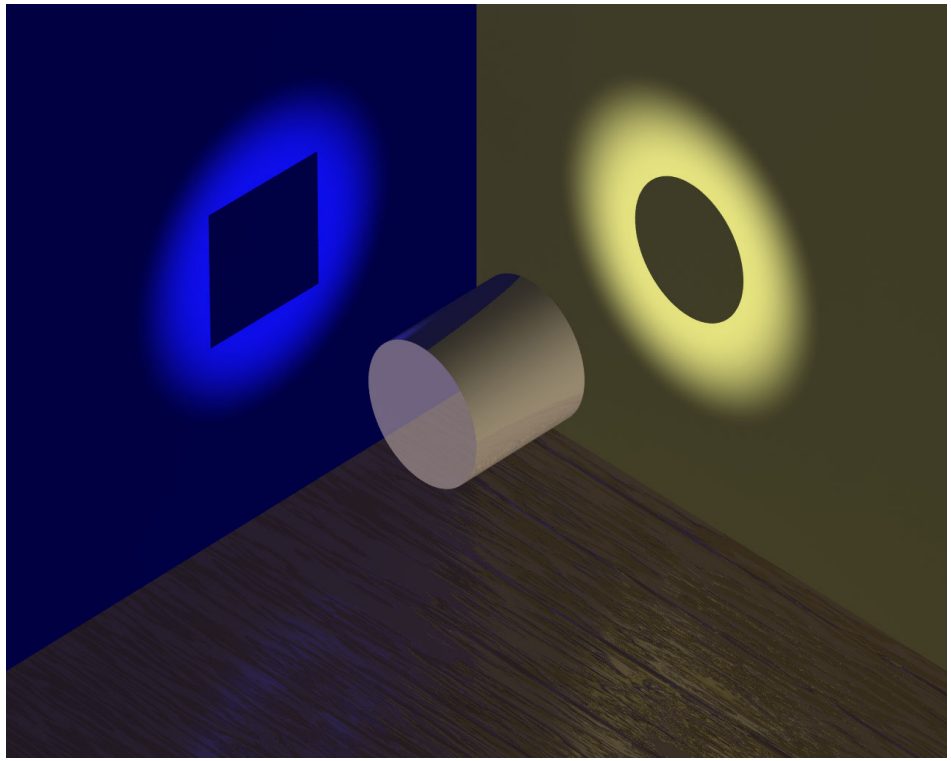
Machine Learning

Volker Roth

Department of Mathematics & Computer Science
University of Basel

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.



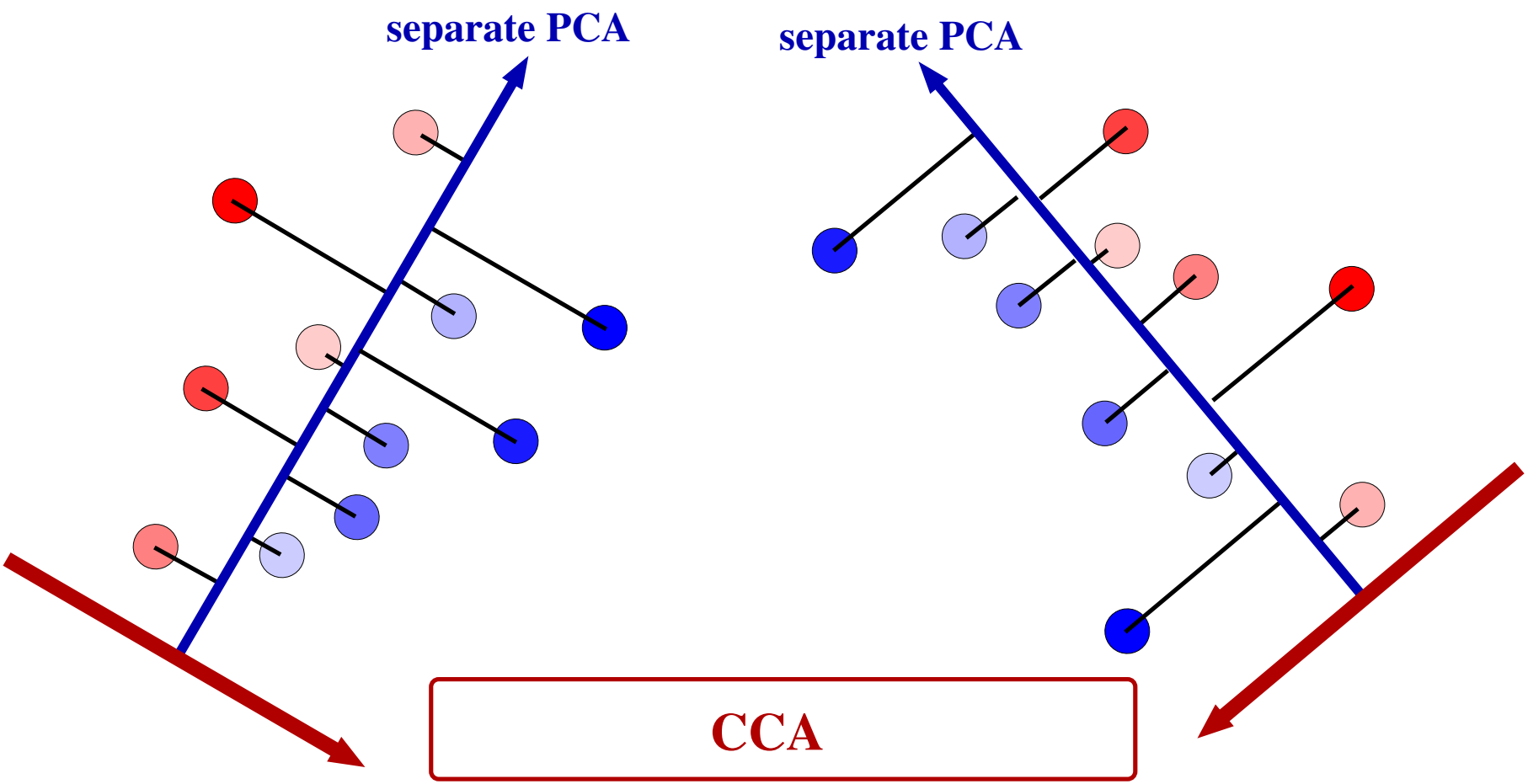
Canonical Correlation Analysis [Hotelling, 1936]

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

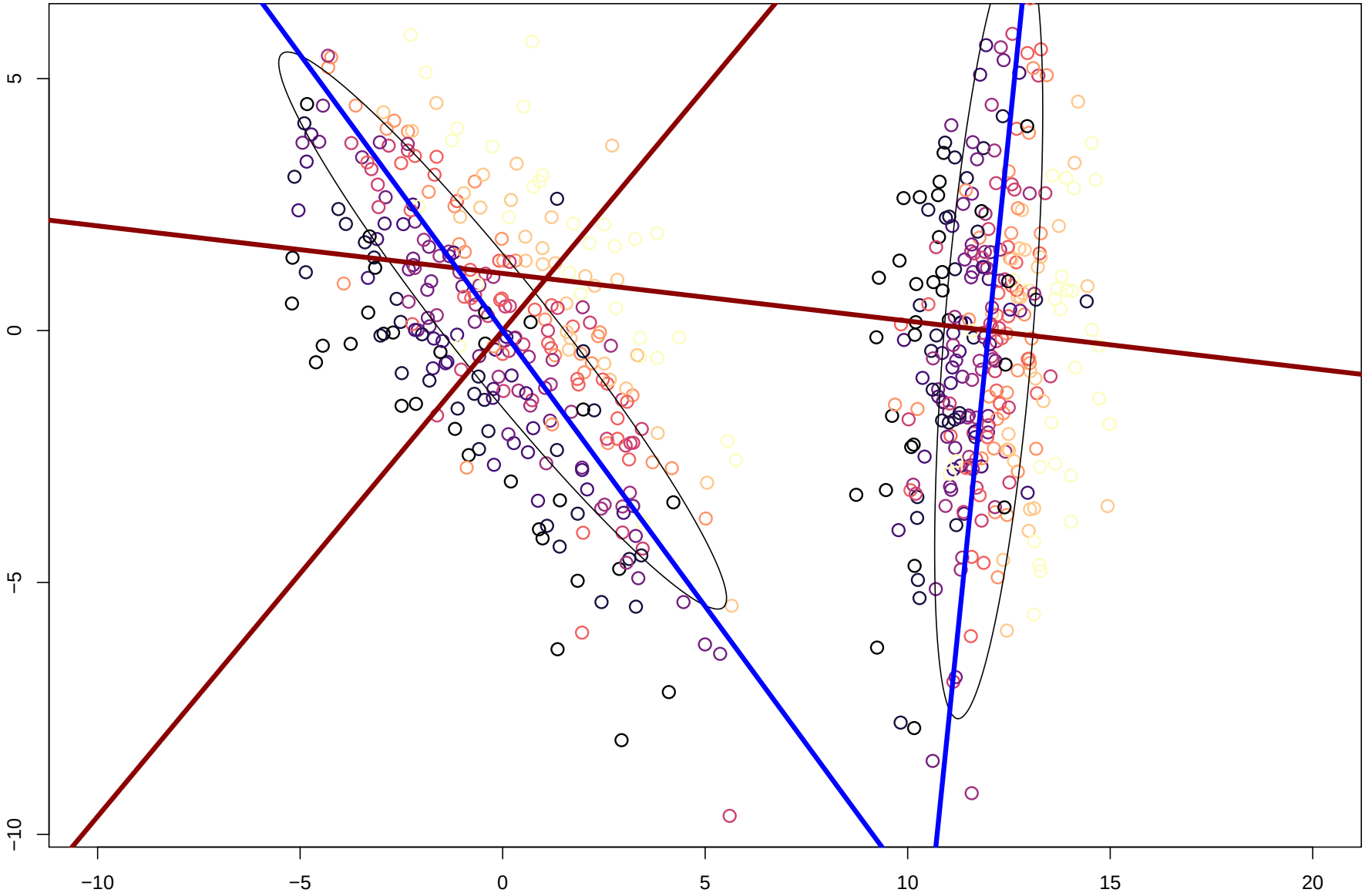
- Image retrieval: for each image, have the following:
  - X: Pixels (or other visual features) Y: Text around the image
- Time series:
  - X: Signal at time t
  - Y: Signal at time $t+1$
- Two-view learning: divide features into two sets
  - X: Features of a word/object, etc.
  - Y: Features of the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

$$\boldsymbol{Z} = \left[\begin{array}{c|c} \Sigma_{XX} \in \mathbb{R}^{p_x \times p_x} & \Sigma_{XY} \in \mathbb{R}^{p_x \times p_y} \\ \Sigma_{YX} \in \mathbb{R}^{p_y \times p_x} & \Sigma_{YY} \in \mathbb{R}^{p_y \times p_y} \end{array}\right]$$

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

$$\mathrm{Z} \approx \frac{1}{n} \left[\begin{array}{c|c} X^t X & X^t Y \\ Y^t X & Y^t Y \end{array}\right]$$

CCA: Setting

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

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

- 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 ($\boldsymbol{x}^t\boldsymbol{y} \leq \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$), we have

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

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

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

- 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 $\boldsymbol{a} = \Sigma_{XX}^{-\frac{1}{2}}\boldsymbol{c}$, $\boldsymbol{b} = \Sigma_{YY}^{-\frac{1}{2}}\boldsymbol{d}$.

Efficient computation of separate PCAs and joint CCA

- Separate SVD of X and Y: $X = U_x S_x V_x^t$, $Y = U_y S_y V_y^t$
  $\rightsquigarrow$ separate PCAs for both views.
- Consider sample covariance matrix

$$\Sigma_{XX} = 1/n \cdot X^t X = 1/n \cdot V_x S_x U_x^t U_x S_x V_x^t = 1/n \cdot V_x S_x^2 V_x^t.$$

- Substitute in matrix G:

$$\begin{aligned} G &= \Sigma_{XX}^{-1/2} \cdot \Sigma_{XY} \cdot \Sigma_{YY}^{-1} \cdot \Sigma_{YX} \cdot \Sigma_{XX}^{-1/2} \\ &= (V_x S_x^{-1} V_x^t)(V_x S_x U_x^t U_y S_y V_y^t)(V_y S_y^{-2} V_y^t)(V_y S_y U_y^t U_x S_x V_x^t)(V_x S_x^{-1} V_x^t) \\ &= V_x U_x^t U_y U_y^t U_x V_x^t. \end{aligned}$$

Efficient computation of separate PCAs and joint CCA

- $\boldsymbol{c}$ is an eigenvector of G
  $\Rightarrow \tilde{\boldsymbol{c}} = V_x^t \boldsymbol{c}$ is an eigenvector of $U_x^t U_y U_y^t U_x$ with the same e.vals:

$$\begin{aligned} G\boldsymbol{c} &= \lambda \boldsymbol{c} \\ V_x U_x^t U_y U_y^t U_x V_x^t \boldsymbol{c} &= \lambda \boldsymbol{c}, \\ U_x^t U_y U_y^t U_x V_x^t \boldsymbol{c} &= \lambda V_x^t \boldsymbol{c}, \\ U_x^t U_y U_y^t U_x \tilde{\boldsymbol{c}} &= \lambda \tilde{\boldsymbol{c}}. \end{aligned}$$

- In matrix form (the columns of matrix C are the eigenvectors $\boldsymbol{c}_j$)

$$\begin{aligned} GC &= C\Lambda, \\ U_x^t U_y U_y^t U_x (V_x^t C) &= (V_x^t C)\Lambda, \\ U_x^t U_y U_y^t U_x \tilde{C} &= \tilde{C}\Lambda. \end{aligned}$$

- Columns of $\tilde{C}$ are the columns of $\tilde{U}$ in SVD of $U_x^t U_y = \tilde{U}\tilde{S}\tilde{V}^t$

Efficient computation of separate PCAs and joint CCA

- Now recover C from $\tilde{C}$: $C = V_x \tilde{C} = V_x \tilde{U}$
  $\Rightarrow A = \Sigma_{XX}^{-1/2} C = (\frac{1}{n} V_x S_x^{-1} V_x^t)(V_x \tilde{U}) = \frac{1}{n} V_x S_x^{-1} \tilde{U}$.
- $B = \frac{1}{n} V_y S_y^{-1} \tilde{V}$ analogous.
- Joint CCA: correlation is maximized between pairs of projections $X\boldsymbol{a}_i$ and $Y\boldsymbol{b}_i$, with $\boldsymbol{a}_i, \boldsymbol{b}_i$ are the i-th columns of A, B.
  The corresponding correlations are $\rho_i = \tilde{S}_{ii}$.
- In summary:
  - Compute individual SVDs of X and $Y \rightsquigarrow V_x, S_X, V_y, S_y$
  - Compute "joint" SVD of $U_x^t U_y = \tilde{U} \tilde{S} \tilde{V}^t$.

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".

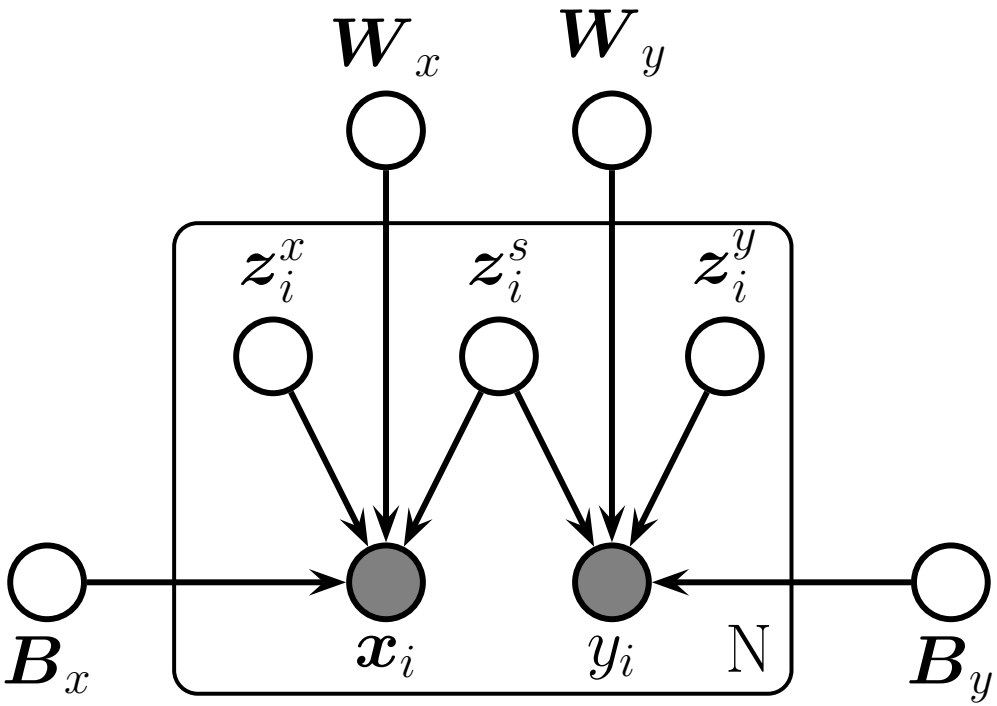
https://webee.technion.ac.il/ yoav/research/pixels-that-sound.html

Probabilistic CCA

(Bach and Jordan 2005): With Gaussian priors

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

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



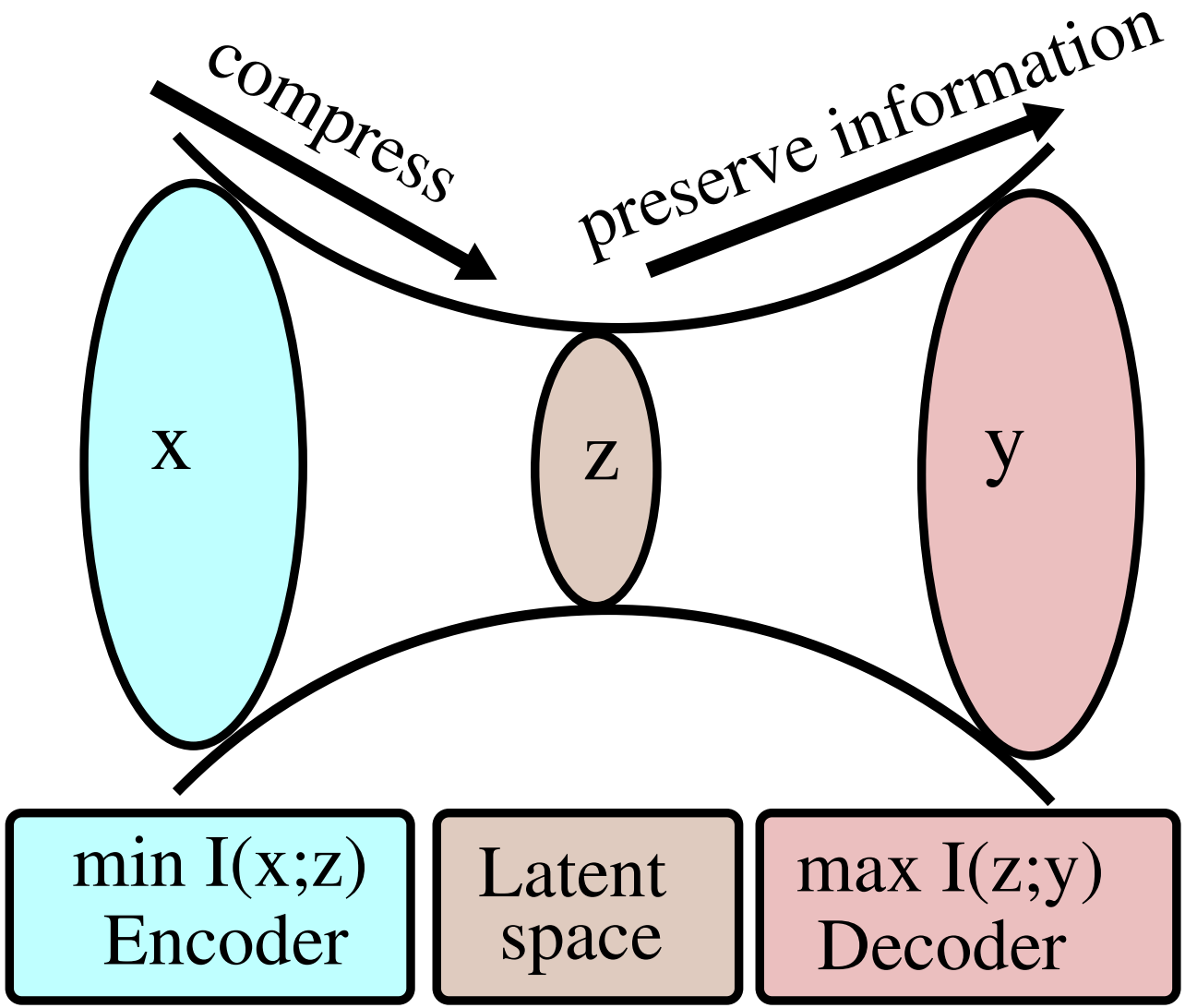
From figure 12.19 in K. Murphy

Further connections

- If y is a **discrete class label** $\rightsquigarrow$ CCA is (essentially) equivalent to Linear Discriminant 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
    $$I(\boldsymbol{x}; \boldsymbol{z}) - \beta I(\boldsymbol{z}; \boldsymbol{y})$$
    where $\beta \geq 0$ is some parameter controlling the trade-off 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.

The Information Bottleneck (Tishby et al., 1999)

FA is powerful, but still limited (Gaussian assumptions etc.). Alternatives?



Mutual Information

- A measure of **mutual dependence** between two random variables: reduction of uncertainty by knowing one variable.
- For continuous RVs:

$$\begin{aligned} I(\boldsymbol{x};\boldsymbol{y}) &= \int\int p(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})\,p(\boldsymbol{y})}\right) d\boldsymbol{x}\,d\boldsymbol{y} \\ &= D_{KL}(p(\boldsymbol{x},\boldsymbol{y})\|p(\boldsymbol{x})\,p(\boldsymbol{y})) \\ &= \int p(\boldsymbol{x}) \int p(\boldsymbol{y}|\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y})}\right) d\boldsymbol{x}\,d\boldsymbol{y} \\ &= E_{p(\boldsymbol{x})} D_{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|p(\boldsymbol{y})). \end{aligned}$$

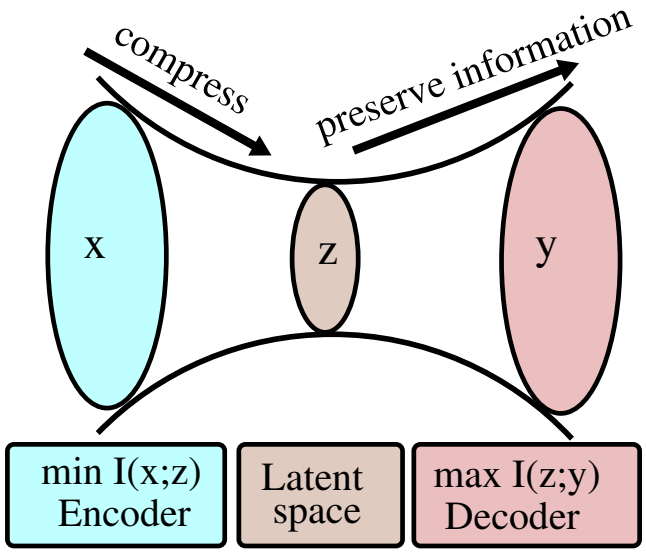
- $\boldsymbol{x}$ and $\boldsymbol{y}$ independent $\rightsquigarrow$ knowing $\boldsymbol{x}$ does not give any information about $\boldsymbol{y}$ $\rightsquigarrow$ $I(\boldsymbol{x};\boldsymbol{y}) = 0$.

Information Bottleneck

- **The IB principle:** compress $\boldsymbol{x}$ into $\boldsymbol{z}$, keep information about $\boldsymbol{y}$.
- Assume $\boldsymbol{y}$ and $\boldsymbol{z}$ are conditionally independent given $\boldsymbol{x}$ an solve:

$$\min_{p(\boldsymbol{z}|\boldsymbol{x})} I(\boldsymbol{x};\boldsymbol{z}) - \beta I(\boldsymbol{z};\boldsymbol{y}).$$

- The original IB formulation is **not a generative model**, nor do we actually condition: $\boldsymbol{x}, \boldsymbol{y}$ are only used for estimating $p(\boldsymbol{x}, \boldsymbol{y})$.



IB as a latent variable model

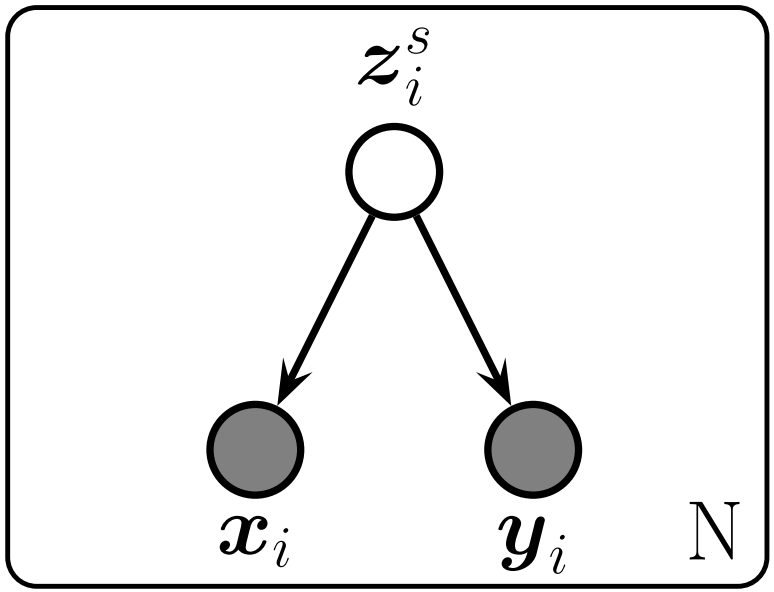
Assume $\boldsymbol{z} = f(\boldsymbol{x}) + \boldsymbol{\xi}$ captures all relevant information about $\boldsymbol{y}$

$\Rightarrow$ $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} | \boldsymbol{z}$

$\rightsquigarrow$ **latent version IB** $^{\text{(lat)}}$, basically an **asymmetric CCA model.**

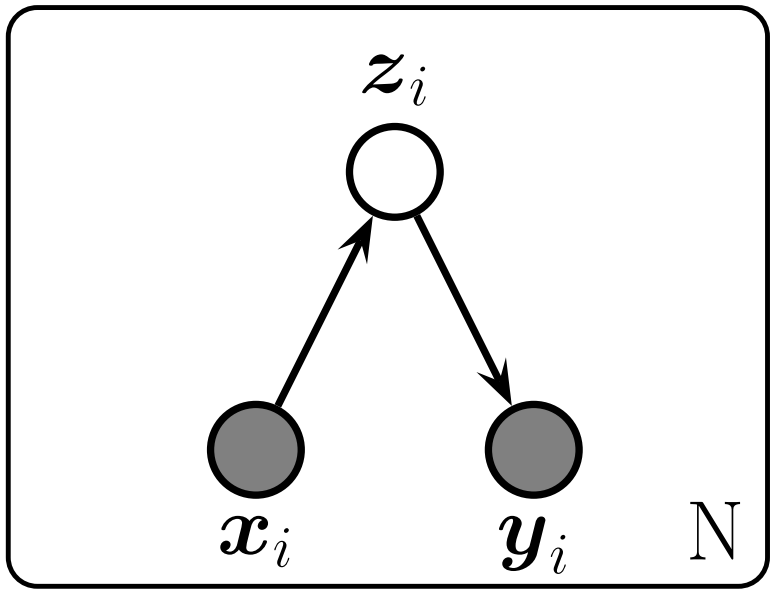
CCA: $p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z})$

$\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \,|\, \boldsymbol{z}$



IB$^{\text{(lat)}}$: $p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{x})$

$\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \,|\, \boldsymbol{z}$



Gaussian IB (Chechnik et al. 2003)

- Assume $\boldsymbol{x}$ and $\boldsymbol{y}$ are **jointly Gaussian-distributed**.

$$(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma_x & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_y \end{pmatrix}\right),$$

- The optimal $\boldsymbol{z}$ is a **noisy projection of $\boldsymbol{x}$**:

$$\boldsymbol{z} = A\boldsymbol{x} + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I) \Rightarrow \boldsymbol{z}|\boldsymbol{x} \sim \mathcal{N}(A\boldsymbol{x}, I),\ \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, A\Sigma_x A^t + I).$$

- **Analytic form of mutual information:**
  $I(\boldsymbol{x}; \boldsymbol{z}) = \frac{1}{2} \log |A\Sigma_x A^t + I|,$
  $I(\boldsymbol{z}; \boldsymbol{y}) = I(\boldsymbol{x}; \boldsymbol{z}) - \frac{1}{2} \log |A\Sigma_{x|y} A^t + I|.$
- general $\boldsymbol{y}$: rows of A are eigenvectors of $\Sigma_x^{-1}\Sigma_{x|y}$ $\rightsquigarrow$ **CCA**
- one-dimensional y: $\rightsquigarrow$ **least squares regression**
- $\boldsymbol{y}$ is noisy version of $\boldsymbol{x}$: $\rightsquigarrow$ **PCA**

Sparse Gaussian IB

- Recall: $\boldsymbol{z} = A\boldsymbol{x} + \boldsymbol{\xi}$, $\Rightarrow$ $\boldsymbol{z}|\boldsymbol{x} \sim \mathcal{N}(A\boldsymbol{x}, I)$, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, A\Sigma_x A^t + I)$.
- Note that $|A\Sigma A^t + I| = |\Sigma A^t A + I| \rightsquigarrow$ only $A^t A$ identifiable.
- **Sparse version:** assume that $A^t A$ is diagonal.
- **Intuition:** for RVs $\boldsymbol{x}, \boldsymbol{x}'$, any full-rank projection $A\boldsymbol{x}'$ of $\boldsymbol{x}'$ would lead to the same mutual information since $I(\boldsymbol{x}, \boldsymbol{x}') = I(\boldsymbol{x}; A\boldsymbol{x}')$, and a reduction can only be achieved by a rank-deficient matrix A.
- If we compress X, A cannot have full rank $\rightsquigarrow$ $A^t A$ diagonal implies that A can be non-zero only in some subspace, $AP = \begin{pmatrix} A' & 0 \end{pmatrix}$.
- The mean $\boldsymbol{\mu} = A\boldsymbol{x}$ is a sparse compression of $\boldsymbol{x}$.
- Basically an information-theoretic lasso variant.

Sparse Gaussian IB: Experiments

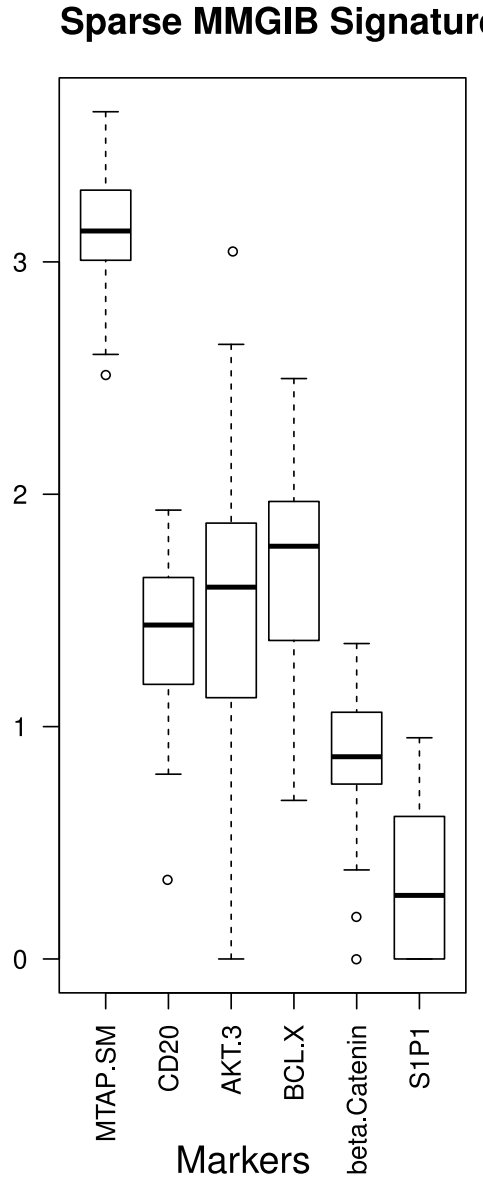
Prognosis of cutaneous **malignant melanoma** (MM) using biomarkers (Fuchs & Buhmann, 2011; Rey & R. 2014).

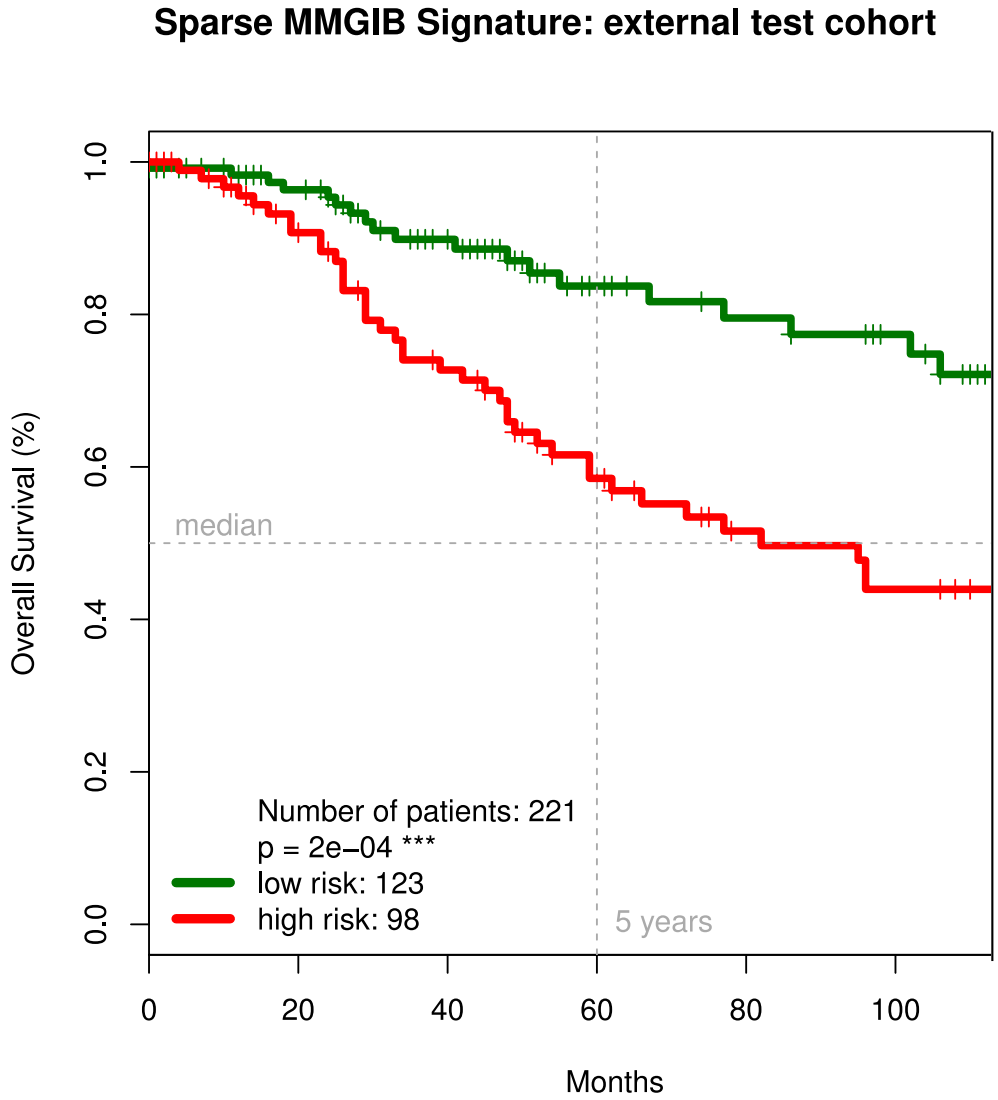
- **Goal: identify biomarkers** relevant to the disease evolution.
- Crucial for cost-effective prognosis and therapy optimization

Data:

- Variable to compress, $\boldsymbol{x}$: **immunohistochemical (IHC) expressions** of 70 candidate biomarkers measured for 364 patients.
- Relevance variable, $\boldsymbol{y}$: 9 **different clinical observations** about the stage of the tumor & survival times.

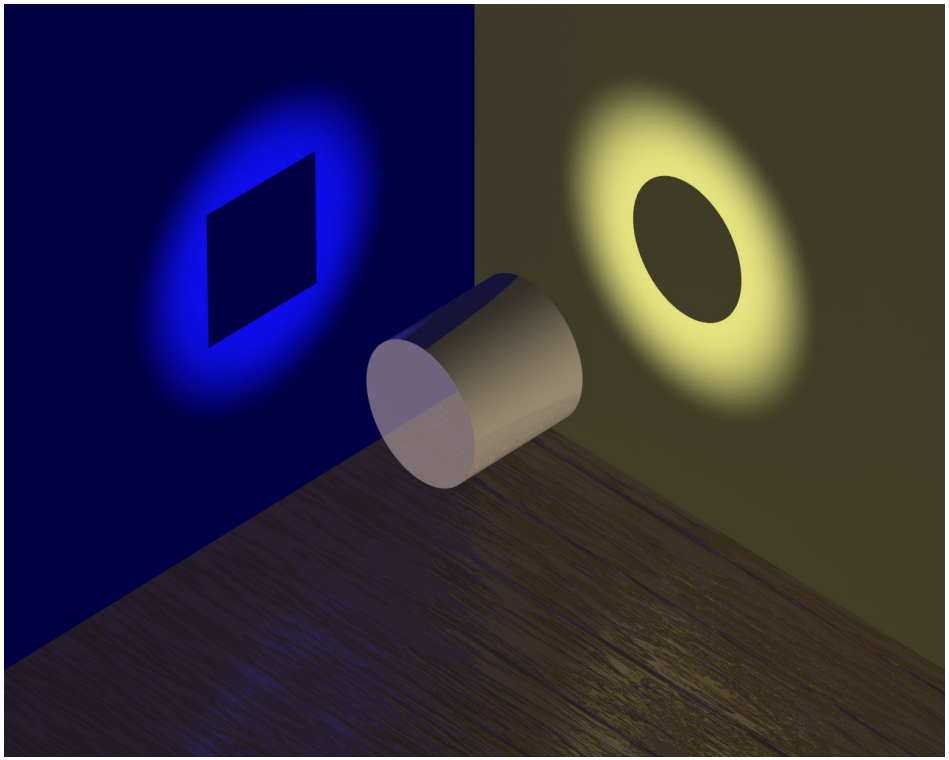
Sparse Gaussian IB: Experiments



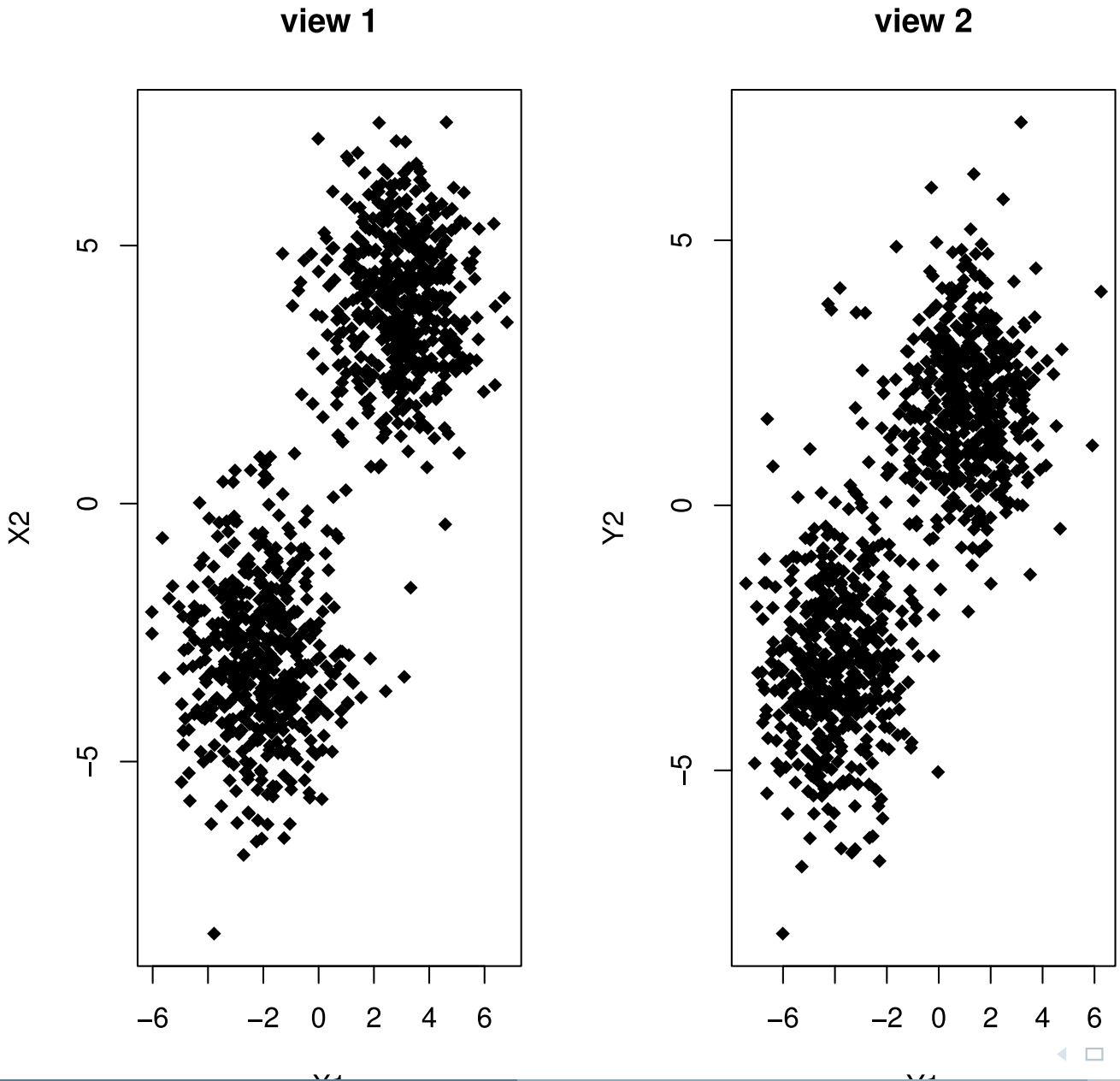


Multiple Views: Dependency Seeking Clustering

- Consider paired samples from different views.
- What is the dependency structure between the views ?
- Standard approach: global linear dependence detected by CCA.
- Generalization $\rightarrow$ dependency-seeking clustering.
- The cluster structure captures dependencies.



Dependency Seeking Clustering



Dependency Seeking Clustering