Review of Statistics

Guillaume Obozinski

Ecole des Ponts - ParisTech

Master MVA 2016-2017

Outline

Outline

Statistical concepts

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p_\theta(x) \mid \theta \in \Theta \right\}$$

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p_\theta(x) \mid \theta \in \Theta \right\}$$

Bernoulli model: $X \sim \text{Ber}(\theta)$ $\quad \Theta = [0, 1]$

$$p_\theta(x) = \theta^x (1 - \theta)^{(1-x)}$$

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p_\theta(x) \mid \theta \in \Theta \right\}$$

Bernoulli model: $X \sim \text{Ber}(\theta) \qquad \Theta = [0, 1]$

$$p_\theta(x) = \theta^x (1 - \theta)^{(1-x)}$$

Binomial model: $X \sim \text{Bin}(n, \theta) \qquad \Theta = [0, 1]$

$$p_\theta(x) = \binom{n}{x} \theta^x (1 - \theta)^{(1-x)}$$

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p_\theta(x) \mid \theta \in \Theta \right\}$$

Bernoulli model: $X \sim \mathrm{Ber}(\theta) \qquad \Theta = [0, 1]$

$$p_\theta(x) = \theta^x (1-\theta)^{(1-x)}$$

Binomial model: $X \sim \mathrm{Bin}(n, \theta) \qquad \Theta = [0, 1]$

$$p_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{(1-x)}$$

Multinomial model: $X \sim \mathcal{M}(n, \pi_1, \pi_2, \ldots, \pi_K) \qquad \Theta = [0, 1]^K$

$$p_\theta(x) = \binom{n}{x_1, \ldots, x_k} {\pi_1}^{x_1} \ldots {\pi_k}^{x_k}$$

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \ldots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \dots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \dots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\mathbf{y} = (0, 0, 0, 1, 0)^\top$.

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \ldots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\mathbf{y} = (0, 0, 0, 1, 0)^\top$.
So $\mathbf{y} \in \{0, 1\}^K$ with $\sum_{k=1}^K y_k = 1$.

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \dots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \dots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\mathbf{y} = (0, 0, 0, 1, 0)^\top$.
So $\mathbf{y} \in \{0, 1\}^K$ with $\sum_{k=1}^K y_k = 1$.

$$\mathbb{P}(C = k) = \mathbb{P}(Y_k = 1) \quad \text{and} \quad \mathbb{P}(Y = y) = \prod_{k=1}^K \pi_k^{y_k}.$$

Bernoulli, Binomial, Multinomial

$Y \sim \text{Ber}(\pi)$	$(Y_1, \ldots, Y_K) \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$
$p(y) = \pi^y (1-\pi)^{1-y}$	$p(\mathbf{y}) = \pi_1^{y_1} \ldots \pi_K^{y_K}$
$N_1 \sim \text{Bin}(n, \pi)$	$(N_1, \ldots, N_K) \sim \mathcal{M}(n, \pi_1, \ldots, \pi_K)$
$p(n_1) = \binom{n}{n_1} \pi^{n_1} (1-\pi)^{n-n_1}$	$p(\mathbf{n}) = \begin{pmatrix} & n & \\ n_1 & \ldots & n_K \end{pmatrix} \pi_1^{n_1} \ldots \pi_K^{n_K}$

with

$$\binom{n}{i} = \frac{n!}{(n-i)!i!} \quad \text{and} \quad \begin{pmatrix} & n & \\ n_1 & \ldots & n_K \end{pmatrix} = \frac{n!}{n_1! \ldots n_K!}$$

Gaussian model

Scalar Gaussian model : $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$

X real valued r.v., and $\theta = \left(\mu, \sigma^2\right) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Gaussian model

Scalar Gaussian model : $X \sim \mathcal{N}(\mu, \sigma^2)$

X real valued r.v., and $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

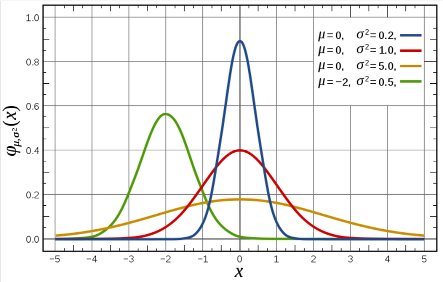
$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Multivariate Gaussian model: $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

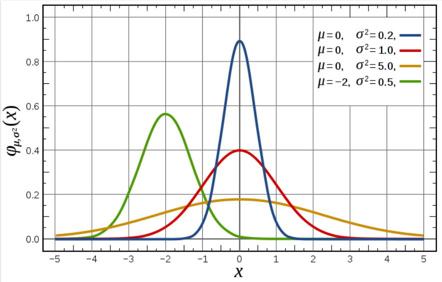
X r.v. taking values in $\mathbb{R}^d$. If $\mathcal{K}_d$ is the set of positive definite matrices of size $d \times d$, and $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Theta = \mathbb{R}^d \times \mathcal{K}_d$.

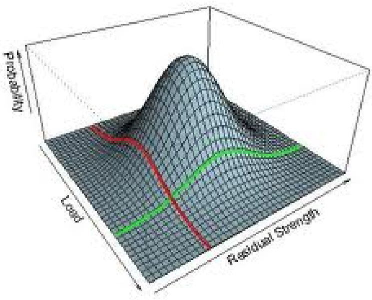
$$p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Gaussian densities



Gaussian densities





Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \dots, X^{(n)}$$

Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \ldots, X^{(n)}$$

A common assumption is that the variables are **i.i.d.**

- **independent**
- **identically distributed**, i.e. have the same distribution P.

Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \ldots, X^{(n)}$$

A common assumption is that the variables are **i.i.d.**

- **independent**
- **identically distributed**, i.e. have the same distribution P.

This collection of observations is called

- the *sample* or the *observations* in statistics
- the *sample***s** in engineering
- the *training set* in machine learning

Outline

The maximum likelihood principle

Maximum likelihood principle

- Let $\mathcal{P}_\Theta = \left\{ p(x;\theta) \mid \theta \in \Theta \right\}$ be a *model*
- Let x be an observation

Maximum likelihood principle

- Let $\mathcal{P}_\Theta = \{p(x;\theta) \mid \theta \in \Theta\}$ be a *model*
- Let x be an observation

Likelihood:

$$\begin{aligned} \mathcal{L} : \Theta &\to \mathbb{R}_+ \\ \theta &\mapsto p(x;\theta) \end{aligned}$$

Maximum likelihood principle

- Let $\mathcal{P}_\Theta = \{p(x;\theta) \mid \theta \in \Theta\}$ be a *model*
- Let x be an observation

Likelihood:

$$\begin{aligned} \mathcal{L} : \Theta &\to \mathbb{R}_+ \\ \theta &\mapsto p(x;\theta) \end{aligned}$$

Maximum likelihood estimator:

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, p(x;\theta)$$

Sir Ronald Fisher
(1890-1962)

Maximum likelihood principle

- Let $\mathcal{P}_\Theta = \{p(x;\theta) \mid \theta \in \Theta\}$ be a *model*
- Let x be an observation

Sir Ronald Fisher
(1890-1962)

Likelihood:

$$\begin{aligned} \mathcal{L} : \Theta &\to \mathbb{R}_+ \\ \theta &\mapsto p(x;\theta) \end{aligned}$$

Maximum likelihood estimator:

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\ p(x;\theta)$$

Case of i.i.d data

If $(x_i)_{1\leq i\leq n}$ is an i.i.d. sample of size n:

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta\in\Theta}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta(x_i) = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log\ p_\theta(x_i)$$

The maximum likelihood estimator

The MLE

- does not always exists

The maximum likelihood estimator

The MLE

- does not always exists
- is not necessarily unique

The maximum likelihood estimator

The MLE

- does not always exists
- is not necessarily unique

MLE for the Bernoulli model

Let $X_1, X_2, \ldots, X_n$ an i.i.d. sample $\sim \text{Ber}(\theta)$.

MLE for the Bernoulli model

Let $X_1, X_2, \ldots, X_n$ an i.i.d. sample $\sim \mathrm{Ber}(\theta)$. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i;\, \theta)$$

MLE for the Bernoulli model

Let $X_1, X_2, \ldots, X_n$ an i.i.d. sample $\sim \text{Ber}(\theta)$. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i;\, \theta) = \sum_{i=1}^{n} \log \left[\theta^{x_i}(1-\theta)^{1-x_i}\right]$$

MLE for the Bernoulli model

Let $X_1, X_2, \dots, X_n$ an i.i.d. sample $\sim \text{Ber}(\theta)$. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i;\, \theta) = \sum_{i=1}^{n} \log \left[\theta^{x_i}(1-\theta)^{1-x_i}\right]$$

$$= \sum_{i=1}^{n} \big(x_i \log\theta + (1-x_i)\log(1-\theta)\big)$$

MLE for the Bernoulli model

Let $X_1, X_2, \ldots, X_n$ an i.i.d. sample $\sim \text{Ber}(\theta)$. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i;\, \theta) = \sum_{i=1}^{n} \log \left[\theta^{x_i}(1-\theta)^{1-x_i}\right]$$
$$= \sum_{i=1}^{n} \big(x_i \log\theta + (1-x_i)\log(1-\theta)\big) = N\log(\theta) + (n-N)\log(1-\theta)$$

with $N := \sum_{i=1}^{n} x_i$.

MLE for the Bernoulli model

Let $X_1, X_2, \ldots, X_n$ an i.i.d. sample $\sim \text{Ber}(\theta)$. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i;\, \theta) = \sum_{i=1}^{n} \log \left[\theta^{x_i}(1-\theta)^{1-x_i}\right]$$

$$= \sum_{i=1}^{n} \big(x_i \log\theta + (1-x_i)\log(1-\theta)\big) = N\log(\theta) + (n-N)\log(1-\theta)$$

with $N := \sum_{i=1}^{n} x_i$.

- $\theta \mapsto \ell(\theta)$ is strongly concave $\Rightarrow$ the MLE exists and is unique.

MLE for the Bernoulli model

Let $X_1, X_2, \ldots, X_n$ an i.i.d. sample $\sim \mathrm{Ber}(\theta)$. The log-likelihood is

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \log p(x_i;\, \theta) = \sum_{i=1}^{n} \log\left[\theta^{x_i}(1-\theta)^{1-x_i}\right] \\ &= \sum_{i=1}^{n} \big(x_i \log\theta + (1-x_i)\log(1-\theta)\big) = N\log(\theta) + (n-N)\log(1-\theta)\end{aligned}$$

with $N := \sum_{i=1}^{n} x_i$.

- $\theta \mapsto \ell(\theta)$ is strongly concave $\Rightarrow$ the MLE exists and is unique.
- since ℓ differentiable + strongly concave its maximizer is the unique stationary point

MLE for the Bernoulli model

Let $X_1, X_2, \ldots, X_n$ an i.i.d. sample $\sim \mathrm{Ber}(\theta)$. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i;\, \theta) = \sum_{i=1}^{n} \log \left[\theta^{x_i}(1-\theta)^{1-x_i}\right]$$

$$= \sum_{i=1}^{n} \big(x_i \log\theta + (1-x_i)\log(1-\theta)\big) = N\log(\theta) + (n-N)\log(1-\theta)$$

with $N := \sum_{i=1}^{n} x_i$.

- $\theta \mapsto \ell(\theta)$ is strongly concave $\Rightarrow$ the MLE exists and is unique.
- since ℓ differentiable + strongly concave its maximizer is the unique stationary point

$$\nabla\ell(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \frac{N}{\theta} - \frac{n-N}{1-\theta}.$$

MLE for the Bernoulli model

Let $X_1, X_2, \ldots, X_n$ an i.i.d. sample $\sim \text{Ber}(\theta)$. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i;\, \theta) = \sum_{i=1}^{n} \log\left[\theta^{x_i}(1-\theta)^{1-x_i}\right]$$
$$= \sum_{i=1}^{n} \big(x_i \log\theta + (1-x_i)\log(1-\theta)\big) = N\log(\theta) + (n-N)\log(1-\theta)$$

with $N := \sum_{i=1}^{n} x_i$.

- $\theta \mapsto \ell(\theta)$ is strongly concave $\Rightarrow$ the MLE exists and is unique.
- since ℓ differentiable + strongly concave its maximizer is the unique stationary point

$$\nabla\ell(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \frac{N}{\theta} - \frac{n-N}{1-\theta}.$$

Thus

$$\hat{\theta}_{\mathrm{ML}} = \frac{N}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

MLE for the multinomial

Done on the board.

Outline

Method of moments (Karl Pearson, 1894)

Consider a statistical model for a *univariate* r.v. parameterized by

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^k.$$

Denote by μ^k the kth moment of a random variable:

$$\mu_1(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[X], \quad \mu_2(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[X^2], \quad \ldots, \quad \mu_K(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[X^K].$$

We have

$$(\mu_1, \ldots, \mu_K) = f(\boldsymbol{\theta}) = f(\theta_1, \ldots, \theta_K).$$

Principle of the method of moments

Given a sample $X_1, \ldots, X_n$

- Estimate the μ_ks with the empirical moments: $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$.
- The moment estimator is $\hat{\boldsymbol{\theta}}$ defined as the solution to the equation

$$(\hat{\mu}_1, \ldots, \hat{\mu}_K) = f(\hat{\theta}_1, \ldots, \hat{\theta}_K).$$

Method of moments: illustration

In many usual cases the *moment estimator* and the *MLE* are equal.

Method of moments: illustration

In many usual cases the *moment estimator* and the *MLE* are equal.

Example where MME $\neq$ MLE

For the family of gamma distribution

$$p(x;\lambda,p) = \frac{x^{p-1}e^{-\lambda x}}{\lambda^p\,\Gamma(p)}1_{\{x>0\}}$$

the MLE is not closed-form (exercise).

Method of moments: illustration

In many usual cases the *moment estimator* and the *MLE* are equal.

Example where MME $\neq$ MLE

For the family of gamma distribution

$$p(x;\lambda,p) = \frac{x^{p-1}e^{-\lambda x}}{\lambda^p\,\Gamma(p)}1_{\{x>0\}}$$

the MLE is not closed-form (exercise). However

$$\mu_1 = \mathbb{E}[X] = \lambda p, \qquad \mu_2 = \mathbb{E}[X^2] = p(p+1)\lambda^2,$$

Method of moments: illustration

In many usual cases the *moment estimator* and the *MLE* are equal.

Example where MME $\neq$ MLE

For the family of gamma distribution

$$p(x;\lambda,p) = \frac{x^{p-1}e^{-\lambda x}}{\lambda^p\,\Gamma(p)}1_{\{x>0\}}$$

the MLE is not closed-form (exercise). However

$\mu_1 = \mathbb{E}[X] = \lambda p, \qquad \mu_2 = \mathbb{E}[X^2] = p(p+1)\lambda^2$, So that

$$\lambda = \frac{\mu_1^2}{\mu_2-\mu_1^2}, \qquad p = \frac{\mu_2-\mu_1^2}{\mu_1},$$

which yields the moment estimators

$$\hat{\lambda} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2-\hat{\mu}_1^2}, \qquad p = \frac{\hat{\mu}_2-\hat{\mu}_1^2}{\hat{\mu}_1}.$$

Outline

Linear regression

Design matrix

Consider a finite collection of vectors $x_i \in \mathbb{R}^d$ pour $i = 1 \ldots n$.

Design Matrix

$$X = \begin{bmatrix} \text{---} & x_1^\top & \text{---} \\ \vdots & \vdots & \vdots \\ \text{---} & x_n^\top & \text{---} \end{bmatrix}.$$

Design matrix

Consider a finite collection of vectors $x_i \in \mathbb{R}^d$ pour $i = 1 \ldots n$.

Design Matrix

$$X = \begin{bmatrix} \text{---} & x_1^\top & \text{---} \\ \vdots & \vdots & \vdots \\ \text{---} & x_n^\top & \text{---} \end{bmatrix}.$$

We assume that the vectors are centered, i.e. that $\sum_{i=1}^n x_i = 0$.

Design matrix

Consider a finite collection of vectors $x_i \in \mathbb{R}^d$ pour $i = 1 \dots n$.

Design Matrix

$$X = \begin{bmatrix} \text{---} & x_1^\top & \text{---} \\ \vdots & \vdots & \vdots \\ \text{---} & x_n^\top & \text{---} \end{bmatrix}.$$

We assume that the vectors are centered, i.e. that $\sum_{i=1}^n x_i = 0$.

If x_i are not centered the design matrix of centered data can be constructed with the rows $x_i - \bar{x}^\top$ with $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Linear regression

- We consider the OLS regression for the linear hypothesis space.
- We have $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}$ and ℓ the square loss.

Linear regression

- We consider the OLS regression for the linear hypothesis space.
- We have $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}$ and ℓ the square loss.

Consider the hypothesis space:

$$S = \{f_{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^p\} \qquad \text{with} \qquad f_{\mathbf{w}} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}.$$

Linear regression

- We consider the OLS regression for the linear hypothesis space.
- We have $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}$ and ℓ the square loss.

Consider the hypothesis space:

$$S = \{f_{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^p\} \qquad \text{with} \qquad f_{\mathbf{w}} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}.$$

Given a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ we have

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2$$

Linear regression

- We consider the OLS regression for the linear hypothesis space.
- We have $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}$ and ℓ the square loss.

Consider the hypothesis space:

$$S = \{f_{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^p\} \qquad \text{with} \qquad f_{\mathbf{w}} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}.$$

Given a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ we have

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

with

- the vector of outputs $\mathbf{y}^\top = (y_1, \ldots, y_n) \in \mathbb{R}^n$
- the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ whose ith row is equal to $\mathbf{x}_i^\top$.

Solving linear regression

To solve $\min_{\mathbf{w}\in\mathbb{R}^p} \widehat{\mathcal{R}}_n(f_{\mathbf{w}})$, we consider that

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n}\left(\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\,\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \|\mathbf{y}\|^2\right)$$

is a differentiable convex function whose minima are thus characterized by the

Solving linear regression

To solve $\min_{\mathbf{w}\in\mathbb{R}^p} \widehat{\mathcal{R}}_n(f_{\mathbf{w}})$, we consider that

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n}\left(\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\,\mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \|\mathbf{y}\|^2\right)$$

is a differentiable convex function whose minima are thus characterized by the

Normal equations

$$\boxed{\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{X}^\top\mathbf{y} = \mathbf{0}}$$

Solving linear regression

To solve $\min_{\mathbf{w}\in\mathbb{R}^p} \widehat{\mathcal{R}}_n(f_{\mathbf{w}})$, we consider that

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n}\left(\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\,\mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \|\mathbf{y}\|^2\right)$$

is a differentiable convex function whose minima are thus characterized by the

Normal equations

$$\boxed{\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{X}^\top\mathbf{y} = \mathbf{0}}$$

If $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\widehat{f}$ is given by:

$$\widehat{f} : \mathbf{x}' \mapsto \mathbf{x}'^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Solving linear regression

To solve $\min_{\mathbf{w}\in\mathbb{R}^p} \widehat{\mathcal{R}}_n(f_\mathbf{w})$, we consider that

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n}\left(\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\,\mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \|\mathbf{y}\|^2\right)$$

is a differentiable convex function whose minima are thus characterized by the

Normal equations

$$\boxed{\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{X}^\top\mathbf{y} = \mathbf{0}}$$

If $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\widehat{f}$ is given by:

$$\widehat{f} : \mathbf{x}' \mapsto \mathbf{x}'^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Problem: $\mathbf{X}^\top\mathbf{X}$ is never invertible for $p > n$ and thus the solution is not unique.

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

- Problem now strongly convex thus well-posed

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

- Problem now strongly convex thus well-posed
- Thus with unique solution:

$$\hat{\mathbf{w}}^{(\text{ridge})} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

- Problem now strongly convex thus well-posed
- Thus with unique solution:

$$\hat{\mathbf{w}}^{(\text{ridge})} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

- Shrinkage effect

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_2^2$$

- Problem now strongly convex thus well-posed
- Thus with unique solution:

$$\hat{\mathbf{w}}^{(\text{ridge})}=(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

- Shrinkage effect
- Regularization improves the conditioning number of the Hessian

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

- Problem now strongly convex thus well-posed
- Thus with unique solution:

$$\hat{\mathbf{w}}^{(\text{ridge})} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

- Shrinkage effect
- Regularization improves the conditioning number of the Hessian

$\Rightarrow$ Problem now easier to solve computationally

Outline

Principal Component Analysis (1901)

Karl Pearson (1857 - 1936)

Empirical covariance and correlation

For centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$$

Empirical covariance and correlation

For centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$$

For non centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top$$

Empirical covariance and correlation

For centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$$

For non centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top$$

Another common operation is to normalize the data by dividing each column of X by its standard deviation. This leads to the empirical covariance matrix.

$$C = \mathrm{Diag}(\widehat{\sigma})^{-1} \widehat{\Sigma}\, \mathrm{Diag}(\widehat{\sigma})^{-1} \qquad \text{avec} \quad \widehat{\sigma}_k^2 = \widehat{\Sigma}_{k,k}.$$

Empirical covariance and correlation

For centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$$

For non centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top$$

Another common operation is to normalize the data by dividing each column of X by its standard deviation. This leads to the empirical covariance matrix.

$$C = \mathrm{Diag}(\widehat{\sigma})^{-1} \widehat{\Sigma}\, \mathrm{Diag}(\widehat{\sigma})^{-1} \qquad \text{avec} \quad \widehat{\sigma}_k^2 = \widehat{\Sigma}_{k,k}.$$

$$C_{k,k'} = \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{x_i^{(k)} - \bar{x}^k}{\widehat{\sigma}_k} \Big) \Big(\frac{x_i^{(k')} - \bar{x}^{k'}}{\widehat{\sigma}_{k'}} \Big).$$

Empirical covariance and correlation

For centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$$

For non centered vectors :

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top$$

Another common operation is to normalize the data by dividing each column of X by its standard deviation. This leads to the empirical covariance matrix.

$$C = \mathrm{Diag}(\widehat{\sigma})^{-1} \widehat{\Sigma}\, \mathrm{Diag}(\widehat{\sigma})^{-1} \qquad \text{avec} \quad \widehat{\sigma}_k^2 = \widehat{\Sigma}_{k,k}.$$

$$C_{k,k'} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i^{(k)} - \bar{x}^k}{\widehat{\sigma}_k} \right) \left(\frac{x_i^{(k')} - \bar{x}^{k'}}{\widehat{\sigma}_{k'}} \right).$$

Normalisation is optional...

PCA from the analysis point of view

Data vectors live in $\mathbb{R}^d$ and one seeks a direction v in $\mathbb{R}^d$ such that the variance along this direction is maximal. Or

$$\begin{aligned} Var((v^\top x_i)_{i=1\ldots n}) &= \frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} v^\top x_i x_i^\top v \end{aligned}$$

PCA from the analysis point of view

Data vectors live in $\mathbb{R}^d$ and one seeks a direction v in $\mathbb{R}^d$ such that the variance along this direction is maximal. Or

$$
\begin{aligned}
Var((v^\top x_i)_{i=1\dots n}) &= \frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} v^\top x_i x_i^\top v \\
&= v^\top\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right) v
\end{aligned}
$$

PCA from the analysis point of view

Data vectors live in $\mathbb{R}^d$ and one seeks a direction v in $\mathbb{R}^d$ such that the variance along this direction is maximal. Or

$$\begin{aligned} Var((v^\top x_i)_{i=1\ldots n}) &= \frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} v^\top x_i x_i^\top v \\ &= v^\top \Big(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\Big) v \\ &= v^\top \widehat{\Sigma} v \end{aligned}$$

PCA from the analysis point of view

Data vectors live in $\mathbb{R}^d$ and one seeks a direction v in $\mathbb{R}^d$ such that the variance along this direction is maximal. Or

$$\begin{aligned} Var((v^\top x_i)_{i=1\dots n}) &= \frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} v^\top x_i x_i^\top v \\ &= v^\top \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right) v \\ &= v^\top \widehat{\Sigma} v \end{aligned}$$

PCA from the analysis point of view

Data vectors live in $\mathbb{R}^d$ and one seeks a direction v in $\mathbb{R}^d$ such that the variance along this direction is maximal. Or

$$
\begin{aligned}
Var((v^\top x_i)_{i=1\dots n}) &= \frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} v^\top x_i x_i^\top v \\
&= v^\top \Big(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\Big) v \\
&= v^\top \widehat{\Sigma} v
\end{aligned}
$$

One needs to solve

$$
\max_{\|v\|_2=1} v^\top \widehat{\Sigma} v
$$

PCA from the analysis point of view

Data vectors live in $\mathbb{R}^d$ and one seeks a direction v in $\mathbb{R}^d$ such that the variance along this direction is maximal. Or

$$\begin{aligned} Var((v^\top x_i)_{i=1\dots n}) &= \frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} v^\top x_i x_i^\top v \\ &= v^\top\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)v \\ &= v^\top \widehat{\Sigma} v \end{aligned}$$

One needs to solve

$$\max_{\|v\|_2=1} v^\top \widehat{\Sigma} v$$

Solution: first eigenvectors of $\widehat{\Sigma}$ say v_1.

Deflation

What is the second best direction to project the data on in order to maximize the variance ?

Deflation

What is the second best direction to project the data on in order to maximize the variance ?

One can perform a deflation

$$\forall i, \quad \widetilde{x}_i \leftarrow x_i - v_1(v_1^\top x_i)$$

Deflation

What is the second best direction to project the data on in order to maximize the variance ?

One can perform a deflation

$$\forall i, \quad \widetilde{x}_i \leftarrow x_i - v_1(v_1^\top x_i)$$

Which translates at the matrix level: $\widetilde{X} \leftarrow X - X v_1 v_1^\top.$

Deflation

What is the second best direction to project the data on in order to maximize the variance ?

One can perform a deflation

$$\forall i, \quad \widetilde{x}_i \leftarrow x_i - v_1(v_1^\top x_i)$$

Which translates at the matrix level: $\widetilde{X} \leftarrow X - X v_1 v_1^\top$.

Then again find the direction of maximal variance

$$\widetilde{\widehat{\Sigma}} = \frac{1}{n} \widetilde{X}^\top \widetilde{X}$$

Deflation

What is the second best direction to project the data on in order to maximize the variance ?

One can perform a deflation

$$\forall i, \quad \widetilde{x}_i \leftarrow x_i - v_1(v_1^\top x_i)$$

Which translates at the matrix level: $\widetilde{X} \leftarrow X - Xv_1 v_1^\top.$

Then again find the direction of maximal variance

$$\widetilde{\widehat{\Sigma}} = \frac{1}{n}\widetilde{X}^\top \widetilde{X}$$

One solves $\max_{\|v\|_2} v^\top \widetilde{\widehat{\Sigma}} v$

Deflation

What is the second best direction to project the data on in order to maximize the variance ?

One can perform a deflation

$$\forall i, \quad \widetilde{x}_i \leftarrow x_i - v_1(v_1^\top x_i)$$

Which translates at the matrix level: $\quad \widetilde{X} \leftarrow X - Xv_1 v_1^\top.$

Then again find the direction of maximal variance

$$\widetilde{\widehat{\Sigma}} = \frac{1}{n}\widetilde{X}^\top \widetilde{X}$$

One solves $\quad \max_{\|v\|_2} v^\top \widetilde{\widehat{\Sigma}} v$

Or equivalently $\quad \max_{\|v\|_2} v^\top \widehat{\Sigma} v \quad$ tel que $\quad v \perp v_1.$

Deflation

What is the second best direction to project the data on in order to maximize the variance ?

One can perform a deflation

$$\forall i, \quad \widetilde{x}_i \leftarrow x_i - v_1(v_1^\top x_i)$$

Which translates at the matrix level: $\widetilde{X} \leftarrow X - X v_1 v_1^\top$.

Then again find the direction of maximal variance

$$\widetilde{\widehat{\Sigma}} = \frac{1}{n} \widetilde{X}^\top \widetilde{X}$$

One solves $\max_{\|v\|_2} v^\top \widetilde{\widehat{\Sigma}} v$

Or equivalently $\max_{\|v\|_2} v^\top \widehat{\Sigma} v$ tel que $v \perp v_1$.

Solution: This yields the second eigenvector of $\widehat{\Sigma}$ say v_2. Etc.

Principal directions

We usually call

- **principal directions (or factors)** of the points cloud the vectors

$$v_1, v_2, \dots, v_k.$$

Principal directions

We usually call

- **principal directions (or factors)** of the points cloud the vectors

$$v_1, v_2, \dots, v_k.$$

- **principal components**:
  the projection of the data on the k principal directions.

Principal directions

We usually call

- **principal directions (or factors)** of the points cloud the vectors

$$v_1, v_2, \dots, v_k.$$

- **principal components**:
  the projection of the data on the k principal directions.

The principal directions are the eigenvectors of $\widehat{\Sigma} = V\,S^2\,V^\top$.

Singular value decomposition and PCA

The SVD of a matrix $X \in \mathbb{R}^{n \times p}$ with $n \leq p$ is of the form $X = USV^\top$, avec

- $U \in \mathbb{R}^{n \times n}$ an orthogonal basis of $\mathbb{R}^n$
- $S \in \mathbb{R}^{n \times p}$ a (rectangular) diagonal matrix .
- $V \in \mathbb{R}^{p \times p}$ une base orthogonale de $\mathbb{R}^p$

Singular value decomposition and PCA

The SVD of a matrix $X \in \mathbb{R}^{n \times p}$ with $n \leq p$ is of the form $X = USV^\top$, avec

- $U \in \mathbb{R}^{n \times n}$ an orthogonal basis of $\mathbb{R}^n$
- $S \in \mathbb{R}^{n \times p}$ a (rectangular) diagonal matrix .
- $V \in \mathbb{R}^{p \times p}$ une base orthogonale de $\mathbb{R}^p$

Reduced SVD

The reduced SVD is more often used: If r is the rank of X then $X = USV^\top$ with,

- $U \in \mathbb{R}^{n \times r}$ whose columns are orthonormal.
- $S \in \mathbb{R}^{r \times r}$ a squared diagonal matrix.
- $V \in \mathbb{R}^{r \times p}$ whose columns are orthonormal.

Singular value decomposition and PCA

The SVD of a matrix $X \in \mathbb{R}^{n\times p}$ with $n \leq p$ is of the form $X = USV^\top$, avec

- $U \in \mathbb{R}^{n\times n}$ an orthogonal basis of $\mathbb{R}^n$
- $S \in \mathbb{R}^{n\times p}$ a (rectangular) diagonal matrix .
- $V \in \mathbb{R}^{p\times p}$ une base orthogonale de $\mathbb{R}^p$

Reduced SVD

The reduced SVD is more often used: If r is the rank of X then $X = USV^\top$ with,

- $U \in \mathbb{R}^{n\times r}$ whose columns are orthonormal.
- $S \in \mathbb{R}^{r\times r}$ a squared diagonal matrix.
- $V \in \mathbb{R}^{r\times p}$ whose columns are orthonormal.

If the diagonal of S is such that $s_1 > s_2 > \ldots > s_r > 0$ and $U_{1k} \geq 0$ for all k the reduced SVD is unique. We have that

- $U\,S^2U^\top$ is a (compact) diagonalisation of $XX^\top$
- $V\,S^2V^\top$ is a (compact) diagonalisation of $X^\top X$

Outline

Bayesian estimation

Bayesians treat the parameter θ as a random variable.

A priori

The Bayesian has to specify an *a priori* distribution $p(\theta)$ for the model parameters θ, which models his prior belief of the relative plausibility of different values of the parameter.

Bayesian estimation

Bayesians treat the parameter θ as a random variable.

A priori

The Bayesian has to specify an *a priori* distribution $p(\theta)$ for the model parameters θ, which models his prior belief of the relative plausibility of different values of the parameter.

A posteriori

The observation contribute through the likelihood: $p(x|\theta)$.
The *a posteriori* distribution on the parameters is then

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)} \propto p(x|\theta)\, p(\theta)\,.$$

→ The Bayesian estimator is therefore a probability distribution on the parameters.

This estimation procedure is called Bayesian inference.

Conjugate priors

A family of prior distribution

$$\mathcal{P}_A = \{p_\alpha(\theta) \mid \alpha \in A\}$$

is said to be **conjugate** to a model $\mathcal{P}_\Theta$, if, for a sample

$$X^{(1)}, \ldots, X^{(n)} \overset{\text{i.i.d.}}{\sim} p_\theta \qquad \text{with} \qquad p_\theta \in \mathcal{P}_\Theta,$$

the distribution q defined by

$$q(\theta) = p(\theta | x^{(1)}, \ldots, x^{(n)}) = \frac{p_\alpha(\theta) \prod_i p_\theta\left(x^{(i)}\right)}{\int p_\alpha(\theta) \prod_i p_\theta\left(x^{(i)}\right) d\theta}$$

is such that

$$q \in \mathcal{P}_A.$$

Dirichlet distribution

We say that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

for

Dirichlet distribution

We say that $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

for $\boldsymbol{\theta}$ in the simplex $\triangle_K = \{\mathbf{u} \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\}$ and

Dirichlet distribution

We say that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

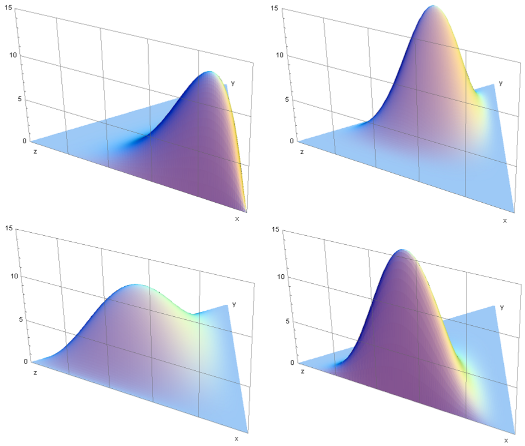
for $\boldsymbol{\theta}$ in the simplex $\triangle_K = \{\mathbf{u} \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\}$ and admitting the density

$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_k \Gamma(\alpha_k)} \, \theta_1^{\alpha_1 - 1} \ldots \theta_K^{\alpha_K - 1}$$

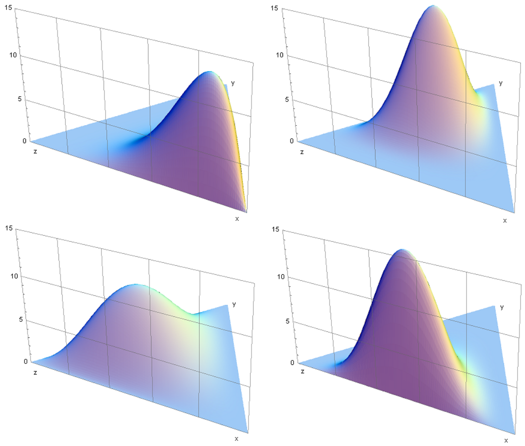
with respect to the uniform measure on the simplex, where

$$\alpha_0 = \sum_k \alpha_k \quad \text{and} \quad \Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt$$

Dirichlet distribution II



Dirichlet distribution II



$$\mathbb{E}[\theta_k] = \frac{\alpha_k}{\alpha_0} \quad , \quad \mathrm{Var}(\theta_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \quad \text{and} \quad \mathrm{Cov}(\theta_j, \theta_k) = \frac{-\alpha_j \alpha_k}{\alpha_0^2(\alpha_0 + 1)}$$

with $\alpha_0 = \sum_k \alpha_k$.

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) =$$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) = \frac{p(\boldsymbol{\theta}) \prod_n p(\mathbf{z}^{(n)}|\theta)}{p(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)})}$$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) = \frac{p(\boldsymbol{\theta}) \prod_n p(\mathbf{z}^{(n)}|\theta)}{p(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)})} \propto \prod_k \theta_k^{\alpha_k + \sum_n z_{nk} - 1}$$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) = \frac{p(\boldsymbol{\theta}) \prod_n p(\mathbf{z}^{(n)}|\theta)}{p(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)})} \quad \propto \quad \prod_k \theta_k^{\alpha_k + \sum_n z_{nk} - 1}$$

So that $(\theta|(Z)) \sim \mathrm{Dir}((\alpha_1 + N_1, \ldots, \alpha_K + N_K))$ with $N_k = \sum_n z_{nk}$

Use of the posterior distribution and posterior mean

The principle of Bayesian estimation is that the prior and posterior distribution model the *uncertainty* that we have in the estimation process. As a consequence, one should always integrate over the uncertainty. So the final estimate for a function $f(\boldsymbol{\theta})$ is

$$\int f(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}.$$

Use of the posterior distribution and posterior mean

The principle of Bayesian estimation is that the prior and posterior distribution model the *uncertainty* that we have in the estimation process. As a consequence, one should always integrate over the uncertainty. So the final estimate for a function $f(\boldsymbol{\theta})$ is

$$\int f(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}.$$

In particular the predictive distribution is

$$p(\mathbf{x}'|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}) = \int p(\mathbf{x}'|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}.$$

Use of the posterior distribution and posterior mean

The principle of Bayesian estimation is that the prior and posterior distribution model the *uncertainty* that we have in the estimation process. As a consequence, one should always integrate over the uncertainty. So the final estimate for a function $f(\boldsymbol{\theta})$ is

$$\int f(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}.$$

In particular the predictive distribution is

$$p(\mathbf{x}'|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}) = \int p(\mathbf{x}'|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}.$$

If a point estimate is needed for $\boldsymbol{\theta}$ then this should be the posterior mean

$$\hat{\boldsymbol{\theta}}_{\mathrm{PM}} = \mathbb{E}\big[\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\big] = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}$$

Maximum A Posteriori estimation

Often, it is too hard or too costly to compute the **posterior mean**

$$\hat{\boldsymbol{\theta}}_{\mathrm{PM}} = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}.$$

Maximum A Posteriori estimation

Often, it is too hard or too costly to compute the **posterior mean**

$$\hat{\boldsymbol{\theta}}_{\mathrm{PM}} = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}.$$

An alternative is to compute the

posterior mode or **maximum a posteriori**:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$$

Maximum A Posteriori estimation

Often, it is too hard or too costly to compute the **posterior mean**

$$\hat{\boldsymbol{\theta}}_{\mathrm{PM}} = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}.$$

An alternative is to compute the

posterior mode or **maximum a posteriori**:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}) \\ &= \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \end{aligned}$$

Maximum A Posteriori estimation

Often, it is too hard or too costly to compute the **posterior mean**

$$\hat{\boldsymbol{\theta}}_{\mathrm{PM}} = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})\, d\boldsymbol{\theta}.$$

An alternative is to compute the

posterior mode or **maximum a posteriori**:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}) \\ &= \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\end{aligned}$$

... corresponds to a form of regularized maximum likelihood.