Probabilistic ICA (IFA –Independent Factor Analysis)

Use probabilities for everything unknown

$$\mathbf{y}(t) = \mathbf{AS}(t) + \mathbf{\Gamma}(t)$$

Probability of observations (given the sources) for Gaussian noise $\mathbf{\Gamma}$

$$p(\mathbf{y}|\mathbf{S}, \mathbf{A}, \mathbf{\Sigma}) = (2\pi \det \mathbf{\Sigma})^{-d/2} \, e^{-\frac{1}{2}(\mathbf{y}-\mathbf{AS})^T \mathbf{\Sigma}^{-1} (\mathbf{y}-\mathbf{AS})} \; .$$

Prior density model of sources

$$p(\mathbf{S}) = \prod_{i=1}^{m} p_i(s_i)$$

Complete Data Likelihood for n datapoints $\{\mathbf{y}\}_{i=1}^{n}$

$$p(\{\mathbf{y}\}_{i=1}^{n}|\mathbf{A}, \mathbf{\Sigma}) = \prod_{i=1}^{n} \int d\mathbf{S} \; p(\mathbf{y}_i|\mathbf{S}, \mathbf{A}, \mathbf{\Sigma}) \; p(\mathbf{S})$$

The Expectation–Maximisation (EM) Algorithm

1. Start with arbitrary $\boldsymbol{\theta}_0$

   Iterate:

2. (E-Step): Compute the expectation

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) \equiv \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_t) \ln p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$$

   with the **posterior probability** (given the observations) of the latent variables

$$p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_t) = \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_t) p(\mathbf{x}|\boldsymbol{\theta}_t)}{p(\mathbf{y}|\boldsymbol{\theta}_t)}$$

3. (M–Step) Maximise

$$\boldsymbol{\theta}_{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$$

Claim: $\ln p(\mathbf{y}|\boldsymbol{\theta}_{t+1}) \geq \ln p(\mathbf{y}|\boldsymbol{\theta}_t)$ Likelihood is not decreasing!

Analysis of EM

The proof requires the *Kullback–Leibler divergence* which fulfils

$$KL(q,p) = \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})} \geq 0 \ .$$

for any $q(\mathbf{x})$. By rearranging we get

$$-\ln p(\mathbf{y}|\boldsymbol{\theta}) \leq F(q,\theta) \equiv \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{y},\mathbf{x}|\boldsymbol{\theta})}$$

For fixed $\boldsymbol{\theta}$, the right is minimal (equality!!!) if $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$.

Let $q_t(\mathbf{x}) \doteq p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}_t)$, then $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\theta}_t) = -F(q_t,\theta) + \sum_{\mathbf{x}} q_t(\mathbf{x}) \ln q_t(\mathbf{x})$

Hence, the EM algorithm can be reformulated as:

1. E–Step: Minimise $F(q,\boldsymbol{\theta}_t)$ w.r.t $q \rightarrow q_t(\mathbf{x})$ and compute $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\theta}_t)$.

2. M–Step Minimise $F(q_t,\boldsymbol{\theta}) = -\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\theta}_t) + \sum_{\mathbf{x}} q_t(\mathbf{x}) \ln q_t(\mathbf{x})$ w.r.t. $\boldsymbol{\theta}$.

We get

$$-\ln p(\mathbf{y}|\boldsymbol{\theta}) \leq F(q_t, \theta)$$

and

$$-\ln p(\mathbf{y}|\boldsymbol{\theta}_t) = F(q_t, \theta_t)$$

Hence,

$$\ln p(\mathbf{y}|\boldsymbol{\theta}_{t+1}) - \ln p(\mathbf{y}|\boldsymbol{\theta}_t) \leq -F(q_t, \theta_{t+1}) + F(q_t, \theta_t) \geq 0$$

Likelihood is not decreasing!

Example: Mixture of Gaussians

- (E-Step): Compute

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) \equiv \sum_{\mathbf{c}} p(\mathbf{c}|\mathbf{y}, \boldsymbol{\theta}_t) \ln \left\{ \prod_i p(y_i, c_i|\boldsymbol{\theta}) \right\}$$

  with

$$p(\mathbf{c}|\mathbf{y}, \boldsymbol{\theta}_t) = \prod_i p(c_i|y_i, \boldsymbol{\theta}_t) = \prod_i \frac{p(y_i|c_i, \boldsymbol{\theta}_t) p(c_i|\boldsymbol{\theta}_t)}{p(y_i|\boldsymbol{\theta}_t)}$$

  and

$$p(y_i, c_i, \boldsymbol{\theta}) = p(y_i|c_i, \boldsymbol{\theta}) p(c_i|\boldsymbol{\theta})$$

- (M–Step) Update $\boldsymbol{\theta}_{t+1} = \text{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$

Explicit Formulas

- Variation with respect to μ_c

$$\sum_i (y_i - \mu_c) p(c|y_i, \boldsymbol{\theta}_t) = 0 \;\rightarrow \mu_{c,t+1} = \frac{\sum_i y_i \, p(c|y_i, \boldsymbol{\theta}_t)}{\sum_i p(c|y_i, \boldsymbol{\theta}_t)}$$

- Variation with respect to σ_c^2

$$\sigma_{c,t+1}^2 = \frac{\sum_i (y_i - \mu_{c,t+1})^2 \, p(c|y_i, \boldsymbol{\theta}_t)}{\sum_i p(c|y_i, \boldsymbol{\theta}_t)}$$

- Variation with respect to $p_{t+1}(c) = p(c|\boldsymbol{\theta}_{t+1})$

$$p_{t+1}(c) \equiv p(c|\boldsymbol{\theta}_{t+1}) = \frac{1}{n} \sum_i p(c|y_i, \boldsymbol{\theta}_t)$$

Low dimensional representations

Observations $\mathbf{y} \in R^d$ live effectively on lower dimensional manifold (+ noise). Introduce latent variables $\mathbf{x} \in R^q \sim \mathcal{N}(0, \mathbf{I})$

<u>Factor analysis</u>:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \mathbf{u}$$

with $\mathbf{W} = d \times q$, $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$ $\mathbf{u} \sim \mathcal{N}(0, \mathbf{D})$ and $\mathbf{D}$ diagonal.

<u>Probabilistic PCA</u> (Tipping & Bishop)

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \mathbf{u}$$

with $\mathbf{u} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.

We have

$$p(\mathbf{x}) \propto \exp\left[-\frac{1}{2}||\mathbf{x}||^2\right]$$

and

$$p(\mathbf{y}|\mathbf{x}) \propto \exp\left[-\frac{1}{2\sigma^2}||\mathbf{y} - (\mathbf{W}\mathbf{x} + \boldsymbol{\mu})||^2\right]$$

Posterior of latent variables

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma^2}||\mathbf{y} - (\mathbf{W}\mathbf{x} + \boldsymbol{\mu})||^2 - \frac{1}{2}||\mathbf{x}||^2\right] \propto$$

$$\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{x} - \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{y} - \boldsymbol{\mu})\right)^T \mathbf{M}\left(\mathbf{x} - \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{y} - \boldsymbol{\mu})\right)\right]$$

with

$$\mathbf{M} = (\sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W})$$

Full probability of data

$$p(\mathbf{y}) \propto \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right]$$

with

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Maximum Likelihood:

Minimise

$$-\ln p(\mathbf{Y}) = -\sum_{i=1}^{n} \ln p(\mathbf{y}_i) = \text{const} + \frac{n}{2}\Big(\ln|\mathbf{C}| + \text{trace}(\mathbf{C}^{-1}\mathbf{S})\Big)$$

with respect to $\mathbf{W}$, $\boldsymbol{\mu}$ and σ^2.

$\mathbf{S}$ is the empirical **data covariance**

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T$$

One can show (not surprisingly) that

$$\boldsymbol{\mu}_{ML} = \bar{\boldsymbol{\mu}} = \frac{1}{N}\sum_i \mathbf{y}_i$$

One can show that optimality is achieved for

$$\mathbf{W} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$$

$\mathbf{U}$ contains the q PCs with eigenvalues in the diagonal $\mathbf{\Lambda}$ of the data covariance $\mathbf{\Sigma}$. $\mathbf{R}$ is an arbitrary orthogonal $(q \times q)$ matrix.

Advantages over conventional PCA

- One can use EM algorithm (in certain cases computationally more efficient)

- can treat missing values

- PPCA can be extended to mixtures of PPCA using

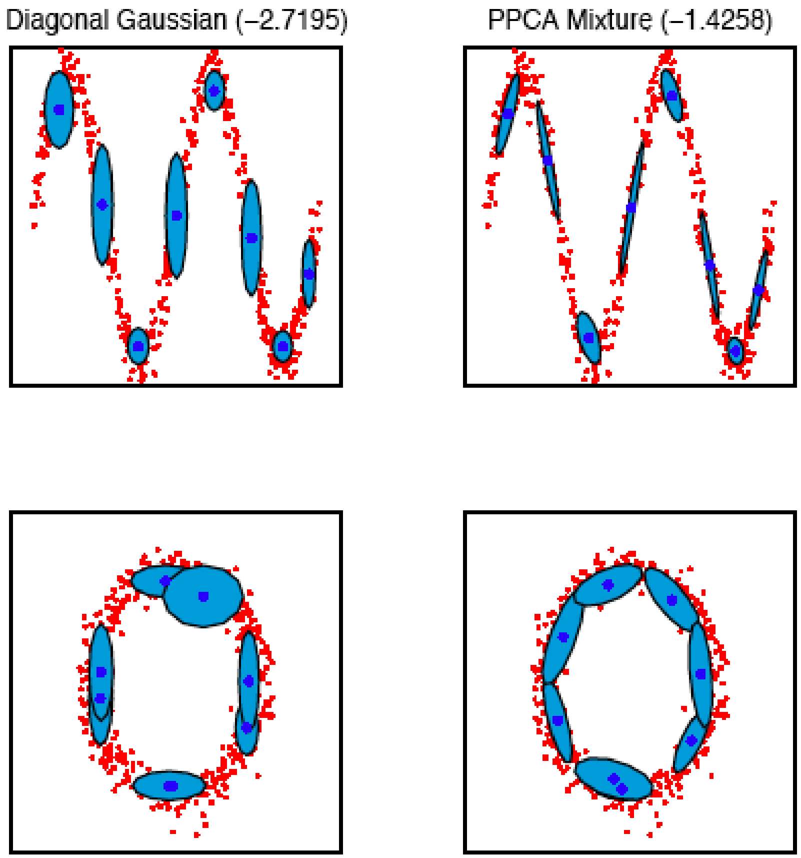
$$p(\mathbf{y}) = \sum_k p_k p(\mathbf{y}|k)$$

  where $p(\mathbf{y}|k)$ is given by PPCA.

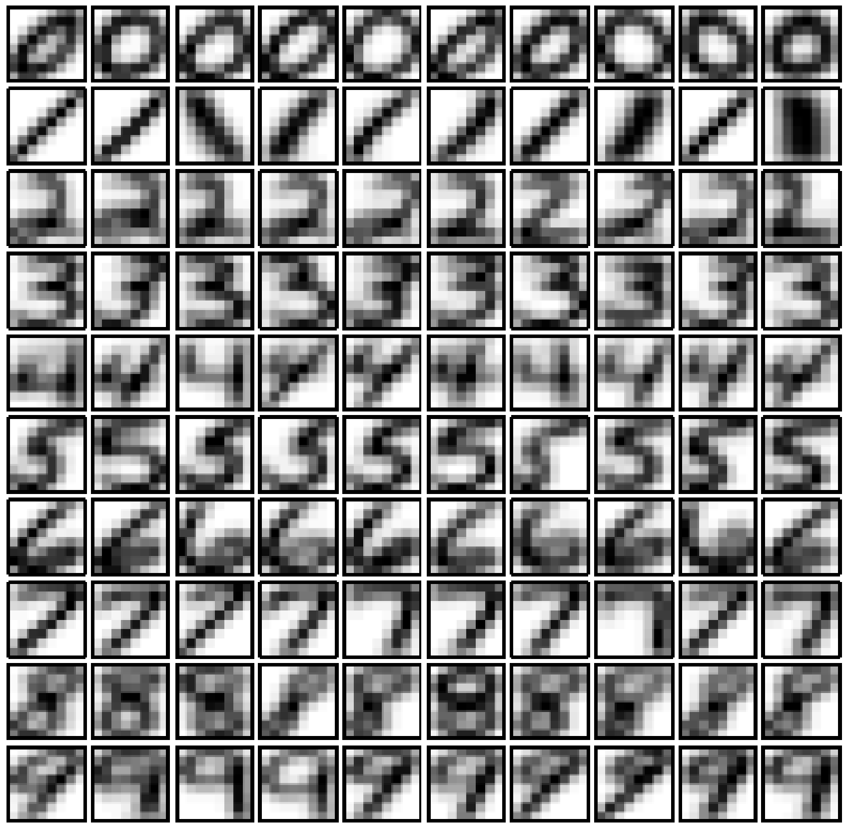
- Can be extended to Bayesian treatment (optimal model order)

- PPCA is a can be used for modelling class conditional densities (classification)

- Likelihood can be used for comparison with other density models



8: Comparison of an 8-component diagonal variance Gaussian mixture model with a mixture of PPCA model. The upper two plots give a view perpendicular to the major axis of the spiral, while the lower two plots show the end elevation. The covariance structure of each mixture component is shown by projection of a unit Mahalanobis distance ellipse and the log-likelihood per data-point is given in brackets above the figures.



'igure 9: The mean vectors $\boldsymbol{\mu}_i$, illustrated as gray-scale digits, for each of the ten digit models. The model for a given digit is a mixture of ten PPCA models, one centred at each of the pixel vectors shown on the corresponding row. Note how different components can capture different styles of digit.

Image compression **Bishop & Tipping**

720×360 pixel image segmented into 8×8 non-overlapping blocks $\rightarrow$ dataset of 4050 64 dim vectors.

Single PCA $q = 4$ versus mixtures of PPCA (12 mixing components, $q = 4$), left half of image used for training. Compress by quantising transform variable and component label.

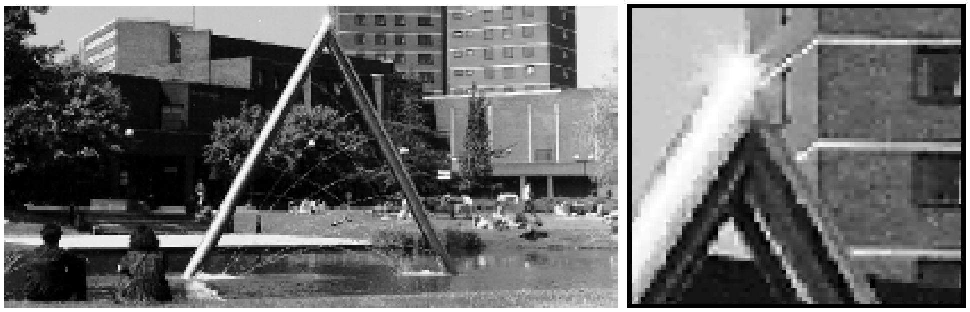


Figure 5: The original image (left), and detail therein (right).

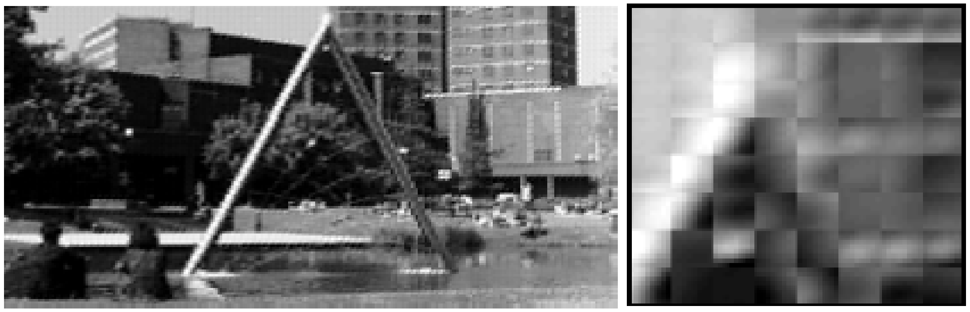


Figure 6: The PCA reconstructed image, at 0.5 bits-per-pixel.

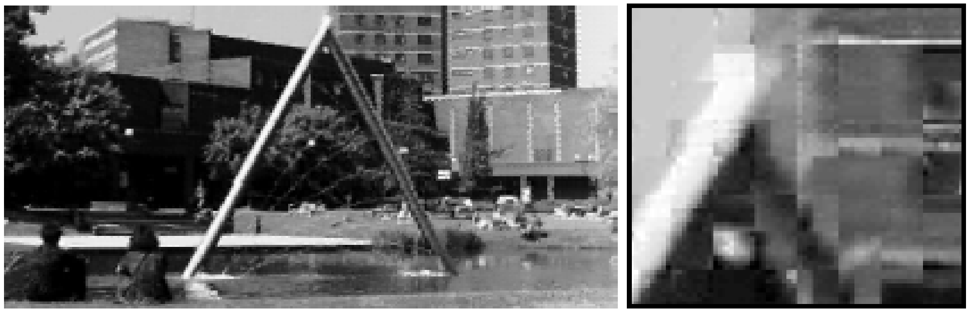


Figure 7: The mixture of PPCA reconstructed image, using the same bit-rate as Figure 6.

Latent variable models for data visualisation

Visualise high (d) dim. data $\mathbf{y}$ in low (~ 2) dim data space $\mathcal{H}$ using latent variables $\mathbf{u}$.

Generative Topographic Mapping (GTM) (Bishop, Svenson & Williams)

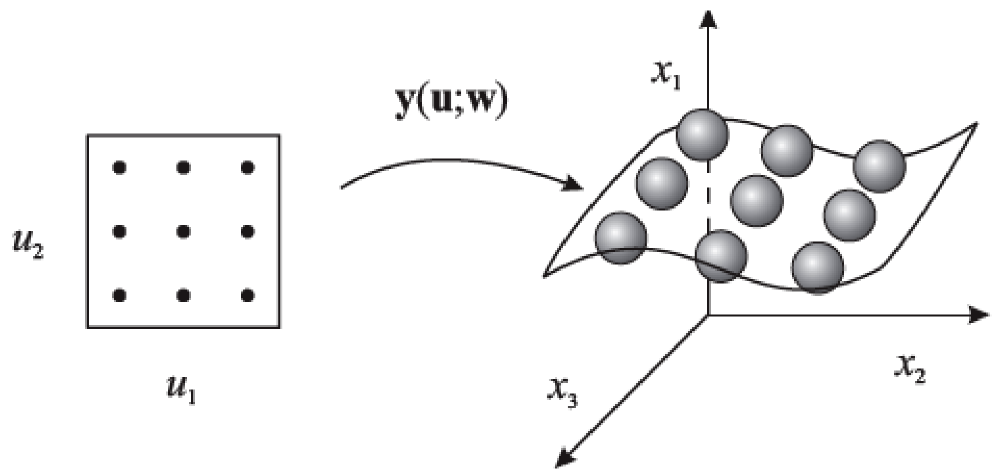
$$p(\mathbf{y}|\mathbf{u}, \mathbf{W}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{\frac{-||\mathbf{f}(\mathbf{u}, \mathbf{W}) - \mathbf{y}||^2}{2\sigma^2}\right\}$$

Latent variables $\mathbf{x}$ are assumed to be on a **discrete grid** with $p(\mathbf{u}) = \frac{1}{K}\sum_k \delta(\mathbf{u} - \mathbf{u}_k)$. $\mathbf{f}$ is a smooth mapping, e.g. $\mathbf{f}(\mathbf{u}, \mathbf{W}) = \mathbf{W}\boldsymbol{\phi}(\mathbf{u})$ with fixed nonlinear (e.g. radial) basis functions $\boldsymbol{\phi}$ and a $d \times M$ matrix $\mathbf{W}$ to be optimised.

The total probability is

$$p(\mathbf{y}|\mathbf{W}) = \frac{1}{K}\sum_k p(\mathbf{y}|\mathbf{u}_k, \mathbf{W})$$

Projection of data points: Use mean or mode of posterior $p(\mathbf{u}|\mathbf{y})$.


u_2
u_1
$\mathbf{y}(\mathbf{u};\mathbf{w})$
x_1
x_2
x_3

Latent trait models

(Kabán, Girolami)

Replace Gaussians by more general exponential families. Helps e.g. to visualise discrete data.

$$p(\mathbf{y}|\mathbf{u}, \mathbf{W}) = p_0(\mathbf{y}) \exp \{\mathbf{y} \cdot \mathbf{f}_{\mathbf{W}}(\mathbf{u}) - g(\mathbf{f}_{\mathbf{W}}(\mathbf{u}))\}$$

with a nonlinear mapping $\mathbf{f}_{\mathbf{W}}$ from latent space to data.

<u>Example 1: Bernoulli distribution for binary data</u>

Let $\mathbf{y} = (y_1, \ldots, y_d) \in \{0, 1\}^d$. Then we define $m_k = \text{sigmo}((\mathbf{W}\boldsymbol{\phi}(\mathbf{u}))_k)$ with $\text{sigmo}(z) = \frac{e^z}{1+e^z}$ Finally:

$$p(\mathbf{y}|\mathbf{u}, \mathbf{W}) = \prod_{k=1}^{d} m_k^{y_k} (1 - m_k)^{1-y_k}$$

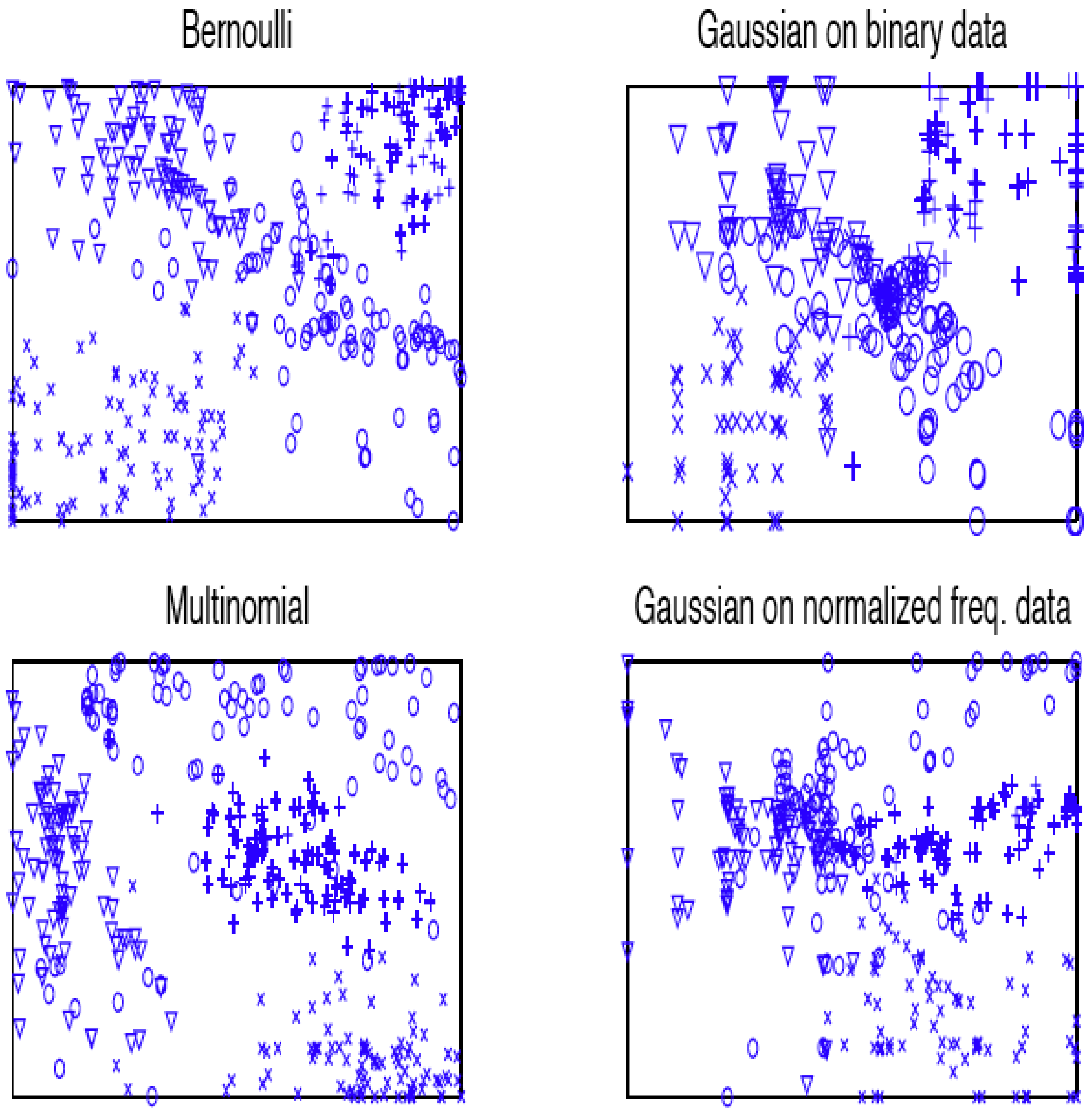
Example II: Multinomial Model

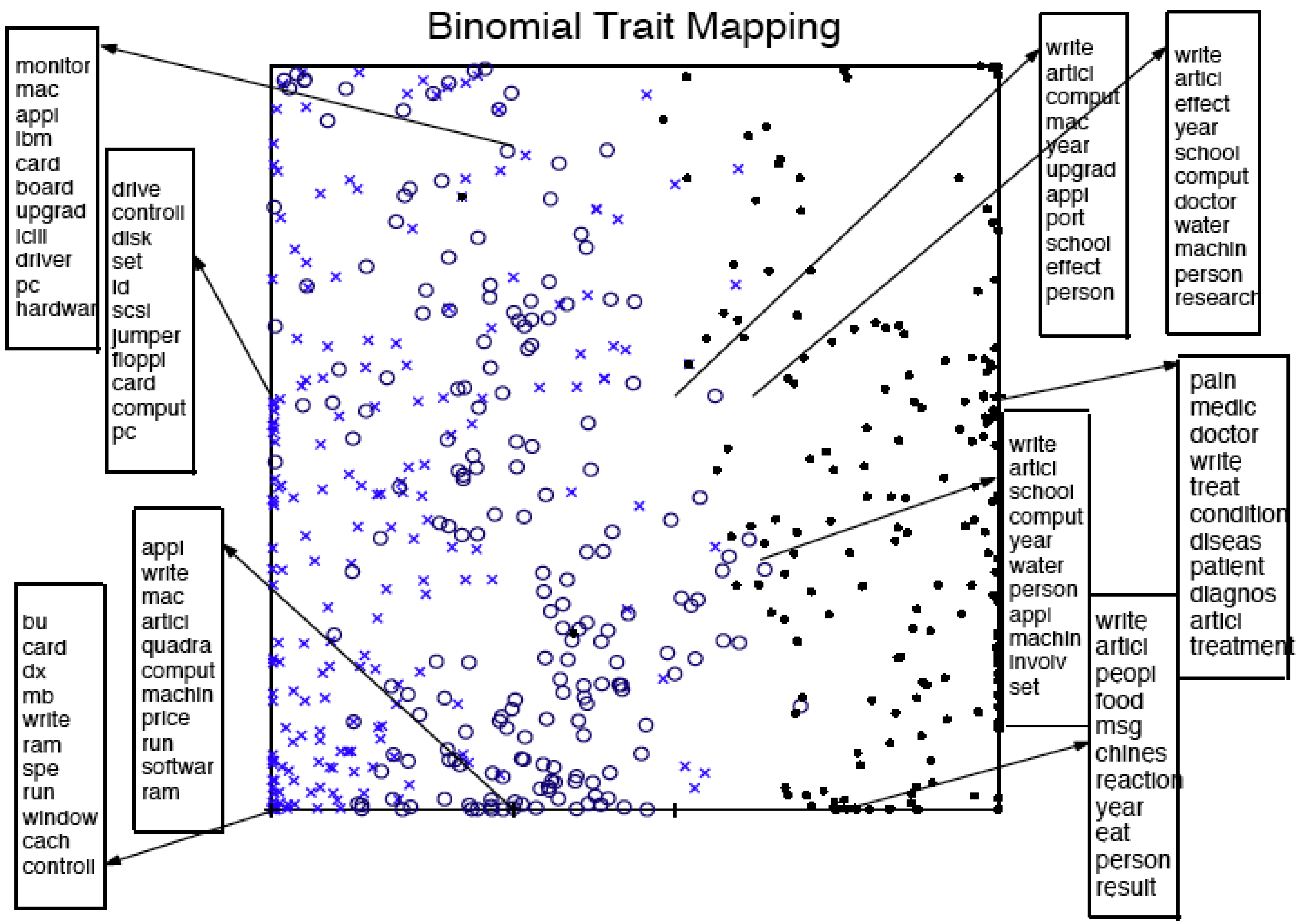
Here, $y_k \in N_0$. We set $m_k = \frac{\exp[(\mathbf{W}\boldsymbol{\phi}(\mathbf{u}))_k)]}{\sum_{k'=1^d} \exp[(\mathbf{W}\boldsymbol{\phi}(\mathbf{u}))_{k'})]}$

$$p(\mathbf{y}|\mathbf{u}, \mathbf{W}) = \prod_{k=1}^{d} m_k^{y_k}$$

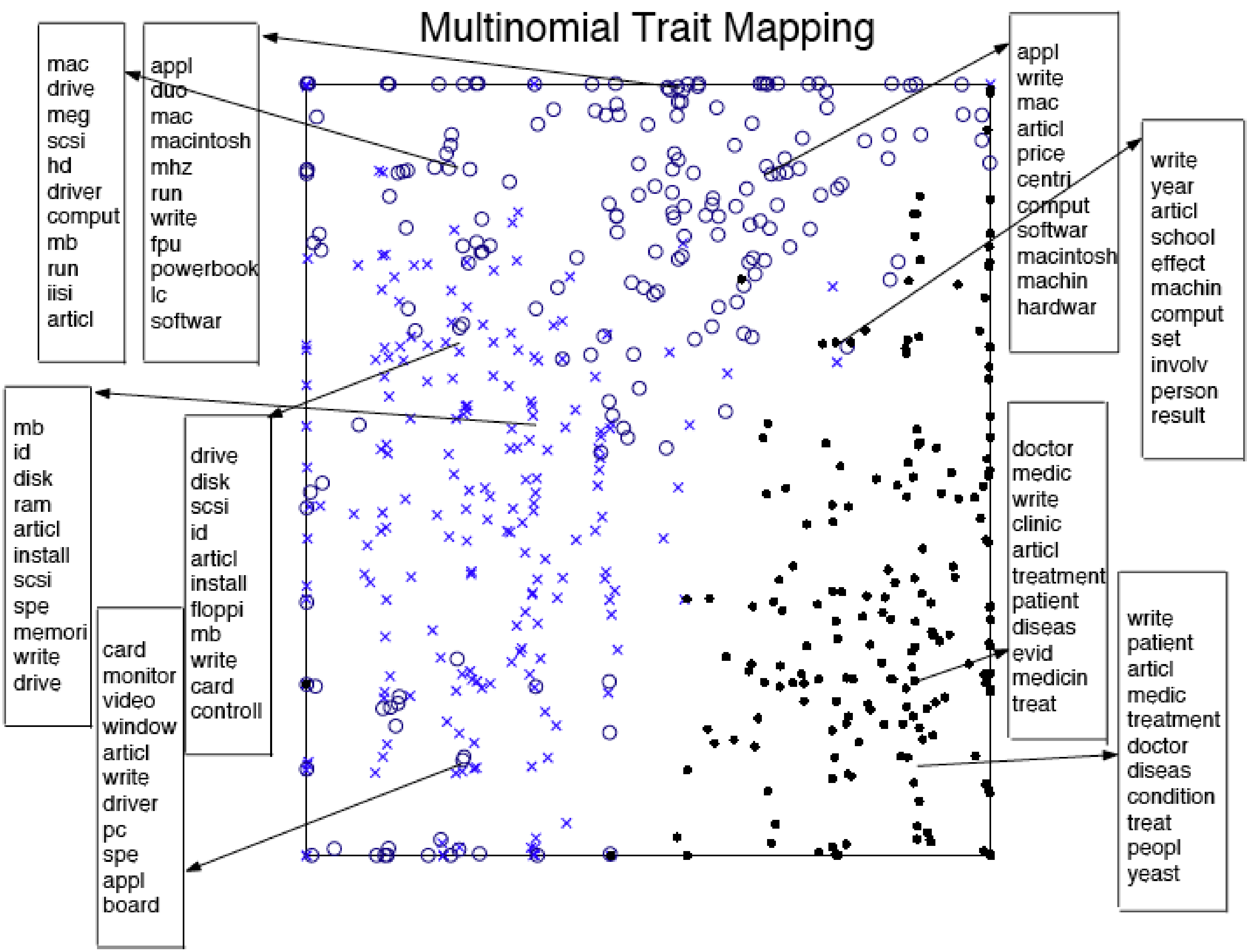
Application:

E.g. in text mining, where for the Bernoulli case $y_k = 1$ indicates that term k is present in a document. The multinomial case is represented by a histogramme of word occurrences. The conditional model represents independent samples from a 'bag of words'. The order of words is irrelevant.


Bernoulli
Gaussian on binary data
Multinomial
Gaussian on normalized freq. data



$\times$ = comp.sys.ibm.pc.hardware, O = comp.sys.mac.hardware, $\cdot$ = sci.med



$\times$ = comp.sys.ibm.pc.hardware, O = comp.sys.mac.hardware, $\cdot$ = sci.med