Summary

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

Bayesian learning

Posterior: $$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})}$$

where $$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

Analytically computed in the conjugate cases,
e.g., Gaussian, Multinomial, etc.

Approximation methods

- Conditionally conjugate

  (Gaussian MF, Mixture of Gaussians, LDA)

  - Gibbs sampling
  - Variational Bayes

- Non-conjugate (likelihood with sigmoid function)
  - Metropolis-Hastings
  - Local variational Bayes, Expectation Propagation

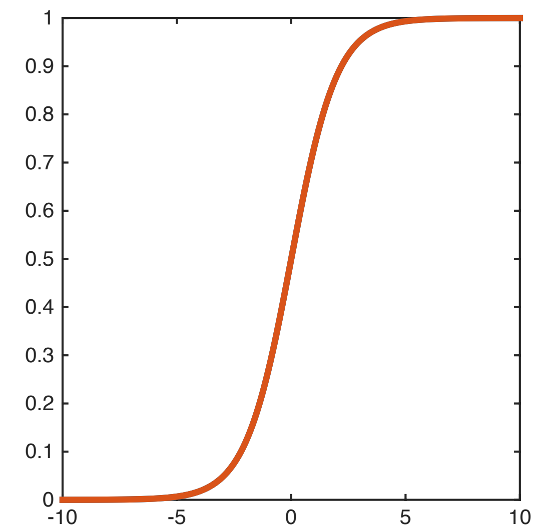
Approximation methods

- Conditionally conjugate
  (Gaussian MF, Mixture of Gaussians, LDA)
  - Gibbs sampling
  - Variational Bayes
- Non-conjugate (likelihood with sigmoid function)
  - Metropolis-Hastings
  - Local variational Bayes, Expectation Propagation

Logistic regression

$$p(y|\boldsymbol{w}) = \psi(y\boldsymbol{w}^\top\boldsymbol{x})$$

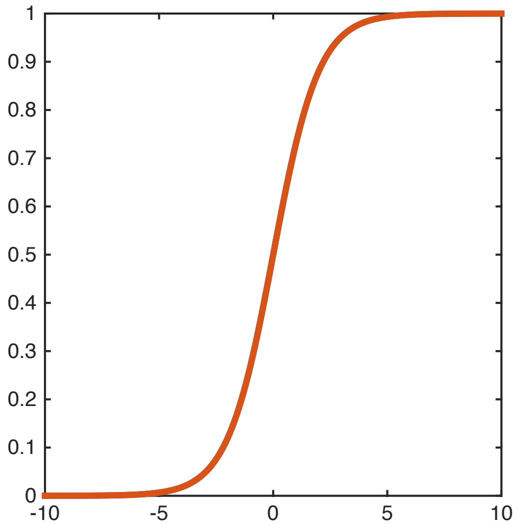
$$y \in \{-1, 1\},$$
$$\boldsymbol{x} \in \mathbb{R}^D,$$
$$\boldsymbol{w} \in \mathbb{R}^D,$$
$$\psi(z) = \frac{1}{1+e^{-z}}$$



Logistic regression

$$y \in \{-1, 1\},$$
$$\boldsymbol{x} \in \mathbb{R}^D,$$
$$\boldsymbol{w} \in \mathbb{R}^D,$$
$$\psi(z) = \frac{1}{1+e^{-z}}$$

$$
\begin{aligned}
p(y|\boldsymbol{w}) &= \psi(y\boldsymbol{w}^\top \boldsymbol{x}) \\
&= \frac{1}{1+e^{-y\boldsymbol{w}^\top \boldsymbol{x}}} \\
&= \frac{e^{\frac{y\boldsymbol{w}^\top \boldsymbol{x}}{2}}}{e^{\frac{y\boldsymbol{w}^\top \boldsymbol{x}}{2}} + e^{-\frac{y\boldsymbol{w}^\top \boldsymbol{x}}{2}}} \\
&= \frac{e^{\frac{y\boldsymbol{w}^\top \boldsymbol{x}}{2}}}{e^{\frac{\boldsymbol{w}^\top \boldsymbol{x}}{2}} + e^{-\frac{\boldsymbol{w}^\top \boldsymbol{x}}{2}}} \\
&\propto e^{\frac{y\boldsymbol{w}^\top \boldsymbol{x}}{2}}
\end{aligned}
$$



Bayesian logistic regression

$$p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})$$

$$= \prod_{n=1}^{N} \underbrace{\psi(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)})}_{\text{No conjugate prior}} \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right)$$

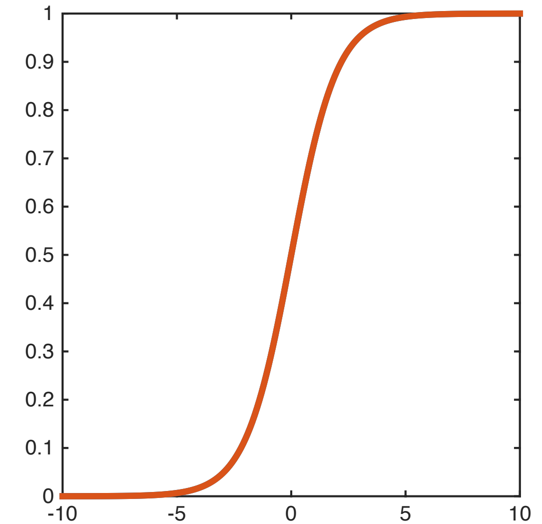
No conjugate prior

$$y \in \{-1, 1\},$$

$$\boldsymbol{x} \in \mathbb{R}^D,$$

$$\boldsymbol{w} \in \mathbb{R}^D,$$

$$\psi(z) = \frac{1}{1+e^{-z}}$$



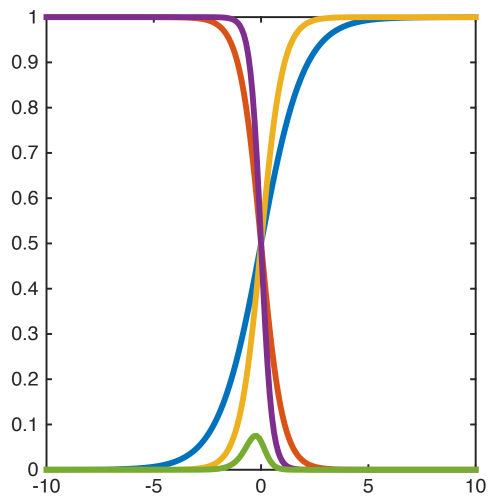
Bayesian logistic regression

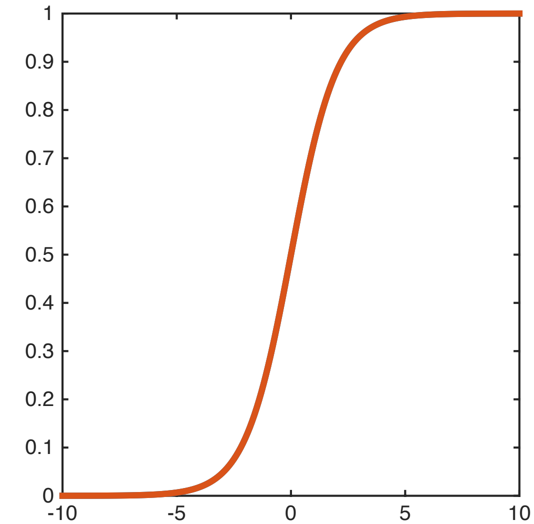
$$p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})$$
$$= \underbrace{\prod_{n=1}^{N} \psi(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)})} \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right)$$

$$y \in \{-1, 1\},$$
$$\boldsymbol{x} \in \mathbb{R}^D,$$
$$\boldsymbol{w} \in \mathbb{R}^D,$$
$$\psi(z) = \frac{1}{1+e^{-z}}$$

No conjugate prior

→ Approximate with Gaussian





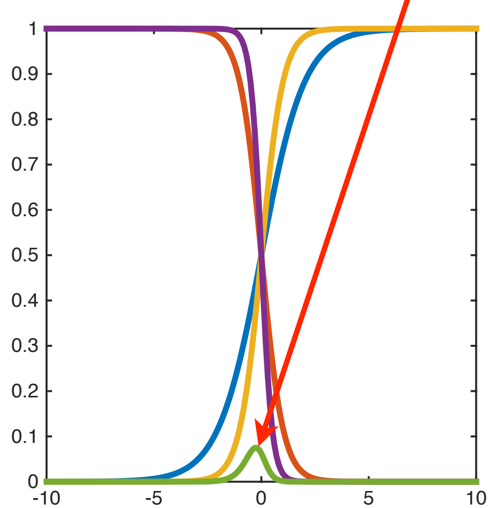
Bayesian logistic regression

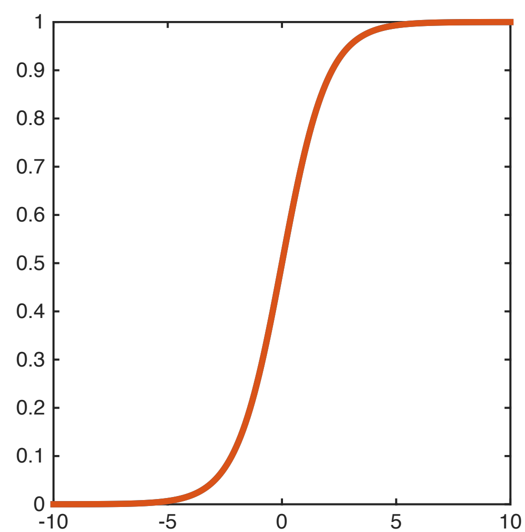
$$p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})$$
$$= \prod_{n=1}^{N} \underbrace{\psi(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)})} \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right)$$

$$y \in \{-1, 1\},$$
$$\boldsymbol{x} \in \mathbb{R}^D,$$
$$\boldsymbol{w} \in \mathbb{R}^D,$$
$$\psi(z) = \frac{1}{1+e^{-z}}$$

No conjugate prior

→ Approximate with Gaussian





Local variational approximation

$$p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})$$
$$= \prod_{n=1}^{N} \underbrace{\psi(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)})} \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right)$$

$$y \in \{-1, 1\},$$
$$\boldsymbol{x} \in \mathbb{R}^D,$$
$$\boldsymbol{w} \in \mathbb{R}^D,$$
$$\psi(z) = \frac{1}{1+e^{-z}}$$

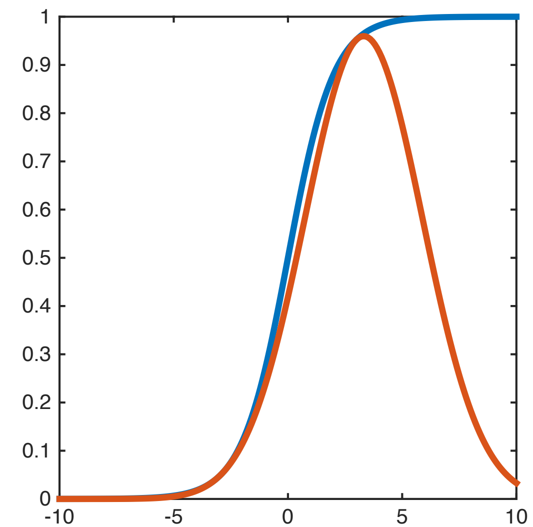
No conjugate prior

Approximate with (unnormalized) Gaussian

$$\underline{\psi}(z;\xi) = \psi(\xi)\exp\left(\frac{z-\xi}{2} + \frac{2\psi(\xi)-1}{4\xi}(z^2-\xi^2)\right)$$

$$\forall \xi > 0, \quad \psi(z) \geq \underline{\psi}(z;\xi)$$

$$p(\boldsymbol{w}|\mathcal{D}) \approx \prod_{n=1}^{N} \underline{\psi}\left(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)}; \xi^{(n)}\right) \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right)$$



Local variational approximation

$$p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})$$
$$= \prod_{n=1}^{N} \underbrace{\psi(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)})} \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right)$$

$$y \in \{-1, 1\},$$
$$\boldsymbol{x} \in \mathbb{R}^D,$$
$$\boldsymbol{w} \in \mathbb{R}^D,$$
$$\psi(z) = \frac{1}{1+e^{-z}}$$

No conjugate prior

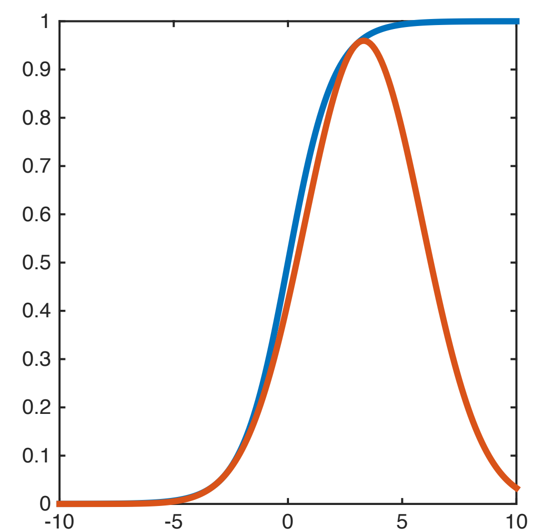
Approximate with (unnormalized) Gaussian

$$\underline{\psi}(z;\xi) = \psi(\xi)\exp\left(\frac{z-\xi}{2} + \frac{2\psi(\xi)-1}{4\xi}(z^2-\xi^2)\right)$$

$$\forall \xi > 0, \quad \psi(z) \geq \underline{\psi}(z;\xi)$$

$$p(\boldsymbol{w}|\mathcal{D}) \approx \prod_{n=1}^{N} \underline{\psi}\left(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)}; \xi^{(n)}\right) \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right)$$

$$\{\xi^{(n)}\} = \underset{\{\xi^{(n)}\}}{\mathrm{argmin}}\, F$$



where $F = -\log \widetilde{p}(\mathcal{D}) = -\log \int \prod_{n=1}^{N} \underline{\psi}\left(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)}; \xi^{(n)}\right) \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right) d\boldsymbol{w}$

Expectation propagation

$$p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})$$
$$= \underbrace{\prod_{n=1}^{N} \psi(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)})}_{} \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right)$$

$$y \in \{-1, 1\},$$
$$\boldsymbol{x} \in \mathbb{R}^D,$$
$$\boldsymbol{w} \in \mathbb{R}^D,$$
$$\psi(z) = \frac{1}{1+e^{-z}}$$

No conjugate prior

Approximate with (unnormalized) Gaussian

$$\widetilde{r}(y\boldsymbol{w}^\top \boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{Z}\mathrm{Gauss}_1(y\boldsymbol{w}^\top \boldsymbol{x}; \mu, \sigma^2)$$

$$\boldsymbol{\theta} = \begin{pmatrix} Z \\ \mu \\ \sigma^2 \end{pmatrix}$$

$$p(\boldsymbol{w}|\mathcal{D}) \approx r(\boldsymbol{w}; \boldsymbol{\Theta})$$
$$= \exp\left(-\frac{\boldsymbol{w}^\top \boldsymbol{C}^{-1}\boldsymbol{w}}{2}\right) \prod_{n=1}^{N} \widetilde{r}\left(y^{(n)}\boldsymbol{w}^\top \boldsymbol{x}^{(n)}; \boldsymbol{\theta}^{(n)}\right)$$

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(n)}\}_{n=1}^{N}$$

where $\boldsymbol{\Theta} = \underset{\boldsymbol{\Theta}}{\mathrm{argmin}}\, \mathrm{KL}(p(\boldsymbol{w}|\mathcal{D}) \| r(\boldsymbol{w}; \boldsymbol{\Theta}))$

Expectation propagation

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -\frac{1}{2\sigma^2} \end{pmatrix}$$

$$\boldsymbol{\Theta} = \operatorname*{argmin}_{\boldsymbol{\Theta}} \mathrm{KL}(p(\boldsymbol{w}|\mathcal{D}) \| r(\boldsymbol{w}; \boldsymbol{\Theta}))$$

Exponential family: $r(\boldsymbol{w}|\boldsymbol{\eta}) = h(\boldsymbol{w}) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{w})\right)$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\eta}} \left\{ \langle \log p(\boldsymbol{w}|\mathcal{D}) \rangle_{p(\boldsymbol{w}|\mathcal{D})} - \langle \log r(\boldsymbol{w}|\boldsymbol{\eta}) \rangle_{p(\boldsymbol{w}|\mathcal{D})} \right\} &= -\frac{\partial}{\partial \boldsymbol{\eta}} \log g(\boldsymbol{\eta}) - \langle \boldsymbol{T}(\boldsymbol{w}) \rangle_{p(\boldsymbol{w}|\mathcal{D})} \\ &= \langle \boldsymbol{T}(\boldsymbol{w}) \rangle_{r(\boldsymbol{w}|\boldsymbol{\eta})} - \langle \boldsymbol{T}(\boldsymbol{w}) \rangle_{p(\boldsymbol{w}|\mathcal{D})} \end{aligned}$$

Moment matching!

$$\begin{aligned} \because -\frac{\partial}{\partial \boldsymbol{\eta}} \log g(\boldsymbol{\eta}) &= \frac{\partial}{\partial \boldsymbol{\eta}} \log \int h(\boldsymbol{w}) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{w})\right) d\boldsymbol{w} \\ &= \frac{\int \boldsymbol{T}(\boldsymbol{w}) h(\boldsymbol{w}) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{w})\right) d\boldsymbol{w}}{\int h(\boldsymbol{w}) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{w})\right) d\boldsymbol{w}} \\ &= \langle \boldsymbol{T}(\boldsymbol{w}) \rangle_{r(\boldsymbol{w}|\boldsymbol{\eta})} \end{aligned}$$

Expectation propagation

$$\langle \boldsymbol{T}(\boldsymbol{w}) \rangle_{r(\boldsymbol{w}|\boldsymbol{\eta})} = \langle \boldsymbol{T}(\boldsymbol{w}) \rangle_{p(\boldsymbol{w}|\mathcal{D})}$$

Moment matching!

current distribution: $r(\boldsymbol{w}|\boldsymbol{\theta}) = p(\boldsymbol{w}) \prod_{n=1}^{N} \widetilde{r}(y^{(n)} \boldsymbol{w}^\top \boldsymbol{x}^{(n)} | \boldsymbol{\theta}^{(n)})$

Step 1: $\widetilde{r}^{\setminus n}(\boldsymbol{w}) = \dfrac{\widetilde{r}(\boldsymbol{w})}{\widetilde{r}(\boldsymbol{w}|\boldsymbol{\theta}^{(n)})}$

Step 2: $\widetilde{r}(\boldsymbol{w}|\boldsymbol{\theta}^{(n)}) = \dfrac{1}{Z^{(n)}} \mathrm{Gauss}(y^{(n)} \boldsymbol{w}^\top \boldsymbol{x}^{(n)}; \mu^{(n)}, \sigma^{2(n)})$

where

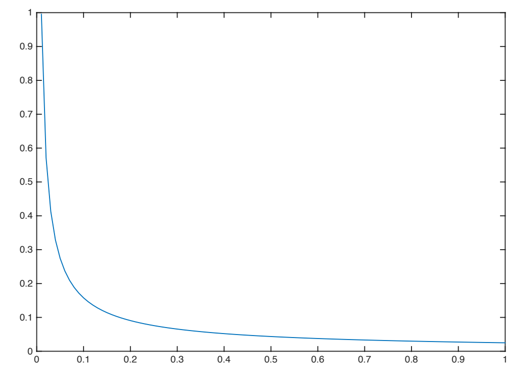
$$Z^{(n)} = \int \widetilde{r}^{\setminus n}(\boldsymbol{w}) \psi(y^{(n)} \boldsymbol{w}^\top \boldsymbol{x}^{(n)}) d\boldsymbol{w},$$

$$\mu^{(n)} = \int y^{(n)} \boldsymbol{w}^\top \boldsymbol{x}^{(n)} \widetilde{r}^{\setminus n}(\boldsymbol{w}) \psi(y^{(n)} \boldsymbol{w}^\top \boldsymbol{x}^{(n)}) d\boldsymbol{w},$$

$$\sigma^{2(n)} = \int (y^{(n)} \boldsymbol{w}^\top \boldsymbol{x}^{(n)})^2 \widetilde{r}^{\setminus n}(\boldsymbol{w}) \psi(y^{(n)} \boldsymbol{w}^\top \boldsymbol{x}^{(n)}) d\boldsymbol{w}.$$

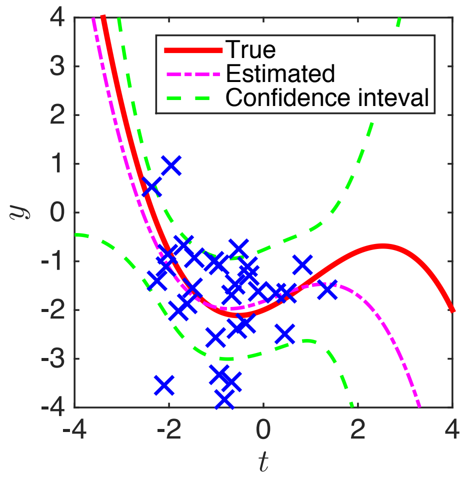
1-D numerical integration is required in each iteration.

Bayesian learning



Pros:

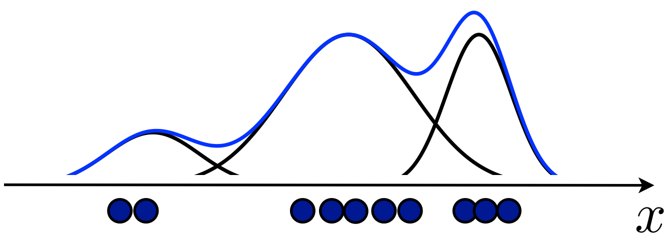
- Less prone to overfitting.
- Information on uncertainty is available.
- All unknowns (hyperparameters) can be estimated from observation through Bayesian model selection.



Cons:

- Integral computation is required.

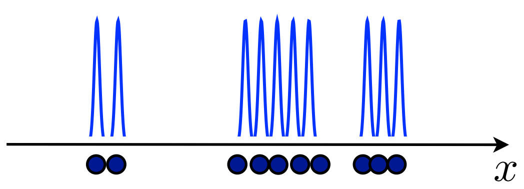
Clustering



Mixture models:

$$p(x) = \sum_{h=1}^{H} a_h \mathcal{N}(x; \mu_h, \sigma_h^2)$$

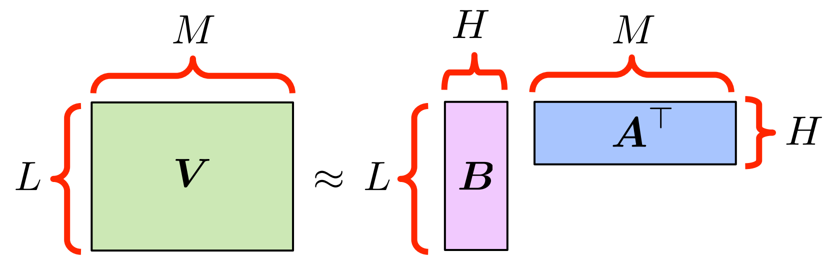
Maximum likelihood estimation results in



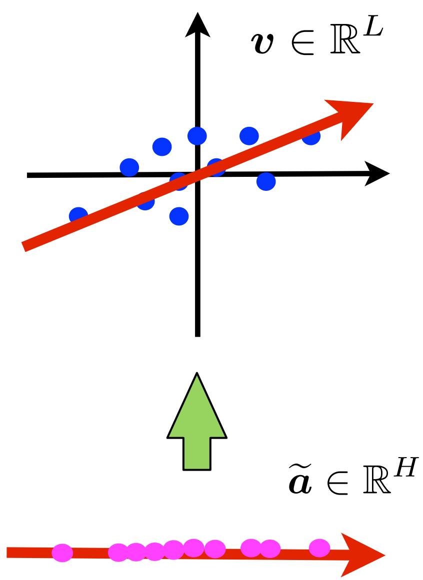
The plausible number of clusters is found.

Matrix factorization

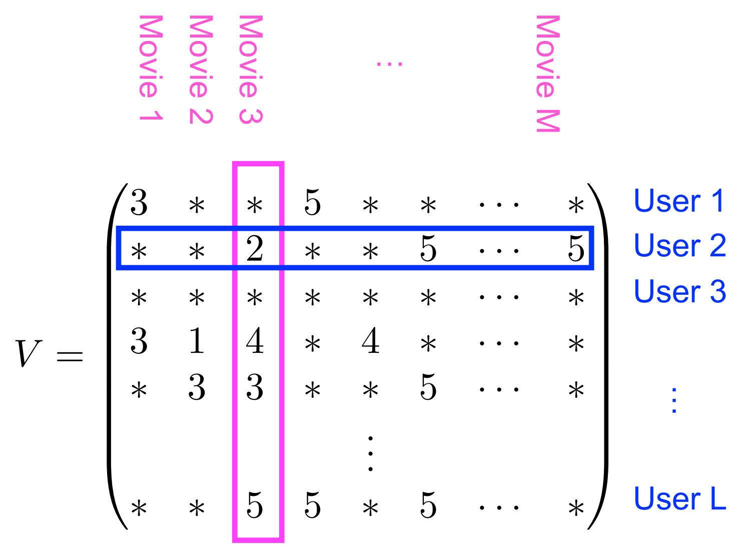
$$V = BA^\top + \mathcal{E}$$



(Probabilistic) PCA

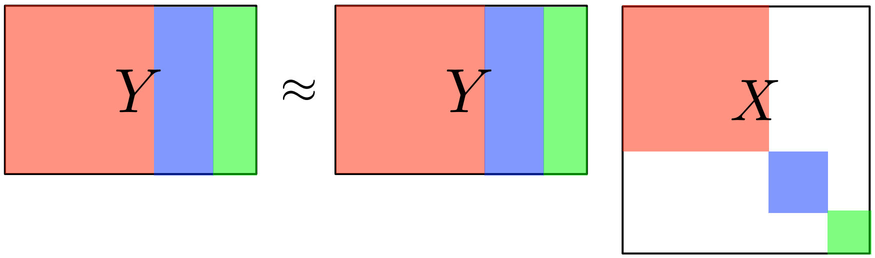


Collaborative filtering



The plausible rank (PCA-dimension) is found

Subspace clustering



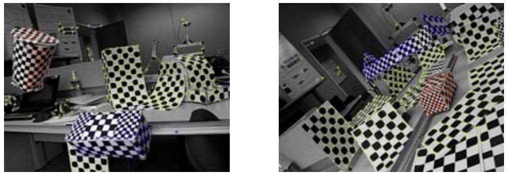
Use Y for dictionary (i.e., $D = Y$):

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

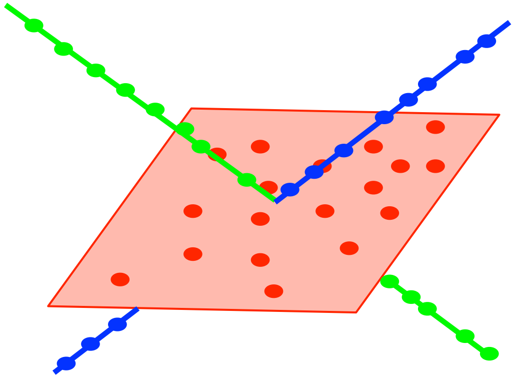
Estimate X, given Y:

$$\min_{X} \|Y - YX\|_{\text{Fro}}^2 + \underline{\lambda\|X\|_{\text{tr}}}.$$

low-rankness inducing penalty



(a) 1R2RCT_B (b) 2T3RCRT (c) cars3 (d) cars10



Spectral clustering with affinity matrix

$$\text{abs}(X) + \text{abs}(X^\top)$$

gives clustering result.

Subspaces with plausible dimensionality is found.

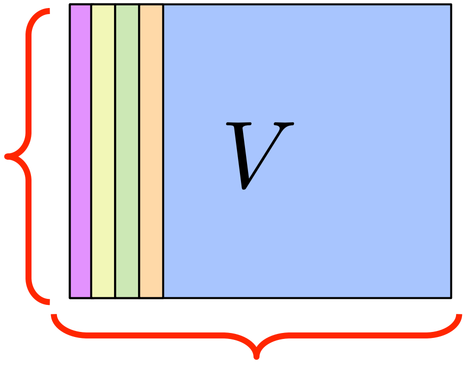
Foreground/Background video separation

$$V = U^{\mathrm{BG}} + U^{\mathrm{FG}} + \mathcal{E}$$

Impose different types of sparsity on U^{BG} and U^{FG}

Robust PCA

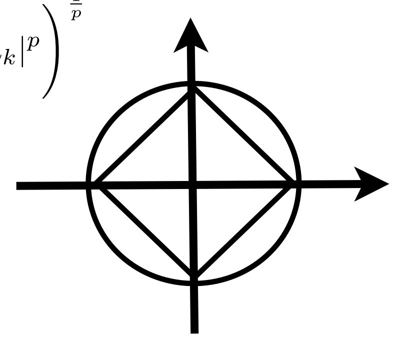
$$V = U^{\text{low-rank}} + U^{\text{element-wise}} + \mathcal{E}$$



FB/BG separation is made without manual tuning parameter.

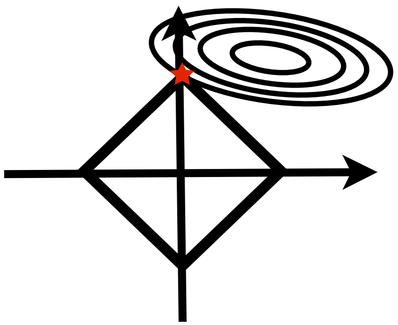
Sparse estimation

$$\|\boldsymbol{u}\|_p = \left(\sum_{k=1}^{K} |u_k|^p \right)^{\frac{1}{p}}$$



- ℓ_1 regularization

$$L(\boldsymbol{x}) = \|\boldsymbol{y} - A\boldsymbol{x}\|^2 + \lambda \|\boldsymbol{x}\|_1$$



  - Convex
  - λ should be tuned.

- Bayesian with automatic relevance determination
  - non-convex (local solver, sparser solution)
  - no hand-tuning parameters (including kernel parameters in GP)

Recent development

- Metropolis Hastings is slow...
    - Hamiltonian Monte Carlo.
- VB approximation can be crude...
    - Theoretical support.
    - Expectation propagation.
- Slow in non-conjugate cases
    - Various variational methods (e.g., proximal gradient).
- Big data
    - stochastic gradient.
    - distributed computation.

Technische Universität Berlin

berlin