\mathbf{XX}}\mathbf{w}$ s.t. $\|\mathbf{w}\|^2 = 1$
	Minimize residuals	$\|(\mathbf{I} - \mathbf{w}\mathbf{w}')\mathbf{X}\|_F^2$
CCA	Maximize correlation	$\frac{\mathbf{w}_{\mathbf{X}}'\mathbf{S}_{\mathbf{XY}}\mathbf{w}_{\mathbf{Y}}}{\sqrt{\mathbf{w}_{\mathbf{X}}'\mathbf{S}_{\mathbf{XX}}\mathbf{w}_{\mathbf{X}}}\sqrt{\mathbf{w}_{\mathbf{Y}}'\mathbf{S}_{\mathbf{YY}}\mathbf{w}_{\mathbf{Y}}}}$
	Maximize fit	$\mathbf{w}_{\mathbf{X}}'\mathbf{S}_{\mathbf{XY}}\mathbf{w}_{\mathbf{Y}}$ s.t. $\|\mathbf{X}\mathbf{w}_{\mathbf{X}}\|^2 = \|\mathbf{Y}\mathbf{w}_{\mathbf{Y}}\|^2 = 1$
	Minimize misfit	$\|\mathbf{w}_{\mathbf{X}}'\mathbf{X} - \mathbf{w}_{\mathbf{Y}}'\mathbf{Y}\|^2$ s.t. $\|\mathbf{X}\mathbf{w}_{\mathbf{X}}\|^2 = \|\mathbf{Y}\mathbf{w}_{\mathbf{Y}}\|^2 = 1$
PLS	Maximize covariance	$\frac{\mathbf{w}_{\mathbf{X}}'\mathbf{S}_{\mathbf{XY}}\mathbf{w}_{\mathbf{Y}}}{\sqrt{\mathbf{w}_{\mathbf{X}}'\mathbf{w}_{\mathbf{X}}}\sqrt{\mathbf{w}_{\mathbf{Y}}'\mathbf{w}_{\mathbf{Y}}}}$
	Maximize fit	$\mathbf{w}_{\mathbf{X}}'\mathbf{S}_{\mathbf{XY}}\mathbf{w}_{\mathbf{Y}}$ s.t. $\|\mathbf{w}_{\mathbf{X}}\|^2 = \|\mathbf{w}_{\mathbf{Y}}\|^2 = 1$
	Minimize misfit	$\|\mathbf{w}_{\mathbf{X}}'\mathbf{X} - \mathbf{w}_{\mathbf{Y}}'\mathbf{Y}\|^2$ s.t. $\|\mathbf{w}_{\mathbf{X}}\|^2 = \|\mathbf{w}_{\mathbf{Y}}\|^2 = 1$

Outline

Motivation

Assume we are in a room where two people are speaking simultaneously. We have two microphones, which we hold in different locations. The microphones give us two recorded time signals, which we could denote by $x_1(t)$ and $x_2(t)$, with x_1 and x_2 the amplitudes, and t the time index. Each of these recorded signals is a weighted sum of the speech signals emitted by the two speakers, which we denote by $z_1(t)$ and $z_2(t)$. We could express this as a linear equation:

$$x_1(t) = a_{11}z_1(t) + a_{12}z_2(t)$$

$$x_2(t) = a_{21}z_1(t) + a_{22}z_2(t)$$

where a_{11},a_{12},a_{21}, and a_{22} are some parameters that depend on the distances of the microphones from the speakers. It would be very useful if we can estimate the two original speech signals $z_1(t)$ and $z_2(t)$, using only the recorded signals $x_1(t)$ and $x_2(t)$.
This is called the cocktail-party problem.

Independent Component Analysis (ICA)

Consider the following model:
Let x_t be the observed signal at the sensors at time t, and z_t be the vector of source signals (non-Gaussian and independent).

$$x_t = Wz_t + \epsilon_t$$

where W is an $D \times L$ matrix, and $\epsilon_t \sim N(0, \Psi)$.
Our goal is to infer the source signals $p(z_t|x_t, \theta)$.

Often we will assume the noise level to be zero and the variance of the source distributions to be 1.

An Example

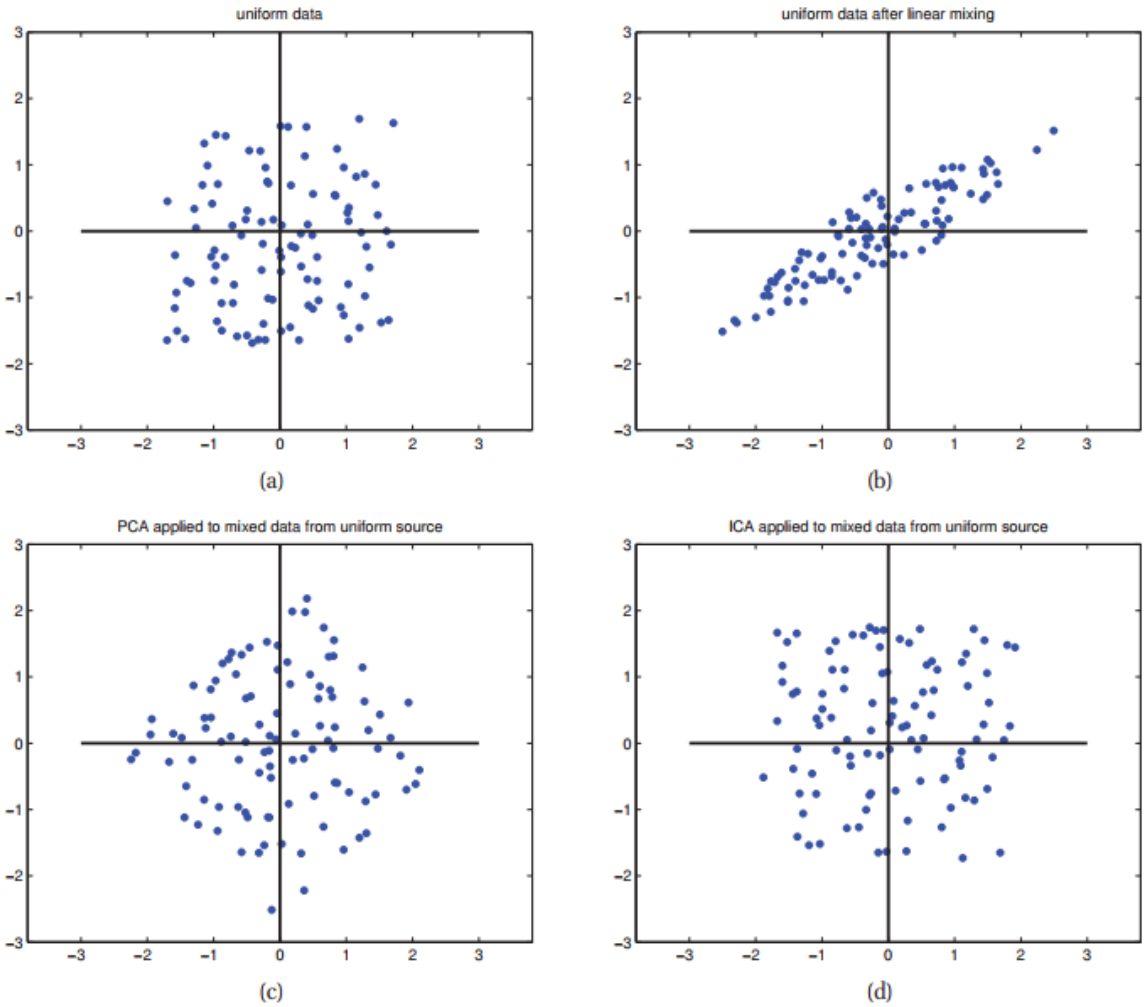


Figure 12.21 Illustration of ICA and PCA applied to 100 iid samples of a 2d source signal with a uniform distribution. (a) Latent signals. (b) Observations. (c) PCA estimate. (d) ICA estimate. Figure generated by `icaDemoUniform`, written by Aapo Hyvarinen.

Independent Component Analysis (ICA)

In a sense, PCA solves half of the problem, since it identifies the linear subspace; all that ICA has to do is then to identify the appropriate rotation.
Hence we see that ICA is not that different from methods such as varimax, which seek good rotations of the latent factors to enhance interpretability.

Outline

Principles of ICA estimation

Principles of ICA estimation

- nongaussianity
- mutual information
- likelihood

Nonguassianity

The Central Limit Theorem tells that the distribution of a sum of independent random variables tends toward a gaussian distribution, under certain conditions. Thus, so a sum of two independent random variables usually has a distribution that is closer to gaussian than any of the two original random variables.
Thus, maximizing the nongaussianity of $w^T x$ gives us one of the independent components.

Measures of nongaussianity:Kurtosis

The classical measure of nongaussianity is kurtosis, which we define:

$$kurt(y) = E[y^4] - 3\{E[y^2]\}^2$$

Random variables that have a negative kurtosis are called subgaussian, and those with positive kurtosis are called supergaussian.
If x_1 and x_2 are two independent random variables, it holds:

$$kurt(x_1 + x_2) = kurt(x_1) + kurt(x_2)$$

$$kurt(\alpha x) = \alpha^4 kurt(x)$$

Measures of nongaussianity:Kurtosis

Let's look at a 2-dimensional model $x = Wz$,
consider $y = Ax = AWz = k_1 z_1 + k_2 z_2$,

$$kurt(y) = k_1^4 kurt(z_1) + k_2^4 kurt(z_2)$$

when

$$1 = E[y^2] = k_1^2 + k_2^2$$

It is trivial that the maximize are at the points when one of the k_i is 1 and others are 0;Thus we only need to find A which maximizes kurt(Ax).

However, kurtosis can be very sensitive to outliers (Huber, 1985). Its value may depend on only a few observations in the tails of the distribution, which may be erroneous or irrelevant observations.

Negentropy

Negentropy is based on the informationtheoretic quantity of entropy.
We define :

$$negentropy(y) = H(N(\mu, \sigma^2)) - H(z)$$

where $\mu = E[y]$ and $\sigma^2 = var(y)$
We can define our objective as maximizing

$$J(V) = \sum_j negentropy(y_j) = \sum_j E[logp(y_j)] + constant$$

Approximations of negentropy

The estimation of negentropy is difficult, therefore this contrast function remains mainly a theoretical one. In practice, some approximation have to be used.

- The classical method of approximating negentropy is using higher-order moments, for example as follows (Jones and Sibson, 1987):

$$negentropy(y) \approx \frac{1}{12}\{E[y^3]\}^2 + \frac{1}{48}[kurt(y)]^2$$

y is zero mean and unit variance. However, the validity of such approximations may be rather limited. Maybe it is not rubust.

Mutual Information

One measure of dependence of a set of random variables is the multi-information:

$$I(x) = KL(p(x)||\prod_j p(x_j)) = \sum_j H(x_j) - H(x)$$

An important property of mutual information (Papoulis, 1991; Cover and Thomas, 1991) is that we have for an invertible linear transformation y = Ax:

$$I(y_1, ..., y_n) = \sum_i H(y_i) - H(x) - log(|det(W)|)$$

There is a strong relation between negentropy and mutual information.

Likelihood

Assume we have known the p_i which are the density functions of the z_i.
Let $V = W^{-1}$; these are often called the recognition weights, as opposed to W, which are the generative weights.
Since x = Wz, we have:

$$p_x(Wz_t) = p_z(z_t)\left|\det\left(W^{-1}\right)\right| = p_z(Vx_t)\,|\det(V)|$$

Hence we can write the log-likelihood, assuming T i.i.d. samples

$$\frac{1}{T}\log p(\mathcal{D} \mid V) = \log|\det(V)| + \frac{1}{T}\sum_{j=1}^{L}\sum_{t=1}^{T}\log p_j\left(v_j^T x_t\right)$$

Since the data z and x are standardized and centralized, W is orthogonal.

Likelihood

So we only need to minimize :

$$\mathrm{NLL}(V) = \sum_{j=1}^{L} E\left[G_j\left(z_j\right)\right]$$

where

$$z_j = v_j^T x$$

and

$$G_j(z) \triangleq -\log p_j(z)$$

A popular alternative is to use an approximate Newton method, which is called The FastICA algorithm.

Outline

The FastICA algorithm

For simplicity of presentation, we initially assume there is only one latent factor.
denote

$$G(z) = -\log p(z)$$

and

$$g(z) = \frac{d}{dz} G(z)$$

then

$$f(v) = E\left[G\left(v^T x\right)\right] + \frac{\beta}{2}\left(1 - v^T v\right)$$

$$\nabla f(v) = E\left[x g\left(v^T x\right)\right] - \beta v$$

$$H(v) = E\left[x x^T g'\left(v^T x\right)\right] - \beta I$$

The FastICA algorithm

Since

$$H(v) = E\left[xx^T g'\left(v^T x\right)\right] - \beta I$$

Let us make the approximation

$$E\left[xx^T g'\left(v^T x\right)\right] \approx E\left[xx^T\right] E\left[g'\left(v^T x\right)\right] = E\left[g'\left(v^T x\right)\right]$$

This makes the Hessian very easy to invert, giving rise to the following Newton update:

$$v^* \triangleq v - \frac{E\left[xg\left(v^T x\right)\right] - \beta v}{E\left[g'\left(v^T x\right)\right] - \beta}$$

The FastICA algorithm

One can rewrite this in the following way

$$v^* \triangleq E\left[xg\left(v^T x\right)\right] - E\left[g'\left(v^T x\right)\right] v$$

and

$$v^{new} \triangleq \frac{v^*}{\|v^*\|}$$

One iterates this algorithm until convergence.

Since the objective is not convex, there are multiple local optima. We can use this fact to learn multiple different weight vectors or features.

Properties of the FastICA Algorithm

Properties of the FastICA Algorithm:

- The convergence is fast;
- The FastICA algorithm is robust;
- Contrary to gradient-based algorithms, there are no step size parameters to choose;
- It finds directly independent components of any non-Gaussian distribution;
- It is parallel, distributed, computationally simple, and requires little memory space;

Application