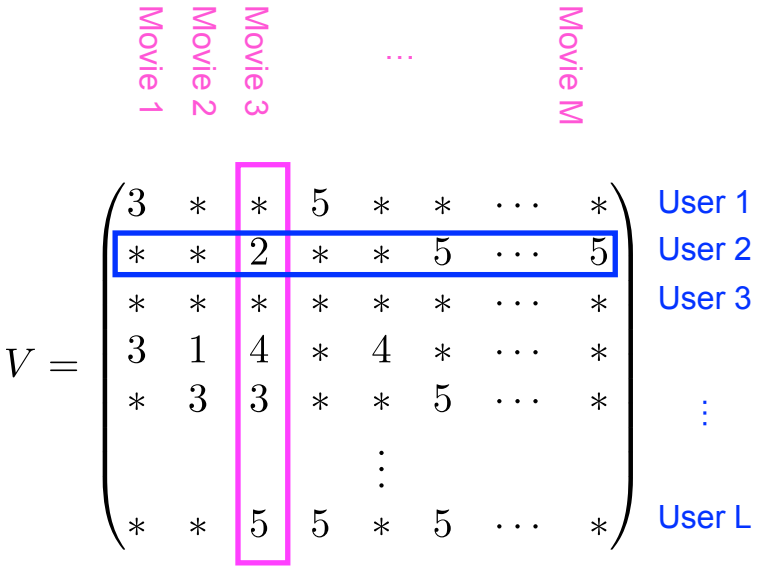
Variational Bayesian Learning

Shinichi nakajima (TU Berlin)
nakajima@tu-berlin.de
http://sites.google.com/site/shinnkj23/

Matrix Factorization Model

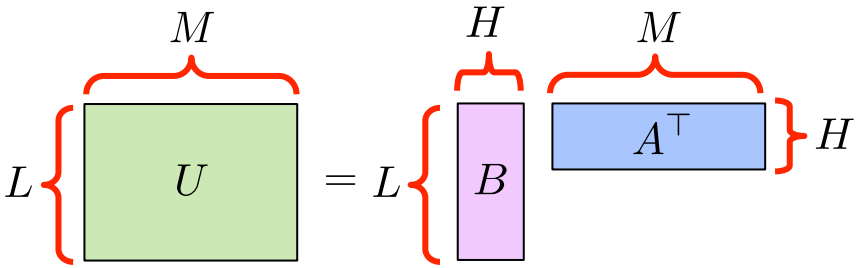
Ex 1. Collaborative Filtering



Predict user preferences, which are not observed.

One can predict * by fitting the data with small degrees of freedom.

Low-rank Approximation



Approximate $V \in \mathbb{R}^{L \times M}$ with a low-rank matrix

$$V = U + \mathcal{E}$$

which can be naturally expressed as the product of two matrices:

$$U = BA^\top$$

Here, $A \in \mathbb{R}^{M \times H}, B \in \mathbb{R}^{L \times H}$

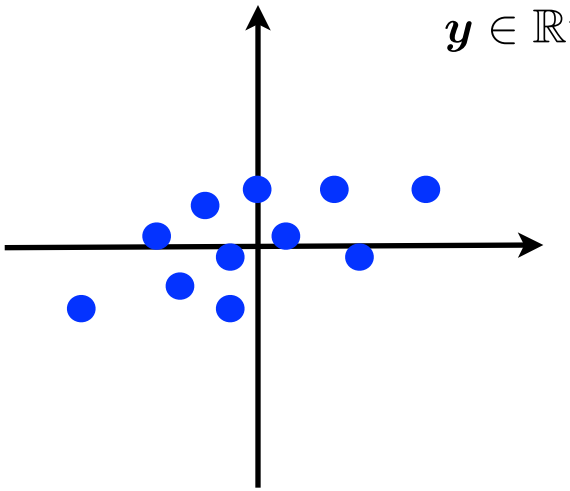
Assuming Gaussian noise, the likelihood is given as

$$p(V|A,B) \propto \exp\left(-\frac{\|V - BA^\top\|_{\mathrm{Fro}}^2}{2\sigma^2}\right)$$

$$\|V\|_{\mathrm{Fro}}^2 = \sum_{l,m} V_{l,m}^2$$

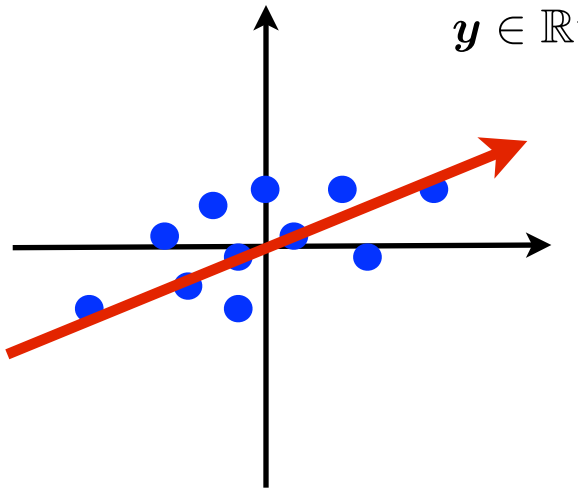
※Missing entries should be ignored.

Ex. 2: Probabilisitic PCA



$\boldsymbol{y} \in \mathbb{R}^L$: observed variable

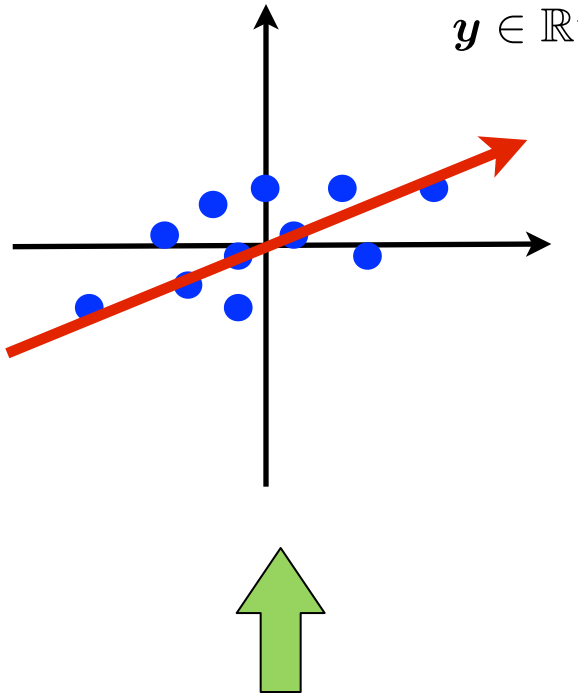
Ex. 2: Probabilisitic PCA

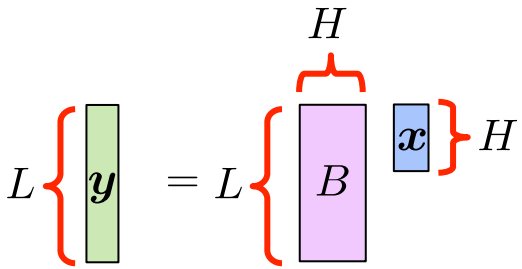


$\boldsymbol{y} \in \mathbb{R}^L$: observed variable

Ex. 2: Probabilisitic PCA

$\boldsymbol{y} \in \mathbb{R}^L$: observed variable



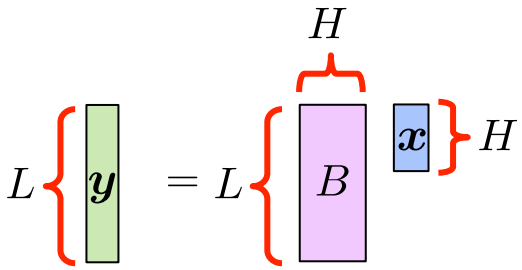


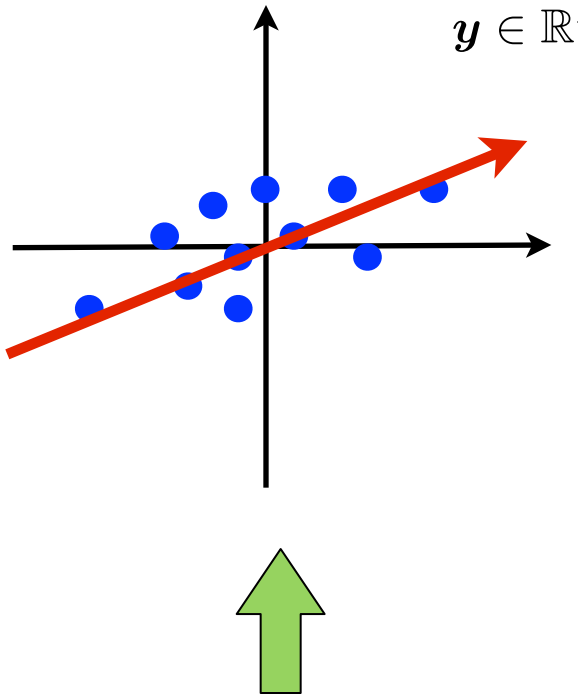
$$\boldsymbol{y} = B\boldsymbol{x} + \boldsymbol{\varepsilon}$$

$B \in \mathbb{R}^{L \times H}$: projection matrix

$\boldsymbol{x} \in \mathbb{R}^H$: hidden variables

Ex. 2: Probabilisitic PCA





$\boldsymbol{y} \in \mathbb{R}^L$: observed variable

For M samples,

$$Y = BX + \mathcal{E}$$

$B \in \mathbb{R}^{L \times H}$: projection matrix

$$Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M) \in \mathbb{R}^{L \times M}$$

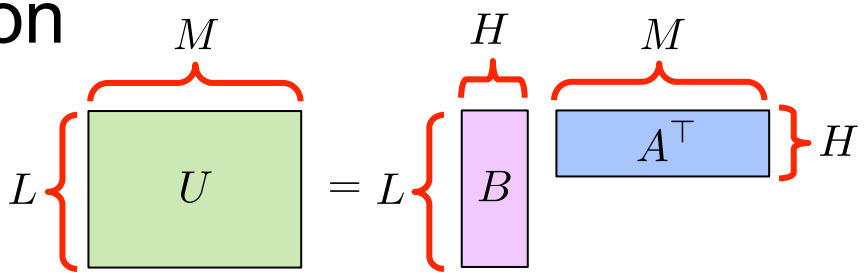
$$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M) \in \mathbb{R}^{H \times M}$$

Since X is also unknown, we should estimate both B and X.

$\boldsymbol{x} \in \mathbb{R}^H$: hidden variables

$Y \to V, X \to A^\top$ gives ...

Bayesian Matrix Factorization

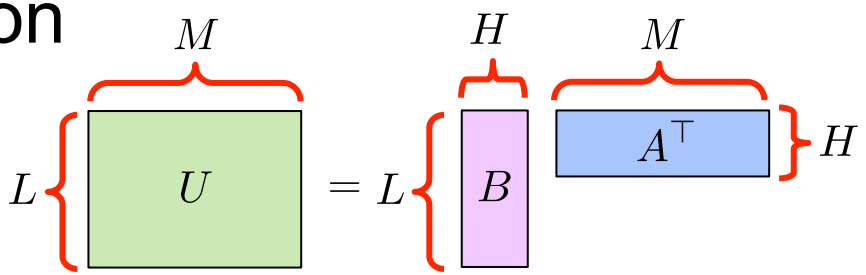


Likelihood: $p(V|A,B) \propto \exp\left(-\frac{\|V - BA^\top\|^2_{\text{Fro}}}{2\sigma^2}\right)$

Observation:
$V \in \mathbb{R}^{L \times M}$

Variational Bayesian Learning

Bayesian Matrix Factorization



Likelihood: $p(V|A,B) \propto \exp\left(-\frac{\|V - BA^\top\|_{\text{Fro}}^2}{2\sigma^2}\right)$

Observation: $V \in \mathbb{R}^{L \times M}$

No conjugate prior because of the quartic term:

$$p(A,B|V) \propto \exp\left(-\frac{\|V - BA^\top\|_{\text{Fro}}^2}{2\sigma^2}\right) \cdot p(A,B)$$

But this is conditionally conjugate:

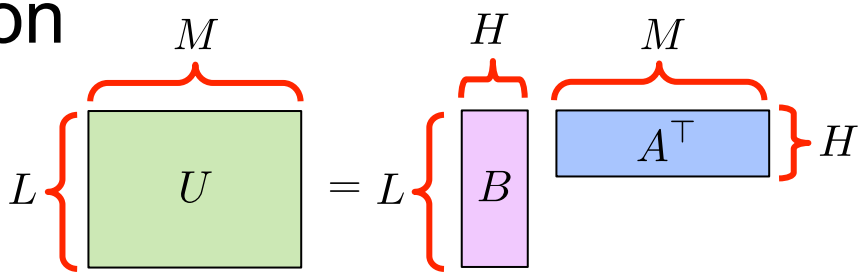
$$p(A|B,V) \propto \exp\left(-\frac{\|V - BA^\top\|_{\text{Fro}}^2}{2\sigma^2}\right) \cdot p(A)$$

$$p(B|A,V) \propto \exp\left(-\frac{\|V - BA^\top\|_{\text{Fro}}^2}{2\sigma^2}\right) \cdot p(B)$$

Independence makes Bayesian learning tractable!

$$\Rightarrow \quad r(A,B) = r_A(A) r_B(B)$$

Bayesian Matrix Factorization



Likelihood: $p(V|A,B) \propto \exp\left(-\frac{\|V - BA^\top\|^2_{\text{Fro}}}{2\sigma^2}\right)$

Prior: $p(A) \propto \exp\left(-\frac{\text{tr}(AC_A^{-1}A^\top)}{2}\right)$

$p(B) \propto \exp\left(-\frac{\text{tr}(BC_B^{-1}B^\top)}{2}\right)$

Observation:

$V \in \mathbb{R}^{L \times M}$

Parameters:

$A \in \mathbb{R}^{M \times H}$

$B \in \mathbb{R}^{L \times H}$

Hyperparameters:

$C_A = \text{diag}(c_{a_1}^2, \ldots, c_{a_H}^2)$

$C_B = \text{diag}(c_{b_1}^2, \ldots, c_{b_H}^2)$

Free energy minimization

Trial distribution: $r(A, B)$

Free energy: $$F(r) = \left\langle \log \frac{r(A,B)}{p(V|A,B)p(A)p(B)} \right\rangle_{r(A,B)}$$

$$= \underbrace{\mathrm{KL}\left(r(A,B) \| p(A,B|V)\right)}_{\text{KL divergence to Bayes posterior}} \underbrace{- \log p(V)}_{\text{marginal likelihood (not depend on r)}}$$

$$\Rightarrow \quad p(A,B|V) = \operatorname*{argmin}_r F(r)$$

Minimizing free energy should give Bayes posterior, but intractable.

Minimize free energy under some constraints!

$$\min_r F(r)$$

$$\text{s.t. } r(A,B) \in \mathcal{G}$$

Free energy minimization

Trial distribution: $r(A, B)$

Free energy: $$F(r) = \left\langle \log \frac{r(A,B)}{p(V|A,B)p(A)p(B)} \right\rangle_{r(A,B)}$$

$$= \underbrace{\mathrm{KL}\left(r(A,B) \| p(A,B|V)\right)}_{\text{KL divergence to Bayes posterior}} - \underbrace{\log p(V)}_{\text{marginal likelihood (not depend on r)}}$$

$$\Rightarrow \quad p(A,B|V) = \operatorname*{argmin}_r F(r)$$

Minimizing free energy should give Bayes posterior, but intractable.

Minimize free energy under some constraints!

$$\min_r F(r)$$

$$\text{s.t. } r(A,B) = r_A(A) r_B(B)$$

Variational method

Free energy

$$F(r) = \left\langle \log \frac{r(A,B)}{p(V|A,B)p(A)p(B)} \right\rangle_{r(A,B)}$$
$$= \int r_A(A) r_B(B) \log \frac{r_A(A) r_B(B)}{p(V|A,B)p(A)p(B)} dA dB$$

$\forall A \in \mathbb{R}^{M \times H}$,

$$0 = \delta I = \frac{\partial F}{\partial r_A} = \int r_B(B) \left(\log \frac{r_A(A) r_B(B)}{p(V|A,B)p(A)p(B)} + 1 \right) dB$$
$$= \left\langle \log \frac{r_A(A) r_B(B)}{p(V|A,B)p(A)p(B)} + 1 \right\rangle_{r_B(B)}$$
$$= \log r_A(A) + \langle \log r_B(B) - \log \left(p(V|A,B)p(A)p(B) \right) \rangle_{r_B(B)} + 1$$
$$= \log r_A(A) - \log p(A) - \langle \log p(V|A,B) \rangle_{r_B(B)} + \text{const.}$$
$$= \log \frac{r_A(A)}{p(A) \exp \left(\langle \log p(V|A,B) \rangle_{r_B(B)} \right)} + \text{const.}$$

Variational method

$$0 = \log \frac{r_A(A)}{p(A) \exp\left(\langle \log p(V|A,B) \rangle_{r_B(B)}\right)} + \text{const.}$$

This must hold for any A!

$$r_A(A) \propto p(A) \exp\left(\langle \log p(V|A,B) \rangle_{r_B(B)}\right)$$

Similarly, considering $\frac{\partial F}{\partial r_B}$ gives

$$r_B(B) \propto p(B) \exp\left(\langle \log p(V|A,B) \rangle_{r_A(A)}\right)$$

Variational method

Local minimizer of $\min_r F(r)$

$$\text{s.t. } r(A,B) = r_A(A) r_B(B)$$

satisfies

$$r_A(A) \propto p(A) \exp\left(\langle \log p(V|A,B) \rangle_{r_B(B)} \right)$$

$$r_B(B) \propto p(B) \exp\left(\langle \log p(V|A,B) \rangle_{r_A(A)} \right)$$

Variational method

$$p(V|A,B) \propto \exp\left(-\frac{\|V - BA^\top\|_{\mathrm{Fro}}^2}{2\sigma^2}\right)$$
$$p(A) \propto \exp\left(-\frac{\mathrm{tr}(AC_A^{-1}A^\top)}{2}\right)$$
$$p(B) \propto \exp\left(-\frac{\mathrm{tr}(BC_B^{-1}B^\top)}{2}\right)$$

Local minimizer of $\min_r F(r)$

$$\text{s.t. } r(A,B) = r_A(A)r_B(B)$$

satisfies

$$r_A(A) \propto p(A)\exp\left(\langle \log p(V|A,B)\rangle_{r_B(B)}\right)$$
$$r_B(B) \propto p(B)\exp\left(\langle \log p(V|A,B)\rangle_{r_A(A)}\right)$$

Focusing on dependence on A, we have

$$\begin{aligned}\langle \log p(V|A,B)\rangle_{r(B)} &= -\frac{1}{2\sigma^2}\left\langle \|V - BA^\top\|_{\mathrm{Fro}}^2\right\rangle_{r(B)} \\ &= -\frac{1}{2\sigma^2}\left\langle \mathrm{tr}(-2V^\top BA^\top + AB^\top BA^\top)\right\rangle_{r(B)} + \text{const.} \\ &= -\frac{1}{2\sigma^2}\mathrm{tr}\left(-2V^\top \langle B\rangle_{r_B(B)} A^\top + A\left\langle B^\top B\right\rangle_{r_B(B)} A^\top\right) + \text{const.}\end{aligned}$$

$$r_A(A) \propto \exp\left(-\frac{1}{2\sigma^2}\mathrm{tr}\left(-2V^\top\langle B\rangle_{r_B(B)}A^\top + A\left\langle B^\top B\right\rangle_{r_B(B)} A^\top + \sigma^2 AC_A^{-1}A^\top\right)\right)$$

$r_A(A)$ is Gaussian!

Variational method

$$r_A(A) \propto \exp\left(-\frac{\operatorname{tr}\left((A-\widehat{A})\widehat{\Sigma}_A^{-1}(A-\widehat{A})\right)}{2}\right)$$

$$\widehat{A} = V^\top \langle B \rangle_{r_B(B)} \frac{\widehat{\Sigma}_A}{\sigma^2} \qquad \widehat{\Sigma}_A = \sigma^2 \left(\left\langle B^\top B \right\rangle_{r_B(B)} + \sigma^2 C_A^{-1} \right)^{-1}$$

Similarly we have

$$r_B(B) \propto \exp\left(-\frac{\operatorname{tr}\left((B-\widehat{B})\widehat{\Sigma}_B^{-1}(B-\widehat{B})\right)}{2}\right)$$

$$\widehat{B} = V^\top \langle A \rangle_{r_A(A)} \frac{\widehat{\Sigma}_B}{\sigma^2} \qquad \widehat{\Sigma}_B = \sigma^2 \left(\left\langle A^\top A \right\rangle_{r_A(A)} + \sigma^2 B_B^{-1} \right)^{-1}$$

Since we now know $r_A(A)$ and $r_B(B)$ are Gaussian, the moments are given as

$$\langle A \rangle_{r_A(A)} = \widehat{A} \qquad \left\langle A^\top A \right\rangle_{r_A(A)} = \widehat{A}^\top \widehat{A} + M\widehat{\Sigma}_A$$

$$\langle B \rangle_{r_B(B)} = \widehat{B} \qquad \left\langle B^\top B \right\rangle_{r_B(B)} = \widehat{B}^\top \widehat{B} + L\widehat{\Sigma}_B$$

VB posterior

$$r_A(A) \propto \exp\left(-\frac{\mathrm{tr}\left((A-\widehat{A})\widehat{\Sigma}_A^{-1}(A-\widehat{A})\right)}{2}\right)$$

$$r_B(B) \propto \exp\left(-\frac{\mathrm{tr}\left((B-\widehat{B})\widehat{\Sigma}_B^{-1}(B-\widehat{B})\right)}{2}\right)$$

$$\widehat{A} = V^\top \widehat{B}\frac{\widehat{\Sigma_A}}{\sigma^2} \qquad \widehat{\Sigma}_A = \sigma^2\left(\widehat{B}^\top\widehat{B} + L\widehat{\Sigma}_B + \sigma^2 C_A^{-1}\right)^{-1}$$

$$\widehat{B} = V\widehat{A}\frac{\widehat{\Sigma}_B}{\sigma^2} \qquad \widehat{\Sigma}_B = \sigma^2\left(\widehat{A}^\top\widehat{A} + M\widehat{\Sigma}_A + \sigma^2 C_B^{-1}\right)^{-1}$$

Iterative algorithm

Starting from initial values for $\widehat{A}, \widehat{\Sigma}_A, \widehat{B}, \widehat{\Sigma}_B$,
they are updated by equations above.

Hyperparameter Estimation

Likelihood: $$p(V|A,B) \propto \exp\left(-\frac{\|V - BA^\top\|_{\mathrm{Fro}}^2}{2\sigma^2}\right)$$

Prior: $$p(A) \propto \exp\left(-\frac{\mathrm{tr}(AC_A^{-1}A^\top)}{2}\right)$$

$$p(B) \propto \exp\left(-\frac{\mathrm{tr}(BC_B^{-1}B^\top)}{2}\right)$$

(σ^2, C_A, C_B) can also be estimated by minimizing free energy

$$F = \underbrace{\mathrm{KL}\left(r(A,B)\|p(A,B|V,\sigma^2,C_A,C_B)\right)}_{\text{minimize KL divergence to Bayes posterior}} \underbrace{- \log p(V|\sigma^2,C_A,C_B)}_{\text{make marginal likelihood larger}}$$

minimize KL divergence to Bayes posterior, and make marginal likelihood larger.

Hyperparameter Estimation

F can be explicitly written as a function of $(A, B, \Sigma_A, \Sigma_B, C_A, C_B, \sigma^2)$.

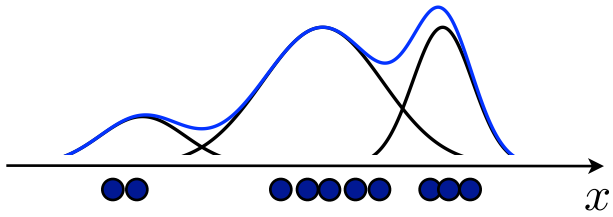
(will be a homework)

Differentiating F, we have update rules for hyperparameters:

$$
\begin{aligned}
c_{a_h}^2 &= \|\widehat{\boldsymbol{a}}_h\|^2/M + (\Sigma_A)_{hh} \\
c_{b_h}^2 &= \|\widehat{\boldsymbol{b}}_h\|^2/L + (\Sigma_B)_{hh} \\
\sigma^2 &= \frac{\|V\|_{\mathrm{Fro}}^2 - \mathrm{tr}(2V^\top \widehat{B}\widehat{A}^\top) + \mathrm{tr}\left((\widehat{A}^\top\widehat{A} + M\Sigma_A)(\widehat{B}^\top\widehat{B} + L\Sigma_B)\right)}{LM}
\end{aligned}
$$

Mixture of Gaussians

Mixture of Gaussians



Observed var.: $\boldsymbol{x} \in \mathbb{R}^M$

Latent var.: $\boldsymbol{z} \in \{\mathbf{1}_k\}_{k=1}^K$

$\mathbf{1}_k = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^K$

↑

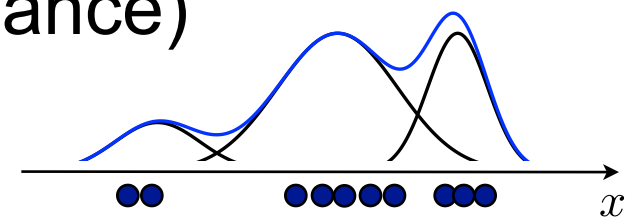
The k-th element

parameters: $\{a_k, \boldsymbol{\mu}_k, \Sigma_k; 0 \leq a_k \leq 1, \sum_{k=1}^K a_k = 1\}$

likelihood: $p(\{\boldsymbol{x}^{(n)}, \boldsymbol{z}^{(n)}\}_{n=1}^N | a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{h=1}^H) = \prod_{n=1}^N \prod_{k=1}^K \left(a_k \mathcal{N}_M(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)^{z_k^{(n)}}$

Prior: $p(\{a_k\}_{k=1}^K | \phi) \propto \prod_{k=1}^K a_k^{\phi - 1}$ $\quad p(\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K) \propto \prod_{k=1}^K \mathcal{N}_M(\boldsymbol{\mu}_k; \boldsymbol{d}_0, c_\mu^2 \boldsymbol{\Sigma}) \mathcal{W}(\boldsymbol{\Sigma}^{-1}; \rho, \Lambda)$

Mixture of Gaussians (known covariance)



Observed var.: $\boldsymbol{x} \in \mathbb{R}^M$

Latent var.: $\boldsymbol{z} \in \{\mathbf{1}_k\}_{k=1}^K$

$\mathbf{1}_k = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^K$

↑

The k-th element

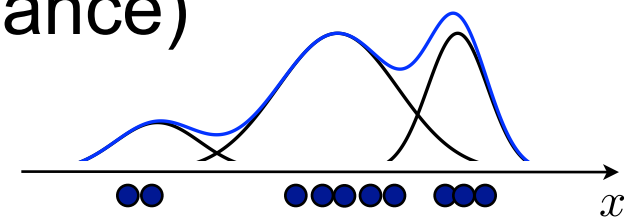
parameters: $\{a_k, \boldsymbol{\mu}_k, \Sigma_k; 0 \leq a_k \leq 1, \sum_{k=1}^K a_k = 1\}$

likelihood: $p(\{\boldsymbol{x}^{(n)}, \boldsymbol{z}^{(n)}\}_{n=1}^N | a_k, \boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{n=1}^N \prod_{k=1}^K \left(a_k \mathcal{N}_M(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{I}_M)\right)^{z_k^{(n)}}$

Prior: $p(\{a_k\}_{k=1}^K | \phi) \propto \prod_{k=1}^K a_k^{\phi - 1}$

$p(\{\boldsymbol{\mu}_k\}_{k=1}^K) \propto \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k; 0, \sigma_0^2 \boldsymbol{I}_M)$

Mixture of Gaussians (known covariance)



Observed var.: $\boldsymbol{x} \in \mathbb{R}^M$

Latent var.: $\boldsymbol{z} \in \{\mathbf{1}_k\}_{k=1}^K$

$\mathbf{1}_k = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^K$

↑ The k-th element

parameters: $\{a_k, \boldsymbol{\mu}_k, \Sigma_k; 0 \leq a_k \leq 1, \sum_{k=1}^K a_k = 1\}$

likelihood: $p(\{\boldsymbol{x}^{(n)}, \boldsymbol{z}^{(n)}\}_{n=1}^N | a_k, \boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{n=1}^N \prod_{k=1}^K \left(a_k \mathcal{N}_M(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{I}_M)\right)^{z_k^{(n)}}$

Prior: $p(\{a_k\}_{k=1}^K | \phi) \propto \prod_{k=1}^K a_k^{\phi - 1}$ $\quad p(\{\boldsymbol{\mu}_k\}_{k=1}^K) \propto \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k; 0, \sigma_0^2 \boldsymbol{I}_M)$

Independence constraint:

$$r(\mathcal{H}, \boldsymbol{w}) = r_{\mathcal{H}}(\mathcal{H}) r_w(\boldsymbol{w})$$

where $\mathcal{H} = \{\boldsymbol{z}^{(n)}\}_{n=1}^N$,

$\boldsymbol{w} = \{a_k, \boldsymbol{\mu}_k\}_{k=1}^K$

Variational Bayesian Learning

$$\begin{aligned} &\min F(r), \\ \text{s.t. } &r(\mathcal{H}, \boldsymbol{w}) = r_{\mathcal{H}}(\mathcal{H}) r_w(\boldsymbol{w}) \end{aligned}$$

where $$F(r) = \left\langle \log \frac{r_{\mathcal{H}}(\mathcal{H}) r_\omega(\boldsymbol{\omega})}{p(\mathcal{D}, \mathcal{H}|\boldsymbol{\omega}) p(\boldsymbol{\omega})} \right\rangle_{r_{\mathcal{H}}(\mathcal{H}) r_\omega(\boldsymbol{\omega})}$$

$$= \sum_{\mathcal{H}} \int r_{\mathcal{H}}(\mathcal{H}) r_\omega(\boldsymbol{\omega}) \log \frac{r_{\mathcal{H}}(\mathcal{H}) r_\omega(\boldsymbol{\omega})}{p(\mathcal{D}, \mathcal{H}|\boldsymbol{\omega}) p(\boldsymbol{\omega})} d\boldsymbol{\omega}$$

Applying variational method gives

$$r_{\mathcal{H}}(\mathcal{H}) \propto \exp \left\langle \log p(\mathcal{D}, \mathcal{H}|\boldsymbol{\omega}) \right\rangle_{r_\omega(\boldsymbol{\omega})}$$

$$r_\omega(\boldsymbol{\omega}) \propto p(\boldsymbol{\omega}) \exp \left\langle \log p(\mathcal{D}, \mathcal{H}|\boldsymbol{\omega}) \right\rangle_{r_{\mathcal{H}}(\mathcal{H})}$$

Variational Bayesian Learning

Thus,

$$r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N) \propto \prod_{n=1}^N \left(\prod_{k=1}^K \left(\overline{z}_k^{(n)}\right)^{z_k^{(n)}} \right)$$

Multinomial

where $$\overline{\boldsymbol{z}}_k^{(n)} = \exp\left(\langle \log \alpha_k \rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha},\{\boldsymbol{\mu}_k\}_{k=1}^K)} - \frac{1}{2} \left\langle \|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\|^2 \right\rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha},\{\boldsymbol{\mu}_k\}_{k=1}^K)} \right)$$

$$r_{\alpha,\mu}(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K) \propto \prod_{k=1}^K \alpha_k^{\overline{N}_k + \phi - 1} \exp\left(- \frac{(\overline{N}_k + \sigma_0^{-2}) \|\boldsymbol{\mu}_k - \frac{\overline{N}_k \overline{\boldsymbol{x}}_k}{\overline{N}_k + \sigma_0^{-2}}\|^2}{2} \right)$$

Dirichlet x Gauss (independent!)

where

$$\overline{N}_k = \sum_{n=1}^N \langle z_k^{(n)} \rangle_{r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N)}$$

$$\overline{\boldsymbol{x}}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^N \boldsymbol{x}^{(n)} \langle z_k^{(n)} \rangle_{r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N)}$$

Variational Bayesian Learning

$$\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$$

Since we now know form of posteriors,

$$r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N) = \prod_{n=1}^N \text{Multinomial}_{K,1}\left(\boldsymbol{z}^{(n)}; \widehat{\boldsymbol{z}}^{(n)}\right)$$

$$r_\alpha(\boldsymbol{\alpha}) = \text{Dirichlet}\left(\boldsymbol{\alpha}; \widehat{\boldsymbol{\alpha}}\right)$$

$$r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{k=1}^K \text{Norm}_M\left(\boldsymbol{\mu}_k; \widehat{\boldsymbol{\mu}}_k, \widehat{\sigma}_k^2 \boldsymbol{I}_M\right)$$

we can compute the expectations

$$\left\langle z_k^{(n)} \right\rangle_{r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N)} = \widehat{z}_k^{(n)}$$

$$\begin{aligned}\left\langle \log \alpha_k \right\rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)} &= \left\langle \log \alpha_k \right\rangle_{r_\alpha(\boldsymbol{\alpha})} \\ &= \Psi(\widehat{\alpha}_k) - \Psi\left(\sum_{k'=1}^K \widehat{\alpha}_{k'}\right)\end{aligned}$$

$$\begin{aligned}\left\langle \|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\|^2 \right\rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)} &= \left\langle \|(\boldsymbol{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k) + (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)\|^2 \right\rangle_{r(\boldsymbol{\mu}_k)} \\ &= \|\boldsymbol{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\|^2 + M\widehat{\sigma}_k^2\end{aligned}$$

Variational Bayesian Learning

$$\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$$

and the update rules:

$$\widehat{z}_k^{(n)} = \frac{\overline{z}_k^{(n)}}{\sum_{k'=1}^{K} \overline{z}_{k'}^{(n)}}$$

$$\widehat{\alpha}_k = \overline{N}_k + \phi$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{\overline{N}_k \overline{\boldsymbol{x}}_k}{\overline{N}_k + \sigma_0^{-2}}$$

$$\widehat{\sigma}_k^2 = \frac{1}{\overline{N}_k + \sigma_0^{-2}}$$

where

$$\overline{\boldsymbol{z}}_k^{(n)} = \exp\left(\Psi(\widehat{\alpha}_k) - \frac{1}{2}\left\|\boldsymbol{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\right\|^2 + M\widehat{\sigma}_k^2 + \text{const.}\right)$$

$$\overline{N}_k = \sum_{n=1}^{N} \widehat{z}_k^{(n)}$$

$$\overline{\boldsymbol{x}}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^{N} \boldsymbol{x}^{(n)} \widehat{z}_k^{(n)}$$

Variational Bayesian Learning

$$\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$$

and the update rules:

$$\widehat{z}_k^{(n)} = \frac{\overline{z}_k^{(n)}}{\sum_{k'=1}^{K} \overline{z}_{k'}^{(n)}}$$

$$\widehat{\alpha}_k = \overline{N}_k + \phi$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{\overline{N}_k \overline{\boldsymbol{x}}_k}{\overline{N}_k + \sigma_0^{-2}}$$

$$\widehat{\sigma}_k^2 = \frac{1}{\overline{N}_k + \sigma_0^{-2}}$$

Iterative algorithm

Iterate these equations until convergence.

where

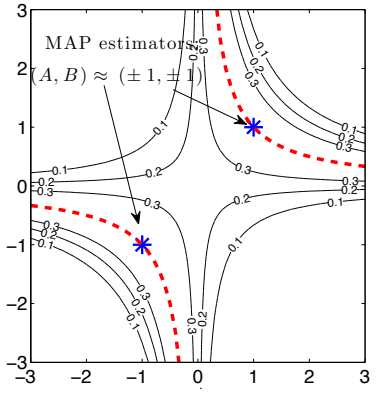
$$\overline{\boldsymbol{z}}_k^{(n)} = \exp\left(\Psi(\widehat{\alpha}_k) - \frac{1}{2}\left\|\boldsymbol{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\right\|^2 + M\widehat{\sigma}_k^2 + \text{const.}\right)$$

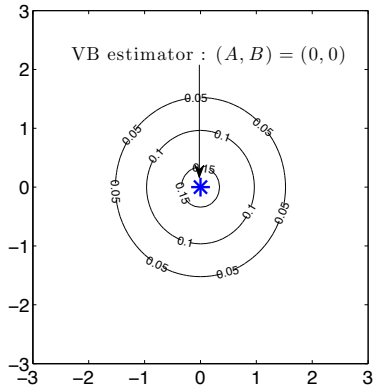
$$\overline{N}_k = \sum_{n=1}^{N} \widehat{z}_k^{(n)}$$

$$\overline{\boldsymbol{x}}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^{N} \boldsymbol{x}^{(n)} \widehat{z}_k^{(n)}$$

Properties of Variational Bayes

- Applicable to many models with conditional conjugacy (typically, models built up with well-known distributions.
- (Generally) faster than sampling methods, and slower than MAP.
- Provides model selection procedure, and tends to give sparse solution.
- (Sometimes crude) approximation. Some theoretical guarantee was obtained for special cases.






Technische
Universität
Berlin
berlin