Machine Learning 2020

Volker Roth

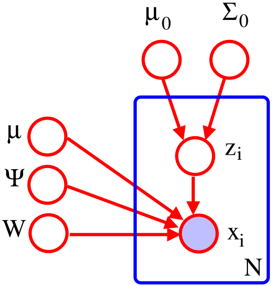
Department of Mathematics & Computer Science
University of Basel

18th May 2020

Section 9

Linear latent variable models

Factor analysis



- One problem with mixture models: **only a single latent variable**. Each observation can only come from one of K prototypes.
- Alternative: $\boldsymbol{z}_i \in \mathbb{R}^k$. Gaussian prior:

$$p(\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{z}_i | \boldsymbol{\mu}_0, \Sigma_0)$$

- For observations $\boldsymbol{x}_i \in \mathbb{R}^p$, we may use a **Gaussian likelihood.**
- As in linear regression, we assume the mean is a **linear** function:

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

  W: **factor loading matrix**, and Ψ: **covariance matrix**.
- We take Ψ to be **diagonal**, since the whole point of the model is to "force" z_i to **explain the correlation.**

Factor analysis: generative process

Generative process ($k = 1$, $p = 2$, diagonal Ψ):

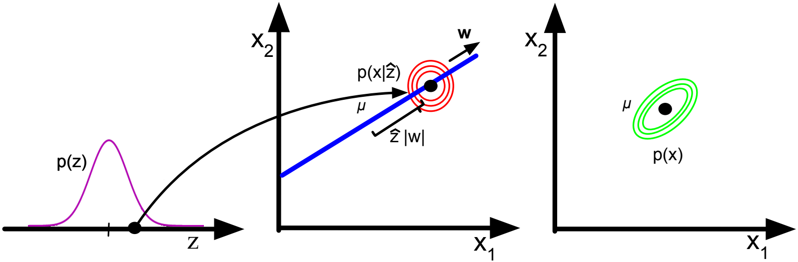


Figure 12.1 in K. Murphy

We take an isotropic Gaussian "spray can" and slide it along the 1d line defined by $wz_i + \mu$. This induces a correlated Gaussian in 2d.

Inference of the latent factors

- We hope that the latent factors z will reveal something interesting about the data $\rightsquigarrow$ compute posterior over the latent variables:

$$\begin{aligned} p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{z}_i|\boldsymbol{m}_i, \Sigma) \\ \Sigma &= (\Sigma_0^{-1} + W^t\Psi^{-1}W)^{-1} \\ \boldsymbol{m}_i &= \Sigma_i(W^t\Psi^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) + \Sigma_0^{-1}\boldsymbol{\mu}_0) \end{aligned}$$

- The posterior means $\boldsymbol{m}_i$ are called the **latent scores, or latent factors.**

Example

- Example from (Shalizi 2009). $p = 11$ variables and $n = 387$ cases describing aspects of cars: engine size, #(cylinders), miles per gallon (MPG), price, etc.
- Fit a $p = 2$ dim model. Plot m_i scores as points in $\mathbb{R}^2$.
- Try to **understand the "meaning" of latent factors:** project unit vectors $e_1 = (1, 0, \ldots, 0), e_2 = (0, 1, 0, \ldots, 0)$, etc. into the low dimensional space (blue lines)
- **Horizontal axis represents price** (features labeled "dealer" and "retail"), with expensive cars on the right.
- **Vertical axis represents fuel efficiency** (measured in terms of MPG) versus size:
  heavy vehicles: less efficient $\rightsquigarrow$ higher up,
  light vehicles: more efficient $\rightsquigarrow$ lower down.

Example

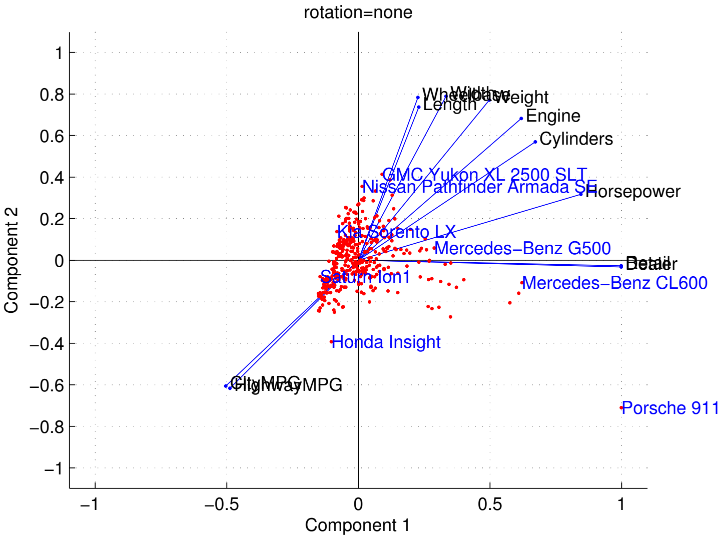
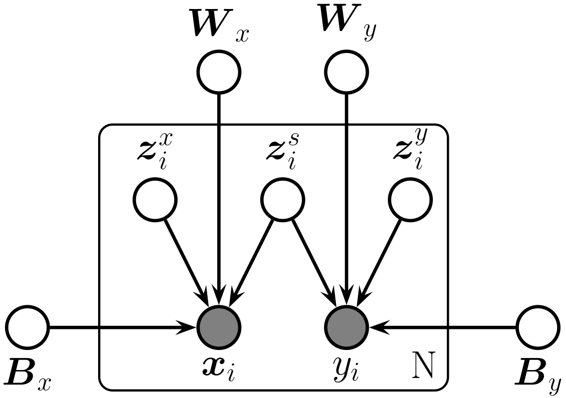


Figure 12.2 in K. Murphy

Special Cases: PCA and CCA

- Covariance matrix $\Psi = \sigma^2 I \rightsquigarrow$ (probabilistic) **PCA**.
- Two-view version involving $\boldsymbol{x}$ and $\boldsymbol{y} \rightsquigarrow$ **CCA**.



From figure 12.19 in K. Murphy

PCA and dimensionality reduction

Given n data points in p dimensions:

$$X = \begin{bmatrix} - & \mathbf{x}_1 & - \\ - & \mathbf{x}_2 & - \\ - & \vdots & - \\ - & \mathbf{x}_n & - \end{bmatrix} \in \mathbb{R}^{n \times p}$$

Want to reduce dimensionality from p to k. Choose k directions $\mathbf{w}_1, \ldots, \mathbf{w}_k$, arrange them as columns in matrix W:

$$W = \begin{bmatrix} \mathbf{w}_1, & \mathbf{w}_2, & \ldots, & \mathbf{w}_k \end{bmatrix} \in \mathbb{R}^{p \times k}$$

For each $\mathbf{w}_j$, compute **similarity** $z_j = \mathbf{w}_j^t \mathbf{x},\ j = 1 \ldots k$.
Project $\mathbf{x}$ down to $\mathbf{z} = (z_1, \ldots, z_k)^t = W^t \mathbf{x}$. How to choose W?

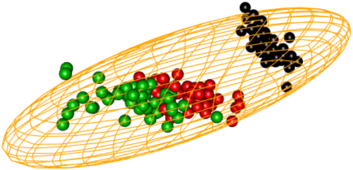
Encoding–decoding model

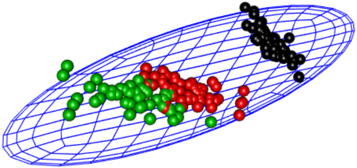
The projection matrix W serves two functions:

- **Encode:** $\mathbf{z} = W^t \mathbf{x}, \;\; \mathbf{z} \in \mathbb{R}^k, \;\; z_j = \mathbf{w}_j^t \mathbf{x}$.
  - The vectors $\mathbf{w}_j$ form a basis of the projected space.
  - We will require that this basis is orthonormal, i.e. $W^t W = I$.
- **Decode:** $\tilde{\mathbf{x}} = W\mathbf{z} = \sum_{j=1}^k z_j \mathbf{w}_j, \;\; \tilde{\mathbf{x}} \in \mathbb{R}^p$.
  - If $k = p$, the above orthonormality condition implies $W^t = W^{-1}$, and encoding can be undone without loss of information.
  - If $k < p$, least-squares problem $\rightsquigarrow$ the reconstruction error will be nonzero.
- Above we assumed that the origin of the coordinate system is in the sample mean, i.e. $\sum_i \mathbf{x}_i = 0$.

Principal Component Analysis (PCA)

In the general case, we want the reconstruction error $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|$ to be small.
Objective: minimize $\min_{W \in \mathbb{R}^{p \times k}:\, W^t W = I} \sum_{i=1}^{n} \|\boldsymbol{x}_i - WW^t \boldsymbol{x}_i\|^2$





Finding the principal components

Projection vectors are orthogonal $\rightsquigarrow$ can treat them separately:

$$\min_{\boldsymbol{w}:\, \|\boldsymbol{w}\|=1} \sum\nolimits_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{w}\boldsymbol{w}^t\boldsymbol{x}_i\|^2$$

$$\begin{aligned}
\sum\nolimits_i \|\boldsymbol{x}_i - \boldsymbol{w}\boldsymbol{w}^t\boldsymbol{x}_i\|^2 &= \sum_{i=1}^{n} [\boldsymbol{x}_i^t\boldsymbol{x}_i - 2\boldsymbol{x}_i^t\boldsymbol{w}\boldsymbol{w}^t\boldsymbol{x}_i + \boldsymbol{x}_i^t\boldsymbol{w}\underbrace{\boldsymbol{w}^t\boldsymbol{w}}_{=1}\boldsymbol{w}^t\boldsymbol{x}_i] \\
&= \sum\nolimits_i [\boldsymbol{x}_i^t\boldsymbol{x}_i - \boldsymbol{x}_i^t\boldsymbol{w}\boldsymbol{w}^t\boldsymbol{x}_i] \\
&= \sum\nolimits_i \boldsymbol{x}_i^t\boldsymbol{x}_i - \boldsymbol{w}^t \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^t\boldsymbol{w} \\
&= \underbrace{\sum\nolimits_i \boldsymbol{x}_i^t\boldsymbol{x}_i}_{\text{const.}} - \boldsymbol{w}^t X^t X \boldsymbol{w}.
\end{aligned}$$

Finding the principal components

- Want to maximize $\mathbf{w}^t X^t X \mathbf{w}$ under the constraint $\|\mathbf{w}\| = 1$
- Can also maximize the ratio $J(\mathbf{w}) = \frac{\mathbf{w}^t X^t X \mathbf{w}}{\mathbf{w}^t \mathbf{w}}$ $\rightsquigarrow$ Rayleigh quotient.
- Optimal projection $\mathbf{w}$ is the eigenvector of $X^t X$ with largest eigenvalue.
- Note that we assumed that $\sum_i \mathbf{x}_i = 0$. Thus, the columns of X are assumed to sum to zero.
  $\rightsquigarrow$ compute SVD of "centered" matrix $X = USV^t$
  $\rightsquigarrow$ singular vectors $\mathbf{v}$ are eigenvectors of $X^t X$
  $\rightsquigarrow$ they are the principal components.

Eigen-faces [Turk and Pentland, 1991]

- p = number of pixels
- Each $\boldsymbol{x}_i \in \mathbb{R}^p$ is a face image
- x_{ji} = intensity of the j-th pixel in image i

$$(X^t)_{p\times n} \approx W_{p\times k} \, (Z^t)_{k\times n}$$

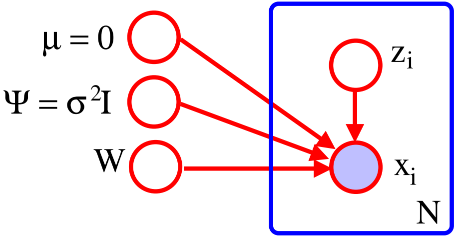
( $\dots$ ) $\approx$ () $\begin{bmatrix} | & & | \\ \boldsymbol{z}_1 & \dots & \boldsymbol{z}_n \\ | & & | \end{bmatrix}$

Idea: $\boldsymbol{z}_i$ more 'meaningful' representation of i-th face than $\boldsymbol{x}_i$

Can use $\boldsymbol{z}_i$ for nearest-neighbor classification

Much faster when $p \gg k$.

Probabilistic PCA



- Assuming $\Psi = \sigma^2 I$ and centered data in the FA model
  $\rightsquigarrow$ likelihood

$$p(\boldsymbol{x}_i | \boldsymbol{z}_i, \boldsymbol{\theta}) = \mathcal{N}(W\boldsymbol{z}_i, \sigma^2 I).$$

Probabilistic PCA

- (Tipping & Bishop 1999): Maxima of the likelihood are given by

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

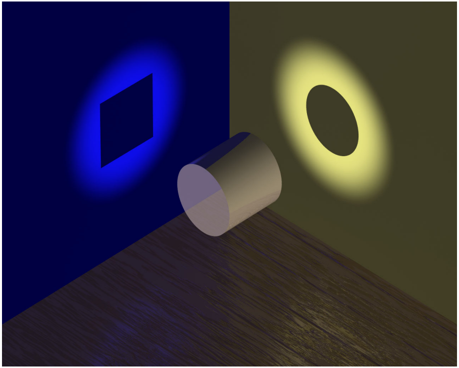
  where R is an arbitrary orthogonal matrix,
  columns of V: first k eigenvectors of $S = \frac{1}{n} X^t X$,
  Λ: diagonal matrix of eigenvalues.
- As $\sigma^2 \to 0$, we have $\hat{W} \to V$, **as in classical PCA** (for $R = \Lambda^{-\frac{1}{2}}$).
- Projections $\boldsymbol{z}_i$: Posterior over the latent factors:

  $$\begin{aligned} p(\boldsymbol{z}_i|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) &= \mathcal{N}(\boldsymbol{z}_i|\hat{\boldsymbol{m}}_i, \sigma^2 \hat{F}^{-1}) \\ \hat{F} &= \sigma^2 I + \hat{W}^t \hat{W} \\ \boldsymbol{m}_i &= \hat{F}^{-1} \hat{W}^t \boldsymbol{x}_i \end{aligned}$$

  For $\sigma^2 \to 0$, $\boldsymbol{z}_i \to \boldsymbol{m}_i$ and $\boldsymbol{m}_i \to (V^t V)^{-1} V^t \boldsymbol{x}_i = V^t \boldsymbol{x}_i$
  $\rightsquigarrow$ orthogonal projection of the data onto the column space of V,
  as in classical PCA.

Multiple Views: CCA

- Consider paired samples from different views.
- What is the dependency structure between the views ?
- Standard approach: global linear dependency detected by CCA.



Canonical Correlation Analysis [Hotelling, 1936]

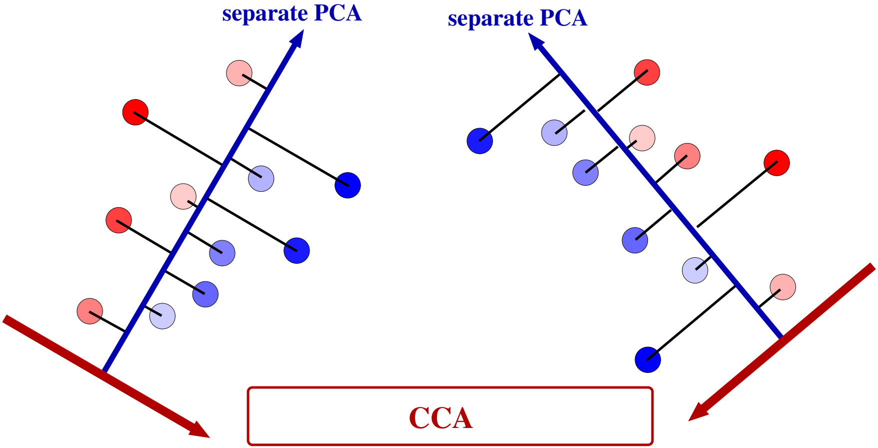
Often, each data point consists of **two views**:

- Image retrieval: for each image, have the following:
  - X: Pixels (or other visual features) Y: Text around the image
- Time series:
  - X: Signal at time t
  - Y: Signal at time $t + 1$
- Two-view learning: divide features into two sets
  - X: Features of a word/object, etc.
  - Y: Features of the context in which it appears

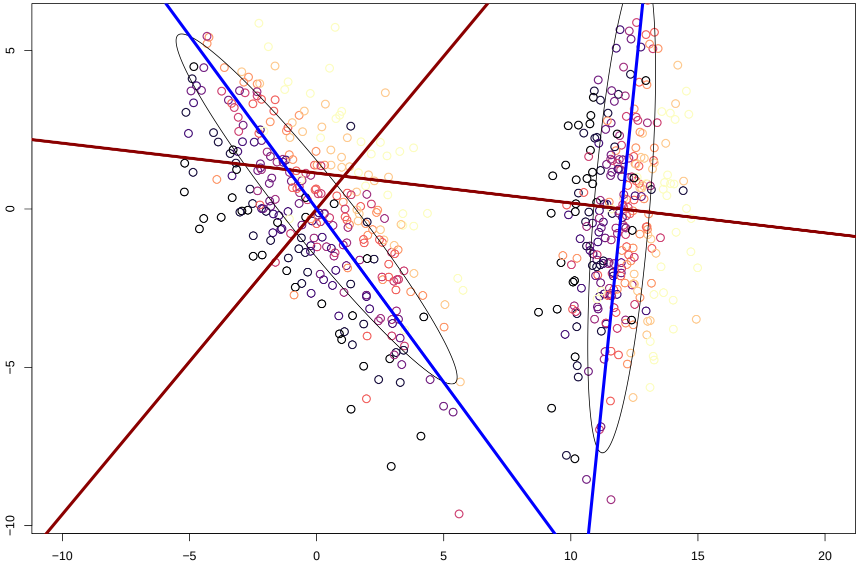
Goal: reduce the dimensionality of the two views **jointly**.

Find projections such that projected views are **maximally correlated.**

CCA vs PCA



CCA vs PCA



CCA: Setting

- Let $\boldsymbol{X}$ be a random vector $\in \mathbb{R}^{p_x}$ and $\boldsymbol{Y}$ be a random vector $\in \mathbb{R}^{p_y}$ Consider the combined $(p := p_x + p_y)$-dimensional random vector $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})^t$. Let its $(p \times p)$ covariance matrix be partitioned into blocks according to:

$$\boldsymbol{Z} = \begin{bmatrix} \Sigma_{XX} \in \mathbb{R}^{p_x \times p_x} & | & \Sigma_{XY} \in \mathbb{R}^{p_x \times p_y} \\ \Sigma_{YX} \in \mathbb{R}^{p_y \times p_x} & | & \Sigma_{YY} \in \mathbb{R}^{p_y \times p_y} \end{bmatrix}$$

- Assuming centered data, the blocks in the covariance matrix can be estimated from observed data sets $X \in \mathbb{R}^{n \times p_x}$, $Y \in \mathbb{R}^{n \times p_y}$:

$$\mathrm{Z} \approx \frac{1}{n} \begin{bmatrix} X^t X & | & X^t Y \\ Y^t X & | & Y^t Y \end{bmatrix}$$

CCA: Setting

- Correlation$(x, y) = \frac{\text{covariance}(x,y)}{\text{standard deviation}(x)\cdot \text{ standard deviation}(y)}$

$$\rho = cor(x, y) = \frac{cov(x, y)}{\sigma(x)\sigma(y)}.$$

- Sample correlation:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})^t}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \overset{\text{centered observations}}{=} \frac{\boldsymbol{x}^t \boldsymbol{y}}{\sqrt{\boldsymbol{x}^t \boldsymbol{x}}\sqrt{\boldsymbol{y}^t \boldsymbol{y}}}.$$

- Want to find maximally correlated 1D-projections $\boldsymbol{x}^t \boldsymbol{a}$ and $\boldsymbol{y}^t \boldsymbol{b}$.
- Projected covariance: $cov(\boldsymbol{x}^t \boldsymbol{a}, \boldsymbol{y}^t \boldsymbol{b}) \overset{\text{zero means}}{=} \boldsymbol{a}^t \Sigma_{XY} \boldsymbol{b}$.
- Define $\boldsymbol{c} = \Sigma_{XX}^{\frac{1}{2}} \boldsymbol{a}$, $\boldsymbol{d} = \Sigma_{YY}^{\frac{1}{2}} \boldsymbol{b}$.
- Thus, the projected correlation coefficient is: $\rho = \frac{\boldsymbol{c}^t \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}} \boldsymbol{d}}{\sqrt{\boldsymbol{c}^t \boldsymbol{c}}\sqrt{\boldsymbol{d}^t \boldsymbol{d}}}$.

CCA: Setting

- By the Cauchy-Schwarz inequality ($\boldsymbol{x}^t\boldsymbol{y} \le \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$), we have

$$\Big(\boldsymbol{c}^t \underbrace{\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-\frac{1}{2}}}_{H}\Big)\boldsymbol{d} \le \Big(\boldsymbol{c}^t \underbrace{\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}}_{G:=HH^t}\boldsymbol{c}\Big)^{\frac{1}{2}} \left(\boldsymbol{d}^t\boldsymbol{d}\right)^{\frac{1}{2}},$$

$$\rho \le \frac{(\boldsymbol{c}^t G\boldsymbol{c})^{\frac{1}{2}}}{(\boldsymbol{c}^t\boldsymbol{c})^{\frac{1}{2}}},$$

$$\rho^2 \le \frac{\boldsymbol{c}^t G\boldsymbol{c}}{\boldsymbol{c}^t\boldsymbol{c}}.$$

- Equality: vectors $\boldsymbol{d}$ and $\Sigma_{YY}^{-\frac{1}{2}}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}\boldsymbol{c}$ are collinear.
- Maximum: $\boldsymbol{c}$ is the eigenvector with the maximum eigenvalue of $G := \Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}$.
  Subsequent pairs $\rightsquigarrow$ using eigenvalues of decreasing magnitudes.
- Collinearity: $\boldsymbol{d} \propto \Sigma_{YY}^{-\frac{1}{2}}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}\boldsymbol{c}$
- Transform back to original variables $\boldsymbol{a} = \Sigma_{XX}^{-\frac{1}{2}}\boldsymbol{c}$, $\boldsymbol{b} = \Sigma_{YY}^{-\frac{1}{2}}\boldsymbol{d}$.

Pixels That Sound [Kidron, Schechner, Elad, 2005]

"People and animals fuse auditory and visual information to obtain robust perception. A particular benefit of such cross-modal analysis is the ability to localize visual events associated with sound sources. We aim to achieve this using computer-vision aided by a single microphone".

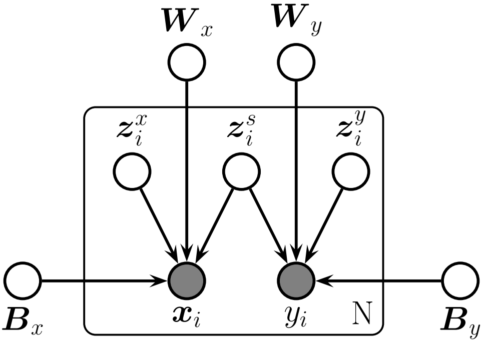
https://webee.technion.ac.il/ yoav/research/pixels-that-sound.html

Probabilistic CCA

(Bach and Jordan 2005): With Gaussian priors

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}^s|\mathbf{0}, I)\mathcal{N}(\boldsymbol{z}^x|\mathbf{0}, I)\mathcal{N}(\boldsymbol{z}^y|\mathbf{0}, I),$$

the MLE in the two-view FA model is equivalent to classical CCA (up to rotation and scaling).



From figure 12.19 in K. Murphy

Further connections

- If y is a **discrete class label** $\rightsquigarrow$ CCA is (essentially) equivalent to Linear Disriminant Analysis (LDA), see (Hastie et al. 1994).
- Arbitrary $\boldsymbol{y}$ $\rightsquigarrow$ CCA is (essentially) equivalent to the **Gaussian Information Bottleneck** (Chechik et al. 2005)
  - Basic idea: **compress $\boldsymbol{x}$** into compact latent representation while **preserving information about $\boldsymbol{y}$.**
  - Information theoretic motivation:
    Find encoding distribution $p(\boldsymbol{z}|\boldsymbol{x})$ by minimizing
    $$I(\boldsymbol{x};\boldsymbol{z}) - \beta I(\boldsymbol{x};\boldsymbol{y})$$
    where $\beta \geq 0$ is some parameter controlling the tradeoff between compression and predictive accuracy.
- Arbitrary $\boldsymbol{y}$, **discrete shared latent $\boldsymbol{z}^s$**
  $\rightsquigarrow$ **dependency-seeking clustering** (Klami and Kaski 2008): find clusters that "explain" the dependency between the two views.