Technische Universität Berlin

Lecture: Bayesian Learning

Shinichi Nakajima
Berlin Big Data Center
Technische Universität Berlin

nakajima@tu-berlin.de
http://sites.google.com/site/shinnkj23/

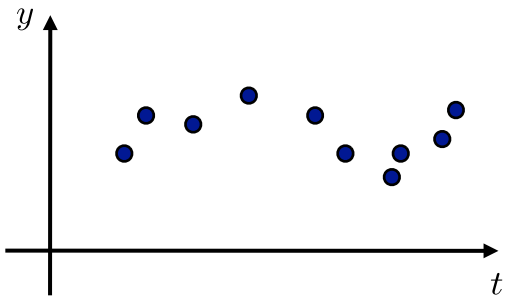
General Information

- Lectures: Thursdays 14-16 (22.10, 29.10, 5.11, 7.1, 14.1, 21.1)
- Homework (5.11 - 7.1, 21.1 - 18.2)
- 3ETC, elective course in Machine Learning I
  (computer science M.Sc.)

What is Bayesian learning?

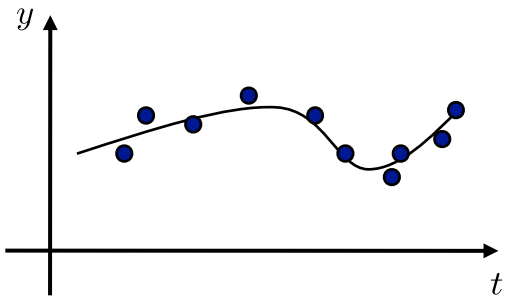
Non-Bayesian (frequentist) approach

A typical machine learning problem



A typical machine learning problem

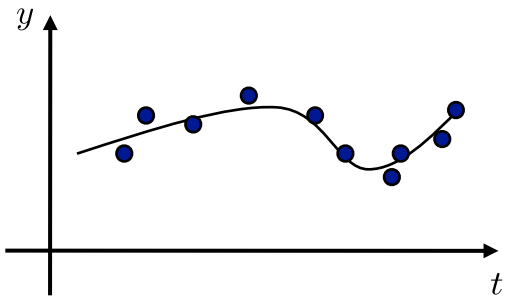
probabilisitic model: $y = f(t) + \varepsilon$



Assumption: y is the sum of a deterministic function of t and a random noise.

A typical machine learning problem

probabilisitic model: $y = f(t) + \varepsilon$



Assumption: y is the sum of a deterministic function of t and a random noise.

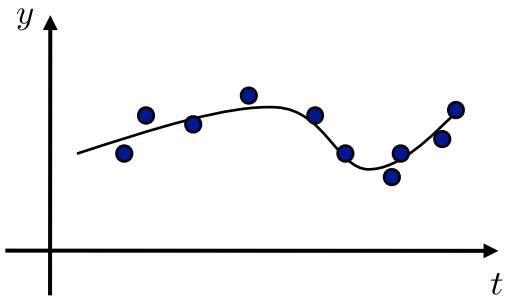
$$\boldsymbol{a} = (a_1, \ldots, a_M)^\top \in \mathbb{R}^M \qquad \boldsymbol{x} = (1, t, \ldots, t^{M-1})^\top \in \mathbb{R}^M$$

linear regression model: $y = \boldsymbol{a}^\top \boldsymbol{x} + \varepsilon,$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2) \quad \Rightarrow \quad p(\varepsilon|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

A typical machine learning problem

probabilisitic model: $y = f(t) + \varepsilon$



Assumption: y is the sum of a deterministic function of t and a random noise.

$$\boldsymbol{a} = (a_1, \ldots, a_M)^\top \in \mathbb{R}^M \qquad \boldsymbol{x} = (1, t, \ldots, t^{M-1})^\top \in \mathbb{R}^M$$

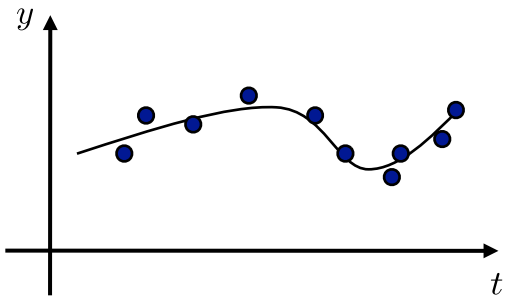
linear regression model: $y = \boldsymbol{a}^\top \boldsymbol{x} + \varepsilon,$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2) \quad \Rightarrow \quad p(\varepsilon|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

model distribution: $p(y|\boldsymbol{x}, \boldsymbol{a}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \boldsymbol{a}^\top \boldsymbol{x})^2}{2\sigma^2}\right)$

A typical machine learning problem

probabilisitic model: $y = f(t) + \varepsilon$



Assumption: y is the sum of a deterministic function of t and a random noise.

$$\boldsymbol{a} = (a_1, \ldots, a_M)^\top \in \mathbb{R}^M \qquad \boldsymbol{x} = (1, t, \ldots, t^{M-1})^\top \in \mathbb{R}^M$$

linear regression model: $y = \boldsymbol{a}^\top \boldsymbol{x} + \varepsilon,$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2) \quad \Rightarrow \quad p(\varepsilon|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

model distribution: $$p(y|\boldsymbol{x}, \boldsymbol{a}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \boldsymbol{a}^\top \boldsymbol{x})^2}{2\sigma^2}\right)$$

For N i.i.d. samples $\mathcal{D} = \{y^{(n)}, \boldsymbol{x}^{(n)}\}_{n=1}^N$

$$p(\mathcal{D}|\boldsymbol{a}, \sigma^2) = \prod_{n=1}^{N} p(y^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{a}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{n=1}^{N} \frac{(y^{(n)} - \boldsymbol{a}^\top \boldsymbol{x}^{(n)})^2}{2\sigma^2}\right)$$

Non-Bayesian approach

Likelihood: $$p(\mathcal{D}|\boldsymbol{a}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{n=1}^{N} \frac{(y^{(n)} - \boldsymbol{a}^\top \boldsymbol{x}^{(n)})^2}{2\sigma^2}\right)$$

Likelihood principle: a model giving higher probability is likely to be the true model.

Maximum likelihood (ML) estimator (least squares curve fitting): $$(\widehat{\boldsymbol{a}}, \widehat{\sigma}^2) = \underset{\boldsymbol{a}, \sigma^2}{\operatorname{argmin}} \left(-\log p(\mathcal{D}|\boldsymbol{a}, \sigma^2)\right)$$

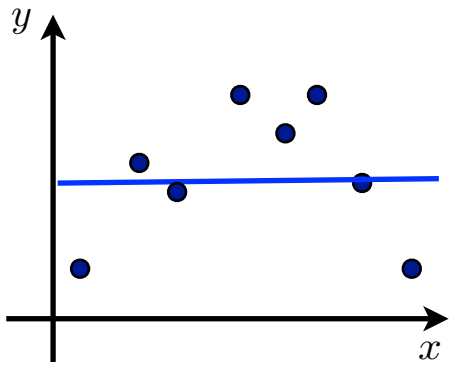
Non-Bayesian approach

Likelihood: $$p(\mathcal{D}|\boldsymbol{a}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{n=1}^{N} \frac{(y^{(n)} - \boldsymbol{a}^\top \boldsymbol{x}^{(n)})^2}{2\sigma^2}\right)$$

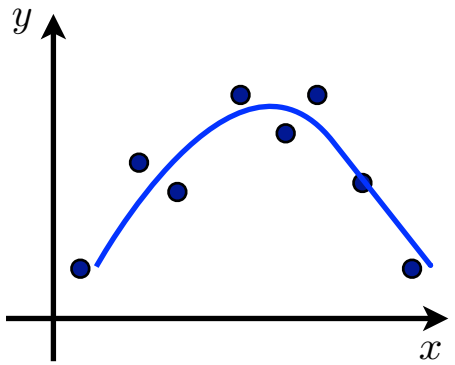
Likelihood principle: a model giving higher probability is likely to be the true model.

Maximum likelihood (ML) estimator (least squares curve fitting): $$(\widehat{\boldsymbol{a}}, \widehat{\sigma}^2) = \underset{\boldsymbol{a}, \sigma^2}{\operatorname{argmin}} \left(-\log p(\mathcal{D}|\boldsymbol{a}, \sigma^2)\right)$$

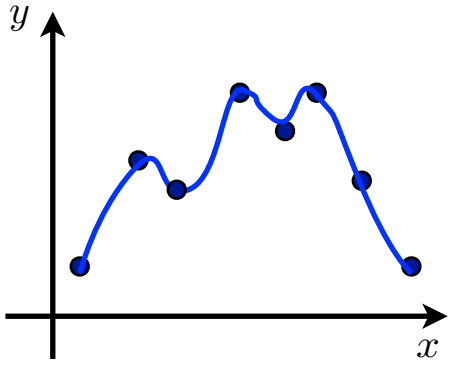
But it can overfit...



$y = a_1 x + a_0$



$y = a_2 x^2 + a_1 x + a_0$



$y = a_7 x^7 + a_6 x^6 + \cdots + a_0$

Non-Bayesian approach

Likelihood: $$p(\mathcal{D}|\boldsymbol{a},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{n=1}^{N} \frac{(y^{(n)} - \boldsymbol{a}^\top \boldsymbol{x}^{(n)})^2}{2\sigma^2}\right)$$

Likelihood principle: a model giving higher probability is likely to be the true model.

Maximum likelihood (ML) estimator (least squares curve fitting): $$(\widehat{\boldsymbol{a}}, \widehat{\sigma}^2) = \operatorname*{argmin}_{\boldsymbol{a},\sigma^2} \left(-\log p(\mathcal{D}|\boldsymbol{a},\sigma^2)\right)$$

But it can overfit...

Penalized ML estimator (ridge regression): $$(\widehat{\boldsymbol{a}}, \widehat{\sigma}^2) = \operatorname*{argmin}_{\boldsymbol{a},\sigma^2} \left(-\log p(\mathcal{D}|\boldsymbol{a},\sigma^2) \underbrace{+\lambda\|\boldsymbol{a}\|_2^2}_{\text{regularizer}}\right)$$

Typically, non-Bayesian method minimizes data fidelity + regularizer

$$\mathcal{L}(\boldsymbol{w}) = F(\mathcal{D};\boldsymbol{w}) + R(\boldsymbol{w}) \qquad \boldsymbol{w} = (\boldsymbol{a},\sigma^2)$$

Bayesian approach

Bayesian learning

Likelihood: $$p(\mathcal{D}|\boldsymbol{a}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{n=1}^{N} \frac{(y^{(n)} - \boldsymbol{a}^\top \boldsymbol{x}^{(n)})^2}{2\sigma^2}\right)$$

Also starts from the likelihood function. Additionally, prior distribution is assumed

Prior: $$p(\boldsymbol{a}) = \frac{1}{(2\pi c^2)^{M/2}} \exp\left(-\frac{\|\boldsymbol{a}\|^2}{2c^2}\right), \qquad p(\sigma^2) \propto 1$$

Bayesian learning

Likelihood: $$p(\mathcal{D}|\boldsymbol{a},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{n=1}^{N} \frac{(y^{(n)} - \boldsymbol{a}^\top \boldsymbol{x}^{(n)})^2}{2\sigma^2}\right)$$

Also starts from the likelihood function. Additionally, prior distribution is assumed

Prior: $$p(\boldsymbol{a}) = \frac{1}{(2\pi c^2)^{M/2}} \exp\left(-\frac{\|\boldsymbol{a}\|^2}{2c^2}\right), \qquad p(\sigma^2) \propto 1$$

and compute the posterior (derived from Bayes Theorem)

Posterior: $$p(\boldsymbol{a},\sigma^2|\mathcal{D}) = \frac{p(\mathcal{D})|\boldsymbol{a},\sigma^2)p(\boldsymbol{a})p(\sigma^2)}{p(\mathcal{D})}$$

where $$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{a},\sigma^2)p(\boldsymbol{a})p(\sigma^2)d\boldsymbol{a}d\sigma^2$$

Bayesian learning computes the distribution of parameters!

Bayesian learning

$$p(\mathcal{D}|\boldsymbol{a},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{n=1}^{N} \frac{(y^{(n)} - \boldsymbol{a}^\top \boldsymbol{x}^{(n)})^2}{2\sigma^2}\right)$$

$$p(\boldsymbol{a}) = \frac{1}{(2\pi c^2)^{M/2}} \exp\left(-\frac{\|\boldsymbol{a}\|^2}{2c^2}\right), \qquad p(\sigma^2) \propto 1$$

Posterior: $p(\boldsymbol{a},\sigma^2|\mathcal{D}) = \dfrac{p(\mathcal{D})|\boldsymbol{a},\sigma^2)p(\boldsymbol{a})p(\sigma^2)}{p(\mathcal{D})}$

where $p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{a},\sigma^2)p(\boldsymbol{a})p(\sigma^2)d\boldsymbol{a}d\sigma^2$

p(D) is constant (w.r.t. unknowns). → shape of posterior is easy to know.

$$p(\boldsymbol{a},\sigma^2|\mathcal{D}) \propto p(\mathcal{D})|\boldsymbol{a},\sigma^2)p(\boldsymbol{a})p(\sigma^2)$$

Bayesian learning

$$p(\mathcal{D}|\boldsymbol{a},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_{n=1}^{N} \frac{(y^{(n)} - \boldsymbol{a}^\top \boldsymbol{x}^{(n)})^2}{2\sigma^2}\right)$$

$$p(\boldsymbol{a}) = \frac{1}{(2\pi c^2)^{M/2}} \exp\left(-\frac{\|\boldsymbol{a}\|^2}{2c^2}\right), \qquad p(\sigma^2) \propto 1$$

Posterior: $p(\boldsymbol{a},\sigma^2|\mathcal{D}) = \dfrac{p(\mathcal{D})|\boldsymbol{a},\sigma^2)p(\boldsymbol{a})p(\sigma^2)}{p(\mathcal{D})}$

where $p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{a},\sigma^2)p(\boldsymbol{a})p(\sigma^2)d\boldsymbol{a}d\sigma^2$

p(D) is constant (w.r.t. unknowns). → shape of posterior is easy to know.

$$p(\boldsymbol{a},\sigma^2|\mathcal{D}) \propto p(\mathcal{D})|\boldsymbol{a},\sigma^2)p(\boldsymbol{a})p(\sigma^2)$$

Maximum a posteriori (MAP) estimation coincides with non-Bayesian method.

$$(\widehat{\boldsymbol{a}},\widehat{\sigma}^2) = \underset{\boldsymbol{a},\sigma^2}{\operatorname{argmin}} \left(-\log p(\mathcal{D}|\boldsymbol{a},\sigma^2) \underbrace{+c^{-2}\|\boldsymbol{a}\|_2^2}_{\text{prior}} \right)$$

MAP is equivalent to non-Bayesian method with appropriate regularizer.

Historically,

Non-Bayesian:

- didn't like the idea of distribution on the parameter (or model). For each problem, w is unknown but fixed (not a random variable).

Bayesian:

- Bayesian interpret the distribution as a "belief".
  - prior distribution: our belief before observation.
  - posterior distribution: our belief after observation.

Nowadays, most people accept the idea of "belief", and know the equivalence between NB and MAP.

Nowadays

Non-Bayesian:

- try to find a point estimates.

Bayesian:

- try to compute the posterior distribution of unknowns as faithful as possible. Typically perform integral computation.

MAP is categorized as non-Bayesian approach.

Bayesian learning

Pros:

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

Bayesian learning

Pros:

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

Posterior mean

Posterior covariance/predictive

Marginal likelihood

Cons:

- Integral computation is required.

Bayesian learning requires integration

Marginal likelihood: $$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

Posterior mean (Bayes estimator): $$\widehat{\boldsymbol{w}} = \int \boldsymbol{w} \cdot p(\boldsymbol{w}|\mathcal{D})d\boldsymbol{w} = \frac{1}{p(\mathcal{D})} \int \boldsymbol{w} \cdot p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

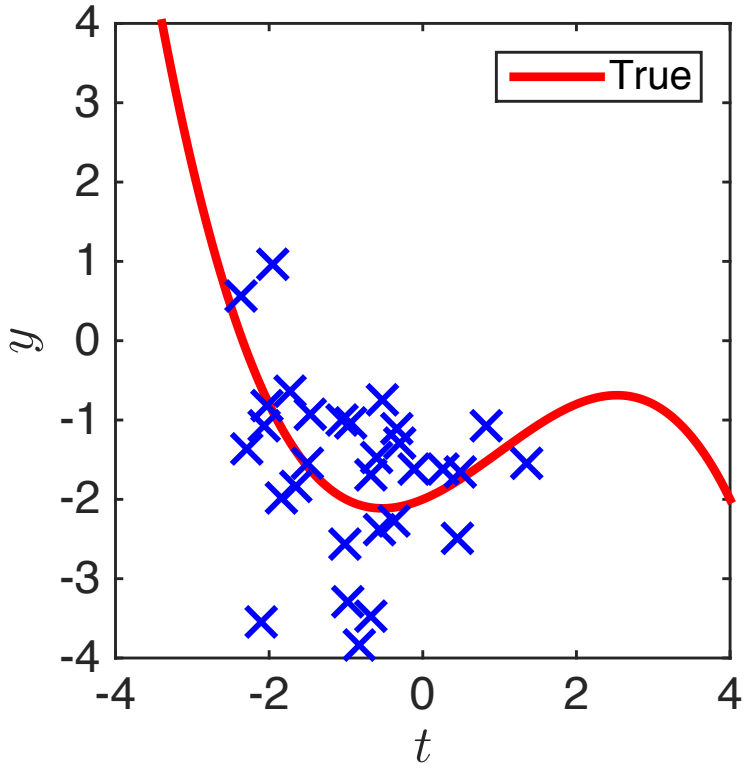
Posterior covariance: $$\widehat{\boldsymbol{\Sigma}}_w = \int \boldsymbol{w}\boldsymbol{w}^\top \cdot p(\boldsymbol{w}|\mathcal{D})d\boldsymbol{w} = \frac{1}{p(\mathcal{D})} \int \boldsymbol{w}\boldsymbol{w}^\top \cdot p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

Predictive distribution:

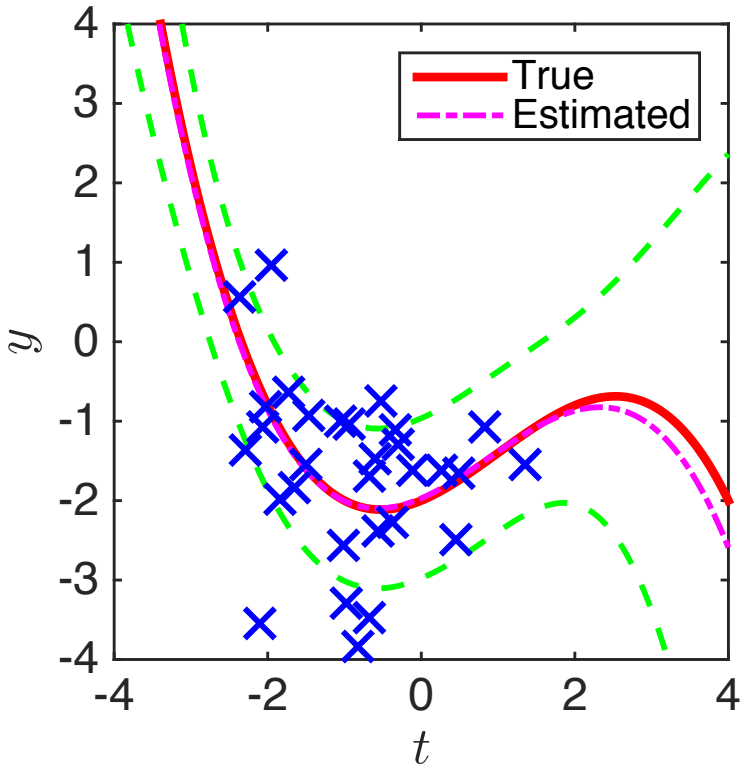
$$p(\mathcal{D}^{\text{new}}|\mathcal{D}) = \int p(\mathcal{D}^{\text{new}}|\boldsymbol{w})p(\boldsymbol{w}|\mathcal{D})d\boldsymbol{w} = \frac{1}{p(\mathcal{D})} \int p(\mathcal{D}^{\text{new}}|\boldsymbol{w}) \cdot p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

Uncertainty assessment

linear regression model



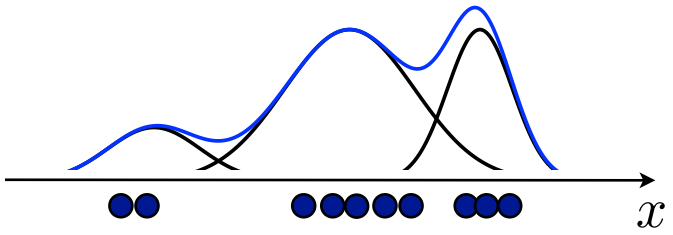
Predictive distribution gives distribution of y*



$$p(y^*|t^*, \mathcal{D})$$

Model selection

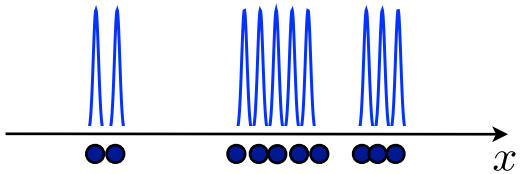
Clustering



Mixture models:

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

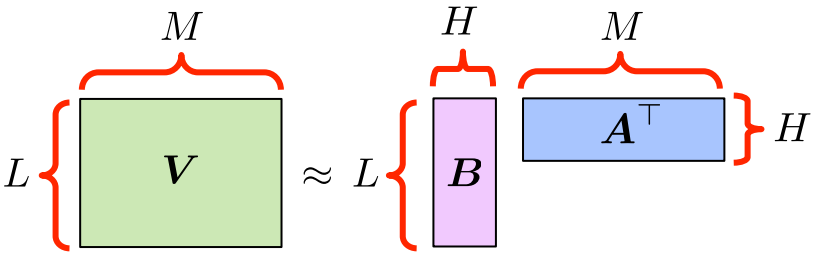
Maximum likelihood estimation results in



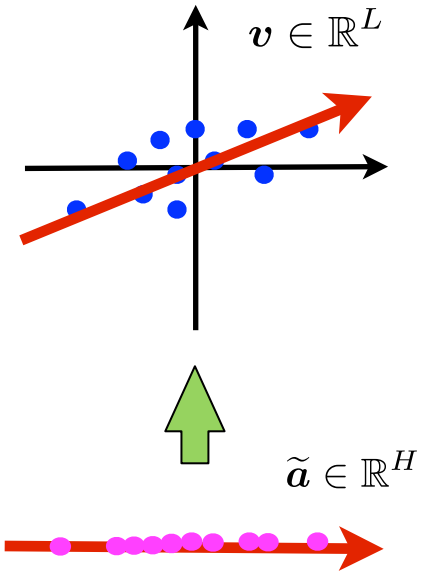
The plausible number of clusters is found.

Matrix factorization

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



(Probabilistic) PCA



Collaborative filtering

$$V = \begin{pmatrix} 3 & * & * & 5 & * & * & \cdots & * \\ * & * & 2 & * & * & 5 & \cdots & 5 \\ * & * & * & * & * & * & \cdots & * \\ 3 & 1 & 4 & * & 4 & * & \cdots & * \\ * & 3 & 3 & * & * & 5 & \cdots & * \\ & & & & \vdots & & & \\ * & * & 5 & 5 & * & 5 & \cdots & * \end{pmatrix}$$

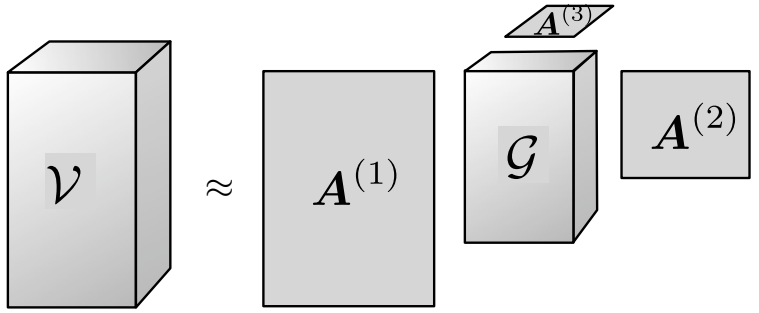
Movie 1, Movie 2, Movie 3, ⋮, Movie M

User 1, User 2, User 3, ⋮, User L

The plausible rank (PCA-dimension) is found

Tensor factorization

$$\mathcal{V} = \mathcal{G} \times_1 \boldsymbol{A}^{(1)} \cdots \times_N \boldsymbol{A}^{(N)} + \mathcal{E}$$



Likelihood: $$p(\mathcal{V}|\mathcal{G}, \{\boldsymbol{A}^{(n)}\}) \propto \exp\left(-\frac{\left\|\mathcal{V} - \mathcal{G} \times_1 \boldsymbol{A}^{(1)} \cdots \times_N \boldsymbol{A}^{(N)}\right\|^2}{2\sigma^2}\right),$$

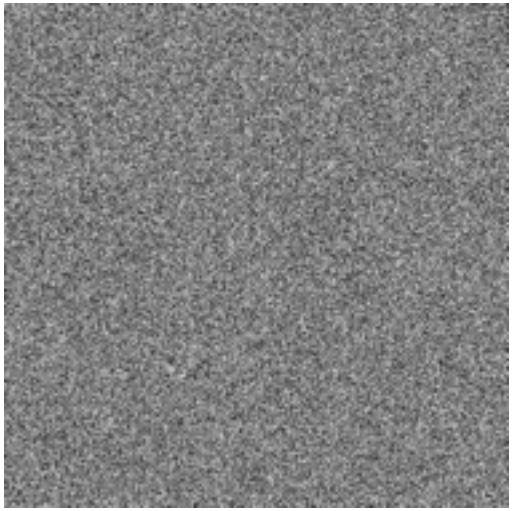
Priors: $$p(\mathcal{G}) \propto \exp\left(-\frac{\mathrm{vec}(\mathcal{G})^\top \left(C_{G^{(N)}} \otimes \cdots \otimes C_{G^{(1)}}\right)^{-1} \mathrm{vec}(\mathcal{G})}{2}\right),$$

$$p(\{A^{(n)}\}) \propto \exp\left(-\frac{\sum_{n=1}^{N} \mathrm{tr}(A^{(n)} C_{A^{(n)}}^{-1} A^{(n)\top})}{2}\right).$$

The plausible tensor ranks are found.

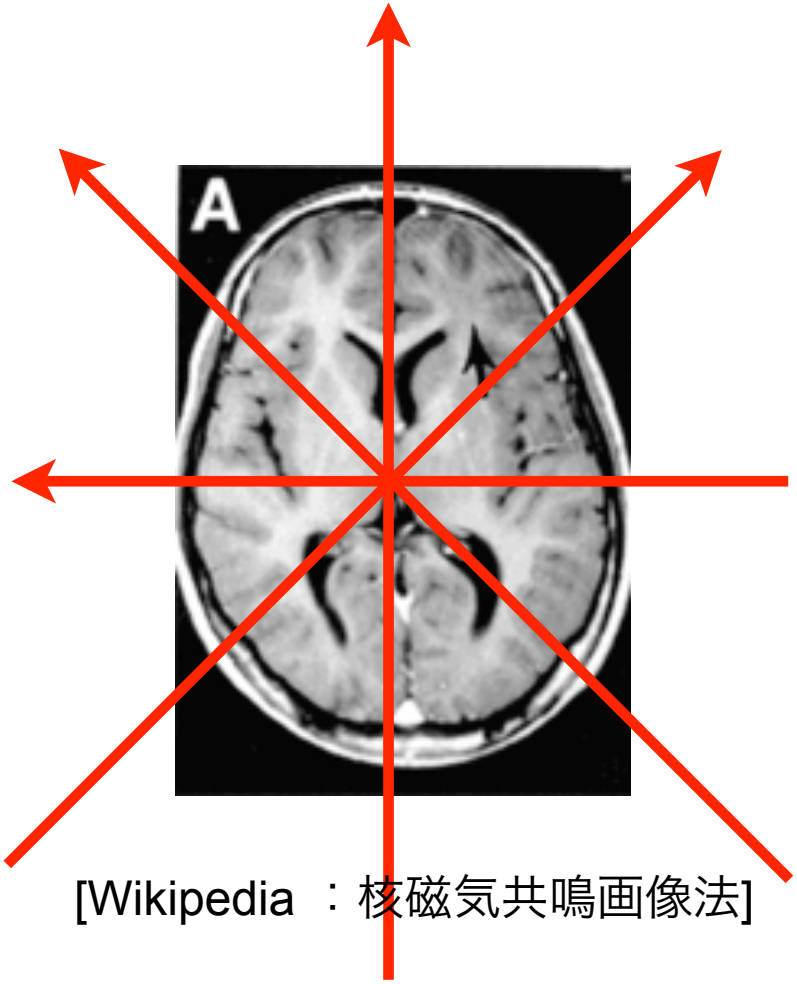
Sparse estimation

Our world is sparse...



# of possible images <<<<<< 256^n

Magnetic Resonance Imaging (MRI)



[Wikipedia ：核磁気共鳴画像法]

Compressed sensing

Natural images are

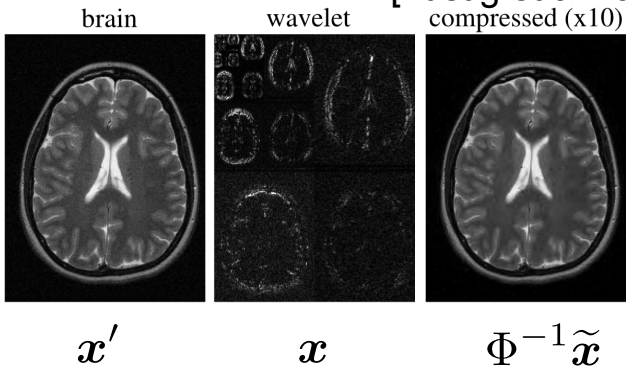
JPEG Compressible → Sparse in wavelet space.

$$\boldsymbol{y} = A'\boldsymbol{x}' + \boldsymbol{\varepsilon}$$
$$= \underbrace{A'\Phi^{-1}}_{A}\underbrace{\Phi\boldsymbol{x}'}_{\boldsymbol{x}} + \boldsymbol{\varepsilon}$$

Φ : wavelet transform

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

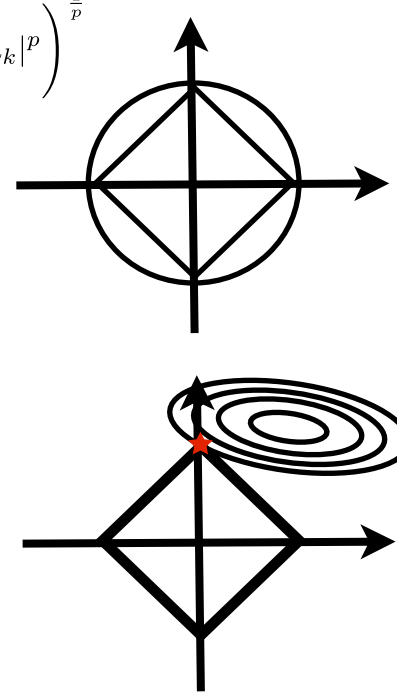
[Lustig et al.2008]



Sparse signal is reconstructed 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

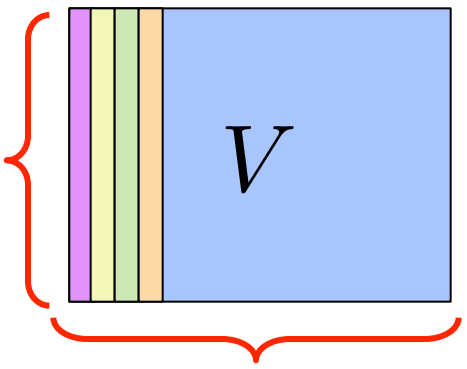
Foreground/Background video separation

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

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

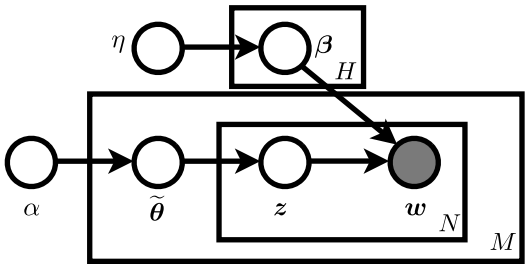
Robust PCA

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



FB/BG separation is made without manual tuning parameter.

Latent Dirichlet allocation



The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Document parameter: $\boldsymbol{\Theta} \in [0,1]^{M \times H}$

Topic parameter: $\boldsymbol{B} \in [0,1]^{L \times H}$

Arts
Budgets
Children
Education

Likelihood: $p(\boldsymbol{w}^{(n,m)}|\boldsymbol{\Theta}, \boldsymbol{B}) = \prod_{l=1}^{L} \left((\boldsymbol{B}\boldsymbol{\Theta}^{\top})_{l,m} \right)^{w_l^{(n,m)}}$

Prior: $p(\boldsymbol{\Theta}|\alpha) \propto \prod_{m=1}^{M} \prod_{h=1}^{H} (\Theta_{m,h})^{\alpha-1}$ $\quad p(\boldsymbol{B}|\eta) \propto \prod_{h=1}^{H} \prod_{l=1}^{L} (B_{l,h})^{\eta-1}$

Overfitting can be avoided, while p-LSA (ML-estimation) suffers from overfitting).

Goal of the lecture

Pros:

- Understand the basic idea of Bayesian learning
- Get able to compute Bayesian learning for tractable cases
- Get able to approximate Bayesian learning for intractable cases.

Tips for calculation

Popular distributions

Isotropic Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$$

Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gamma: $$p(x|\alpha, \beta) = \frac{\beta^\rho}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

Wishart: $$p(\boldsymbol{X}|\boldsymbol{V}, \nu) = \frac{1}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)$$

Bernoulli: $$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

binomial: $$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Multinomail: $$p(\boldsymbol{x}|\boldsymbol{\theta}) = n!\prod_{k=1}^{K}\frac{\theta_k^{x_k}}{x_k!}$$

Beta: $$p(\theta|a, b) = \frac{1}{\mathcal{B}(a, b)}\theta^{a-1}(1-\theta)^{b-1}$$

Dirichlet: $$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^{K}\theta_k^{\phi_k-1}$$

Popular distributions

Isotropic Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$$

Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gamma: $$p(x|\alpha, \beta) = \frac{\beta^\rho}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

Wishart: $$p(\boldsymbol{X}|\boldsymbol{V}, \nu) = \frac{1}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)$$

Bernoulli: $$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

binomial: $$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Multinomail: $$p(\boldsymbol{x}|\boldsymbol{\theta}) = n!\prod_{k=1}^{K}\frac{\theta_k^{x_k}}{x_k!}$$

Beta: $$p(\theta|a, b) = \frac{1}{\mathcal{B}(a, b)}\theta^{a-1}(1-\theta)^{b-1}$$

Dirichlet: $$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^{K}\theta_k^{\phi_k-1}$$

Looks complicated?

Popular distributions

Isotropic Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$$

Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gamma: $$p(x|\alpha, \beta) = \frac{\beta^\rho}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

Wishart: $$p(\boldsymbol{X}|\boldsymbol{V}, \nu) = \frac{1}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)$$

Bernoulli: $$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

binomial: $$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Multinomail: $$p(\boldsymbol{x}|\boldsymbol{\theta}) = n!\prod_{k=1}^{K}\frac{\theta_k^{x_k}}{x_k!}$$

Beta: $$p(\theta|a, b) = \frac{1}{\mathcal{B}(a, b)}\theta^{a-1}(1-\theta)^{b-1}$$

Dirichlet: $$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^{K}\theta_k^{\phi_k-1}$$

Looks complicated?

Mainly because of normalization factors.
You can often neglect them!

Popular distributions

Isotropic Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$$

Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gamma: $$p(x|\alpha, \beta) = \frac{\beta^\rho}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

Wishart: $$p(\boldsymbol{X}|\boldsymbol{V}, \nu) = \frac{1}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)$$

Bernoulli: $$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

binomial: $$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Multinomail: $$p(\boldsymbol{x}|\boldsymbol{\theta}) = n!\prod_{k=1}^{K}\frac{\theta_k^{x_k}}{x_k!}$$

Beta: $$p(\theta|a, b) = \frac{1}{\mathcal{B}(a, b)}\theta^{a-1}(1-\theta)^{b-1}$$

Dirichlet: $$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^{K}\theta_k^{\phi_k-1}$$

Too many distributions?

Popular distributions

Isotropic Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$$

Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gamma: $$p(x|\alpha, \beta) = \frac{\beta^\rho}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

Wishart: $$p(\boldsymbol{X}|\boldsymbol{V}, \nu) = \frac{1}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)$$

Bernoulli: $$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

binomial: $$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Multinomail: $$p(\boldsymbol{x}|\boldsymbol{\theta}) = n!\prod_{k=1}^{K}\frac{\theta_k^{x_k}}{x_k!}$$

Beta: $$p(\theta|a, b) = \frac{1}{\mathcal{B}(a, b)}\theta^{a-1}(1-\theta)^{b-1}$$

Dirichlet: $$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^{K}\theta_k^{\phi_k-1}$$

Too many distributions?

No only 4 types.

Popular distributions

Isotropic Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$$

Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gamma: $$p(x|\alpha, \beta) = \frac{\beta^\rho}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

Wishart: $$p(\boldsymbol{X}|\boldsymbol{V}, \nu) = \frac{1}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)$$

Bernoulli: $$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

binomial: $$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Multinomail: $$p(\boldsymbol{x}|\boldsymbol{\theta}) = n!\prod_{k=1}^{K}\frac{\theta_k^{x_k}}{x_k!}$$

Beta: $$p(\theta|a, b) = \frac{1}{\mathcal{B}(a, b)}\theta^{a-1}(1-\theta)^{b-1}$$

Dirichlet: $$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^{K}\theta_k^{\phi_k-1}$$

Too many distributions?

Even only 2 pairs!

Popular distributions

Isotropic Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$$

Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gamma: $$p(x|\alpha, \beta) = \frac{\beta^\rho}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

Wishart: $$p(\boldsymbol{X}|\boldsymbol{V}, \nu) = \frac{1}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)$$

Bernoulli: $$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

binomial: $$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Multinomail: $$p(\boldsymbol{x}|\boldsymbol{\theta}) = n!\prod_{k=1}^{K}\frac{\theta_k^{x_k}}{x_k!}$$

Beta: $$p(\theta|a, b) = \frac{1}{\mathcal{B}(a, b)}\theta^{a-1}(1-\theta)^{b-1}$$

Dirichlet: $$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^{K}\theta_k^{\phi_k-1}$$

Do I have to integrate black part?

Popular distributions

Isotropic Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$$

Gauss: $$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gamma: $$p(x|\alpha, \beta) = \frac{\beta^\rho}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

Wishart: $$p(\boldsymbol{X}|\boldsymbol{V}, \nu) = \frac{1}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)$$

Bernoulli: $$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

binomial: $$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Multinomail: $$p(\boldsymbol{x}|\boldsymbol{\theta}) = n!\prod_{k=1}^{K}\frac{\theta_k^{x_k}}{x_k!}$$

Beta: $$p(\theta|a, b) = \frac{1}{\mathcal{B}(a, b)}\theta^{a-1}(1-\theta)^{b-1}$$

Dirichlet: $$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^{K}\theta_k^{\phi_k-1}$$

Do I have to integrate black part?

Yes, but the value is equal to the inverse of the blue part.

Domain of random variable and parameters

$p(\boldsymbol{x}\|\boldsymbol{w})$	$\boldsymbol{x} \in \mathcal{X}$	$\boldsymbol{w} \in \mathcal{W}$
$\frac{\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)}{(2\pi\sigma^2)^{M/2}}$	$\boldsymbol{x} \in \mathbb{R}^M$	$\boldsymbol{\mu} \in \mathbb{R}^M, \sigma^2 > 0$
$\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{(2\pi)^{M/2}\|\boldsymbol{\Sigma}\|^{1/2}}$	$\boldsymbol{x} \in \mathbb{R}^M$	$\boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\Sigma} \in \mathbb{S}^M_{++}$
$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$	$x > 0$	$\alpha > 0, \beta > 0$
$\frac{\|\boldsymbol{X}\|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)}{(2^\nu\|\boldsymbol{V}\|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}$	$\boldsymbol{X} \in \mathbb{S}^M_{++}$	$\boldsymbol{V} \in \mathbb{S}^M_{++}, \nu > M-1$
$\theta^x(1-\theta)^{1-x}$	$x \in \{0,1\}$	$\theta \in [0,1]$
$\binom{N}{x}\theta^x(1-\theta)^{N-x}$	$x \in \{0,\ldots,N\}$	$\theta \in [0,1]$
$N!\prod_{k=1}^K (x_k!)^{-1}\theta_k^{x_k}$	$\boldsymbol{x} \in \mathbb{H}_N^{K-1}$	$\boldsymbol{\theta} \in \Delta^{K-1}$
$\frac{1}{\mathcal{B}(a,b)}x^{a-1}(1-x)^{b-1}$	$x \in [0,1]$	$a > 0, b > 0$
$\frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^K x_k^{\phi_{k-1}}$	$\boldsymbol{x} \in \Delta^{K-1}$	$\boldsymbol{\phi} \in \mathbb{R}^K_{++}$

Moments are known!

$$\mathrm{Norm}_M(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}}$$

$$\mathrm{Gamma}(x;\alpha,\beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$$

$$\mathrm{Wishart}_M(\boldsymbol{X};\boldsymbol{V},\nu) \equiv \frac{|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}$$

$$\mathrm{Multinomial}_{K,N}(\boldsymbol{x};\boldsymbol{\theta}) \equiv N!\prod_{k=1}^{K}(x_k!)^{-1}\theta_k^{x_k}$$

$$\mathrm{Dirichlet}_K(\boldsymbol{x};\boldsymbol{\phi}) \equiv \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)}\prod_{k=1}^{K} x_k^{\phi_{k-1}}$$

$p(\boldsymbol{x}\|\boldsymbol{w})$	1st order moment	2nd order moment
$\mathrm{Norm}_M(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$	$\mathbf{Mean}(\boldsymbol{x}) = \boldsymbol{\mu}$	$\mathbf{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$
$\mathrm{Gamma}(x;\alpha,\beta)$	$\mathrm{Mean}(x) = \frac{\alpha}{\beta}$, $\mathrm{Mean}(\log x) = \Psi(\alpha) - \log\beta$,	$\mathrm{Var}(x) = \frac{\alpha}{\beta^2}$, $\mathrm{Var}(\log x) = \Psi_1(\alpha)$
$\mathrm{Wishart}_M(\boldsymbol{X};\boldsymbol{V},\nu)$	$\mathbf{Mean}(\boldsymbol{X}) = \nu\boldsymbol{V}$	$\mathrm{Var}(X_{m,m'}) = \nu(V_{m,m'}^2 + V_{m,m}V_{m',m'})$
$\mathrm{Multinomial}_{K,N}(\boldsymbol{x};\boldsymbol{\theta})$	$\mathbf{Mean}(\boldsymbol{x}) = N\boldsymbol{\theta}$	$(\mathbf{Cov}(\boldsymbol{x}))_{k,k'} = \begin{cases} N\theta_k(1-\theta_k) & (k=k') \\ -N\theta_k\theta_{k'} & (k\neq k') \end{cases}$
$\mathrm{Dirichlet}_K(\boldsymbol{x};\boldsymbol{\phi})$	$\mathbf{Mean}(\boldsymbol{x}) = \frac{1}{\sum_{k=1}^K \phi_k}\boldsymbol{\phi}$ $\mathbf{Mean}(\log x_k) = \Psi(\phi_k) - \Psi(\sum_{k'=1}^K \phi_{k'})$	$(\mathbf{Cov}(\boldsymbol{x}))_{k,k'} = \begin{cases} \frac{\phi_k(\tau-\phi_k)}{\tau^2(\tau+1)} & (k=k') \\ -\frac{\phi_k\phi_{k'}}{\tau^2(\tau+1)} & (k\neq k') \end{cases}$ ただし $\tau = \sum_{k=1}^K \phi_k$

Let’s try!


Technische
Universität
Berlin
berlin