Review of Statistics

Guillaume Obozinski

Ecole des Ponts - ParisTech

Master MVA

Outline

Outline

Statistical concepts

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Expectation of $f(X)$, pour f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Expectation of $f(X)$, pour f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

- Variance :

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

- Covariance :

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$$

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Expectation of $f(X)$, pour f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

- Variance :

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

- Covariance :

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$$

- Conditional expectation of X given Y :

$$\mathbb{E}[X|Y] = \sum_x x \cdot p(x|y)$$

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Expectation of $f(X)$, pour f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

- Variance :

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

- Covariance :

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$$

- Conditional expectation of X given Y :

$$\mathbb{E}[X|Y] = \sum_x x \cdot p(x|y)$$

- Variance decomposition

Independence concepts

Independence: $X \perp\!\!\!\perp Y$

We will say that X and Y are independent and write $X \perp\!\!\!\perp Y$ iff:

$$\forall x, y, \qquad P(X = x, Y = y) = P(X = x)\, P(Y = y)$$

Independence concepts

Independence: $X \perp\!\!\!\perp Y$

We will say that X and Y are independent and write $X \perp\!\!\!\perp Y$ iff:

$$\forall x, y, \qquad P(X = x, Y = y) = P(X = x)\, P(Y = y)$$

Conditional Independence: $X \perp\!\!\!\perp Y \mid Z$

- We will say that X and Y are independent conditionally on Z and
- write $X \perp\!\!\!\perp Y \mid Z$ ssi:

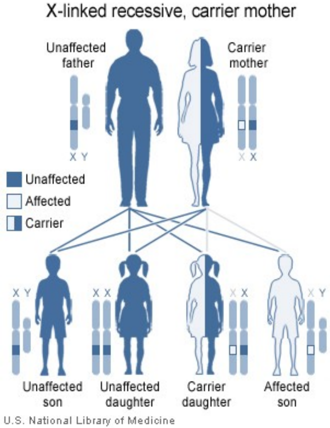
$\forall x, y, z,$

$$P(X = x, Y = y \mid Z = z) = P(X = x|Z = z)\, P(Y = y|Z = z)$$

Conditional Independence exemple

Example of
"X-linked recessive inheritance":

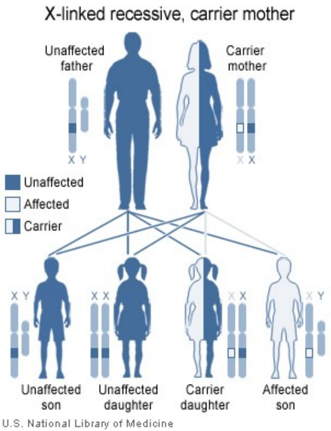
Transmission of the gene
responsible for hemophilia



Conditional Independence exemple

Example of
“X-linked recessive inheritance”:

Transmission of the gene
responsible for hemophilia



Risk for sons from an unaffected father:

- dependance between the situation of the two brothers.
- conditionally independent given that the mother is a carrier of the gene or not.

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

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

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 $\boldsymbol{y} = (0, 0, 0, 1, 0)^\top$.

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 $\boldsymbol{y} = (0, 0, 0, 1, 0)^\top$.
So $\boldsymbol{y} \in \{0, 1\}^K$ 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 $\boldsymbol{y} = (0, 0, 0, 1, 0)^\top$.
So $\boldsymbol{y} \in \{0, 1\}^K$ 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 $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

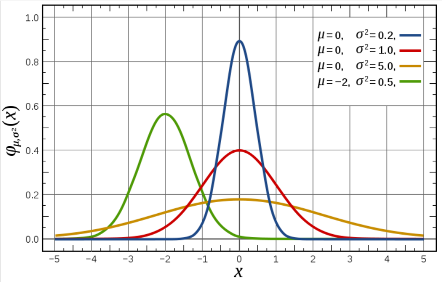
$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

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

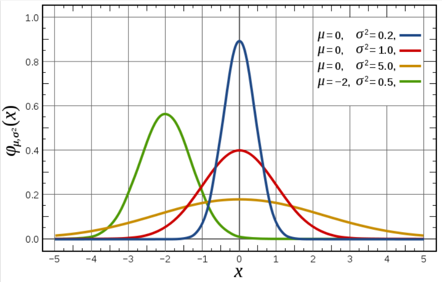
X r.v. taking values in $\mathbb{R}^d$. If $\mathcal{K}_d$ is the set of positive definite matrices of size $d \times d$, and $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Theta = \mathbb{R}^d \times \mathcal{K}_d$.

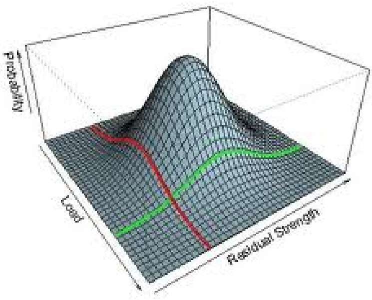
$$p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

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.

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

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

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

Case of i.i.d data

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

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

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

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

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

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 $\sim \mathrm{Ber}(\theta)$. 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 $N := \sum_{i=1}^{n} x_i$.

MLE: Existence? Uniqueness?

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: Existence? Uniqueness?

- $\theta \mapsto \ell(\theta)$ 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

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

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

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

MLE: Existence? Uniqueness?

- $\theta \mapsto \ell(\theta)$ 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 $N := \sum_{i=1}^{n} x_i$.

MLE: Existence? Uniqueness?

- $\theta \mapsto \ell(\theta)$ 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 $N := \sum_{i=1}^{n} x_i$.

MLE: Existence? Uniqueness?

- $\theta \mapsto \ell(\theta)$ 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:

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

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

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

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: $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

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 $i = 1 \dots n$.

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 $\sum_{i=1}^n x_i = 0$.

Design matrix

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

Design Matrix

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

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:

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

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:

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

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

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

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:

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

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

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

with

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

Solving linear regression

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

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

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

Solving linear regression

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

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

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

Normal equations

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

Solving linear regression

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

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

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

Normal equations

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

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

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

Solving linear regression

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

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

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

Normal equations

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

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

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

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

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

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

- Problem now strongly convex thus well-posed

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

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

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

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

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

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

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

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

- Shrinkage effect

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

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

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

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

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

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

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

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

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

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

$\Rightarrow$ Problem now easier to solve computationally

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.

Dirichlet distribution

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

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

for

Dirichlet distribution

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

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

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

Dirichlet distribution

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

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

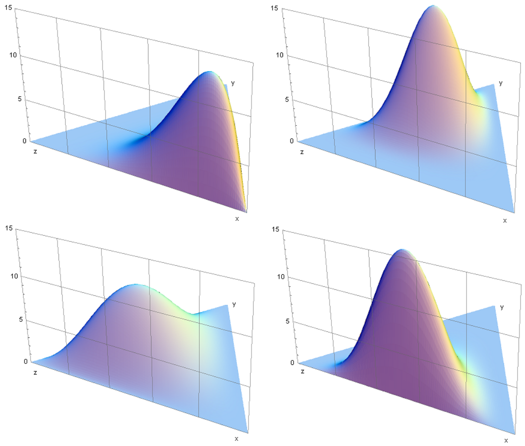
for $\boldsymbol{\theta}$ in the simplex $\triangle_K = \{\boldsymbol{u} \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\}$ and admitting the density

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

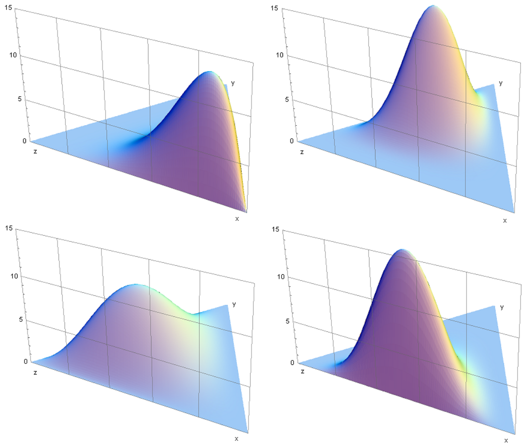
with respect to the uniform measure on the simplex, where

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

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: $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\boldsymbol{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

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 $\boldsymbol{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

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

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

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

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

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

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

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 $\boldsymbol{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

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

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

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

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 $\boldsymbol{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

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

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

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

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

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

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

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

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

So that $(\theta|(Z)) \sim \text{Dir}((\alpha_1 + N_1, \ldots, \alpha_K + N_K))$ 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

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

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

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

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.

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

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

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.

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

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

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.

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

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

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

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

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

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

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

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

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

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

One needs to solve

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

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: $\quad \widetilde{X} \leftarrow X - X v_1 v_1^\top.$

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: $\widetilde{X} \leftarrow X - Xv_1 v_1^\top$.

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: $\quad \widetilde{X} \leftarrow X - X v_1 v_1^\top.$

Then again find the direction of maximal variance

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

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

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: $\quad \widetilde{X} \leftarrow X - X v_1 v_1^\top.$

Then again find the direction of maximal variance

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

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

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

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

$$v_1, v_2, \ldots, v_k.$$

Principal directions

We usually call

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

$$v_1, v_2, \ldots, v_k.$$

- **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

$$v_1, v_2, \ldots, v_k.$$

- **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

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

Eckart-Young theorem

Let $X = USV^\top$ be the SVD of X.
Let

- $U_{[k]} \in \mathbb{R}^{n \times k}$ the matrix formed by the k first columns of U
- $V_{[k]} \in \mathbb{R}^{p \times k}$ the matrix formed by the k first columns of V
- $S_{[k]} \in \mathbb{R}^{k \times k}$ the diagonal matrix with the k first singular values in S

The solution of

$$\min_Z \|X - Z\|_F^2 \quad \text{s.t.} \quad \mathrm{rank}(Z) \leq k$$

is

$$Z = X_{[k]} \qquad \text{with} \quad X_{[k]} := U_{[k]} S_{[k]} V_{[k]}^\top.$$

Two different views of PCA

Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Two different views of PCA

Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Analysis view

Find projection $v \in \mathbb{R}^p$ maximizing variance:

$$\begin{aligned} \max_{v \in \mathbb{R}^p} \quad & v^\top X^\top X\, v \\ \text{s.t.} \quad & \|v\|_2 \leq 1 \end{aligned}$$

$\rightarrow$ deflate and iterate to obtain more components.

Two different views of PCA

Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Analysis view

Find projection $v \in \mathbb{R}^p$ maximizing variance:

$$\begin{aligned} \max_{v \in \mathbb{R}^p} \quad & v^\top X^\top X\, v \\ \text{s.t.} \quad & \|v\|_2 \leq 1 \end{aligned}$$

$\rightarrow$ deflate and iterate to obtain more components.

Synthesis view

Find $V = [v_1, \ldots, v_k]$ s.t. x_i have low reconstruction error on span(V):

$$\min_{u_i, v_i \in \mathbb{R}^p} \|X - \sum_{i=1}^{k} u_i v_i^\top\|_F^2$$

Two different views of PCA

Given data matrix $X = (x_1^\top, \dots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Analysis view

Find projection $v \in \mathbb{R}^p$ maximizing variance:

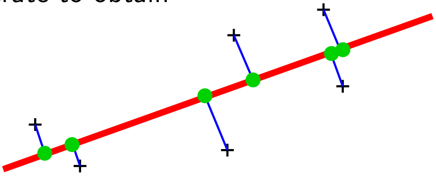
$$\begin{aligned} \max_{v \in \mathbb{R}^p} \quad & v^\top X^\top X\, v \\ \text{s.t.} \quad & \|v\|_2 \leq 1 \end{aligned}$$

$\rightarrow$ deflate and iterate to obtain more components.

Synthesis view

Find $V = [v_1, \dots, v_k]$ s.t. x_i have low reconstruction error on span(V):

$$\min_{u_i, v_i \in \mathbb{R}^p} \|X - \sum_{i=1}^{k} u_i v_i^\top\|_F^2$$



Two different views of PCA

Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Analysis view

Find projection $v \in \mathbb{R}^p$ maximizing variance:

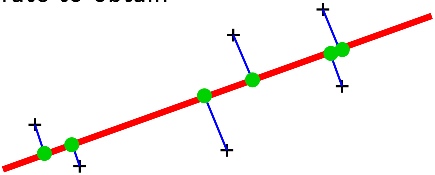
$$\begin{array}{ll} \max_{v \in \mathbb{R}^p} & v^\top X^\top X v \\ \text{s.t.} & \|v\|_2 \leq 1 \end{array}$$

$\rightarrow$ deflate and iterate to obtain more components.

Synthesis view

Find $V = [v_1, \ldots, v_k]$ s.t. x_i have low reconstruction error on span(V):

$$\min_{u_i, v_i \in \mathbb{R}^p} \|X - \sum_{i=1}^{k} u_i v_i^\top\|_F^2$$



- For regular PCA, the two views are equivalent!

Outline

1. Statistical concepts
2. The maximum likelihood principle
3. Method of moments
4. Linear regression
5. Bayesian Inference
6. Principal Component Analysis
7. Mesures of performance for binary classifiers

Sensitivity, precision and co.

		Predicted	
Actual		"P"	"N"
	P	TP	FN
	N	FP	TN

- **sensitivity**, true positive rate or **recall** $r_{TP} = \frac{|TP|}{|P|}$

Sensitivity, precision and co.

		Predicted	
		"P"	"N"
Actual	P	TP	FN
	N	FP	TN

- **sensitivity**, true positive rate or **recall** $r_{TP} = \frac{|TP|}{|P|}$
- **specificity** or true negative rate $r_{TN} = \frac{|TN|}{|N|}$

Sensitivity, precision and co.

	Predicted	
Actual	"P"	"N"
P	TP	FN
N	FP	TN

- **sensitivity**, true positive rate or **recall** $r_{TP} = \frac{|TP|}{|P|}$
- **specificity** or true negative rate $r_{TN} = \frac{|TN|}{|N|}$
- **false positive rate** (type I error) $\alpha = r_{FP} = \frac{|FP|}{|N|} = 1 - r_{TN}$

Sensitivity, precision and co.

	Predicted	
Actual	**"P"**	**"N"**
P	**TP**	**FN**
N	**FP**	**TN**

- **sensitivity**, true positive rate or **recall** $r_{TP} = \frac{|TP|}{|P|}$
- **specificity** or true negative rate $r_{TN} = \frac{|TN|}{|N|}$
- **false positive rate** (type I error) $\alpha = r_{FP} = \frac{|FP|}{|N|} = 1 - r_{TN}$
- **false negative rate** (type II error) $\beta = r_{FN} = \frac{|FN|}{|P|} = 1 - r_{TP}$

Sensitivity, precision and co.

	Predicted	
Actual	**"P"**	**"N"**
P	**TP**	**FN**
N	**FP**	**TN**

- **sensitivity**, true positive rate or **recall** $r_{TP} = \frac{|TP|}{|P|}$
- **specificity** or true negative rate $r_{TN} = \frac{|TN|}{|N|}$
- **false positive rate** (type I error) $\alpha = r_{FP} = \frac{|FP|}{|N|} = 1 - r_{TN}$
- **false negative rate** (type II error) $\beta = r_{FN} = \frac{|FN|}{|P|} = 1 - r_{TP}$
- **precision** $prec = \frac{|TP|}{|FP|+|TP|}$

Sensitivity, precision and co.

	Predicted	
Actual	"P"	"N"
P	TP	FN
N	FP	TN

- **sensitivity**, true positive rate or **recall** $r_{TP} = \frac{|TP|}{|P|}$
- **specificity** or true negative rate $r_{TN} = \frac{|TN|}{|N|}$
- **false positive rate** (type I error) $\alpha = r_{FP} = \frac{|FP|}{|N|} = 1 - r_{TN}$
- **false negative rate** (type II error) $\beta = r_{FN} = \frac{|FN|}{|P|} = 1 - r_{TP}$
- **precision** $prec = \frac{|TP|}{|FP|+|TP|}$

Sensitivity, specificity, etc

	P	N
$\widehat{P}$	Sensitivity (TPR) $\frac{TP}{P}$	FPR $\frac{FP}{N}$
$\widehat{N}$	FNR $\frac{FN}{P}$	Specificity (TNR) $\frac{TN}{N}$

TPR True Positive Rate

FPR False Positive Rate

FNR False Negative Rate

NPR True Negative Rate

Precision, FDR, etc

	P	N
$\widehat{P}$	Precision (PPV) $\frac{TP}{\widehat{P}}$	FDR $\frac{FP}{\widehat{P}}$
$\widehat{N}$	FOR $\frac{FN}{\widehat{N}}$	NPV $\frac{TN}{\widehat{N}}$

PPV Positive Predictive Value

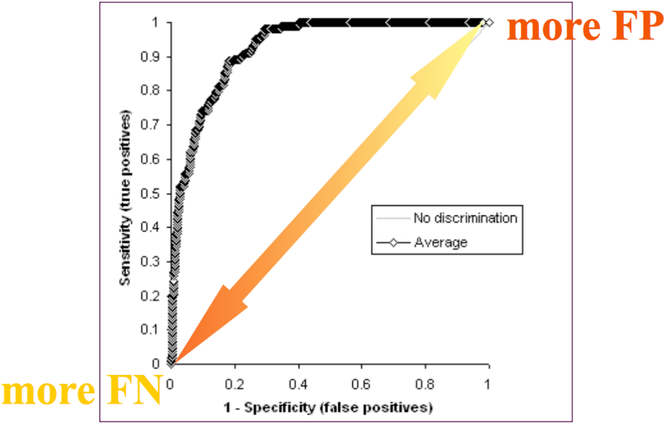
FDR False Discovery Rate

FOR False Omission Rate

NPV Negative Predictive Value

ROC curve: definition

The **Receiver Operating Characteristic** is a representation of the trade-off between r_{TP} *vs* r_{FP} as one changes the parameter controlling the sensitivity of the classifier, such as the offset b.



ROC curve: more rigorously a convex curve

More precisely, the ROC plot is a representation of the attainable regimes.

ROC curve: more rigorously a convex curve

More precisely, the ROC plot is a representation of the attainable regimes.

Convexity property of the ROC plot: Given two points

- (α_0, β_0) for classifier c_0
- (α_1, β_1) for classifier c_1.

Consider the classifier c_λ that uses c_1 with probability λ and c_0 with probability $1 - \lambda$. Then

$$(\alpha_\lambda, \beta_\lambda) = \lambda(\alpha_1, \beta_1) + (1 - \lambda)(\alpha_0, \beta_0).$$

ROC curve: more rigorously a convex curve

More precisely, the ROC plot is a representation of the attainable regimes.

Convexity property of the ROC plot: Given two points

- (α_0, β_0) for classifier c_0
- (α_1, β_1) for classifier c_1.

Consider the classifier c_λ that uses c_1 with probability λ and c_0 with probability $1 - \lambda$. Then

$$(\alpha_\lambda, \beta_\lambda) = \lambda(\alpha_1, \beta_1) + (1 - \lambda)(\alpha_0, \beta_0).$$

Attainable points form a convex set.

ROC curve: more rigorously a convex curve

More precisely, the ROC plot is a representation of the attainable regimes.

Convexity property of the ROC plot: Given two points

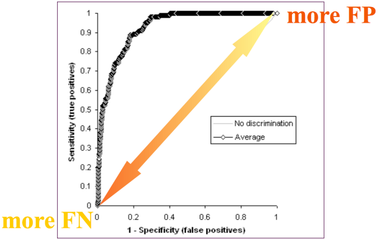
- (α_0, β_0) for classifier c_0
- (α_1, β_1) for classifier c_1.

"Ideal" ROC curve is the concave envelope of the attainable points.

Consider the classifier c_λ that uses c_1 with probability λ and c_0 with probability $1 - \lambda$. Then

$$(\alpha_\lambda, \beta_\lambda) = \lambda(\alpha_1, \beta_1) + (1 - \lambda)(\alpha_0, \beta_0).$$

Attainable points form a convex set.



Precision Recall curve