Probabilities and Statistical Estimation

Chapter 3

University of Amsterdam

Introduction

Frequentist approach

- "The data comes from a distribution, let us find as best as we can which distribution that was"
- Find "estimators" for parameters, and try to figure out how good these estimators are.

Bayesian approach

- "The data could have come from any number of distributions, let us find what those distributions could have been, and how likely they are"
- Using Bayes' rule does not make an approach Bayesian

A few definitions (scalar variables)

- Expectation

$$E[f(x)] \triangleq \int_{-\infty}^{\infty} f(x)p(x)\mathrm{d}x \tag{1}$$

- Mean (Expectation of x):

$$E[x] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x \tag{2}$$

- Variance

$$V[x] \triangleq E[(x - E[x])^2] = \int_{-\infty}^{\infty} (x - E[x])^2 p(x)\mathrm{d}x \tag{3}$$

- Covariance

$$V[x, y] \triangleq E[(x - E[x])(y - E[y]) \tag{4}$$

Multivariate version

- Expectation

$$E[f(\mathbf{x})] \triangleq \int_{\mathbb{R}^n} f(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{5}$$

- Mean (Expectation of x):

$$E[\mathbf{x}] = \int_{\mathbb{R}^n} \mathbf{x}p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{6}$$

- Covariance matrix

$$V[\mathbf{x}] \triangleq E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^\top] \tag{7}$$

Law of large numbers

Law of large numbers:

As N grows large,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \rightarrow E[\mathbf{x}] \tag{8}$$

Change of variables

If $\mathbf{x}$ is an n-vector and $\mathbf{y} = \mathbf{A}\mathbf{x}$ is an m-vector, for an arbitrary mn-matrix $\mathbf{A}$

$$E[\mathbf{y}] = \mathbf{A}E[\mathbf{x}] \tag{9}$$

$$V[\mathbf{y}] = \mathbf{A}V[\mathbf{x}]\mathbf{A}^\top \tag{10}$$

If we perform a functional transformation of variables, $\mathbf{y} = \mathbf{y}(\mathbf{x})$, then by defining $\mathbf{x} = \bar{\mathbf{x}} + \Delta\mathbf{x}$ and $\mathbf{y} = \bar{\mathbf{y}} + \Delta\mathbf{y}$, we obtain to a first approximation

$$\bar{\mathbf{y}} = \mathbf{y}(\bar{\mathbf{x}}) \quad \Delta\mathbf{y} = \left.\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right|_{\bar{x}} \Delta x \quad V[\mathbf{y}] = \left.\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right|_{\bar{x}} V[\mathbf{x}] \left.\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right|_{\bar{x}}^{\top} \tag{11}$$

Principal Component Analysis

PCA is probably the most well-known method for dimensionality reduction.

- Also known as the *Karhunen-Loève transform*
- Orthogonal projection of the data into a lower-dimensional subspace, so that the variance of the projected data is maximised
- Equivalently: linear projection that minimises the mean-squared distance between data points and their projection

Maximising the projected variance

Consider projecting on $\mathbf{u}_1$, with unit length for convenience.

- Each vector $\mathbf{x}_n$ is then projected into $\mathbf{u}_1^\top \mathbf{x}_n$
- The mean of the projected data equals the projected mean $\mathbf{u}_1 \bar{\mathbf{x}}$
- To maximise the variance of the projected data, we maximise

$$\frac{1}{N} \sum_{n=1}^{N} (\mathbf{u}_1^\top \mathbf{x}_n - \mathbf{u}_1^\top \bar{\mathbf{x}}_n)^2 = \mathbf{u}_1^\top V[\mathbf{x}] \mathbf{u}_1 \tag{12}$$

- Using a Lagrange multiplier to constrain $\mathbf{u}_1^\top \mathbf{u}_1 = 1$, we get

$$\mathbf{u}_1^\top V[\mathbf{x}] \mathbf{u}_1 + \lambda_1 (1 - \mathbf{u}_1^\top \mathbf{u}_1) \tag{13}$$

resulting in $V[\mathbf{x}]\mathbf{u}_1 = \lambda \mathbf{u}_1$. That is, $\mathbf{u}_1$ is an eigenvector of $V[\mathbf{x}]$ and the maximum is obtained for the largest eigenvalue.

Interesting properties

- The eigenvalues indicate the variance of the data along the orientation of the corresponding eigenvector
- Since the eigenvectors form an orthonormal basis, the data is uncorrelated along the projection orientations (3.24)

Local distributions

- Definition of manifold
- Definition of dimension / codimension
- Non-singular: codimension $= 1$
- Singular: codimension $\neq 1$
- Tangent space / Normal space
- Local distribution: Assume that the distribution is sufficiently "localised", so that all observations can be approximated as lying in the tangent space of the manifold

3D rotation

Consider a rotation $\mathbf{R}$ as a random variable, representing a rotation $\bar{\mathbf{R}}$ perturbed by some small noise $\Delta\mathbf{R}$: $\mathbf{R} = \bar{\mathbf{R}} + \Delta\mathbf{R}$.
$\bar{\mathbf{R}}$, $\mathbf{R}$ and $\Delta\mathbf{R}$ are rotations, so that we can write:

$$\mathbf{R} = (\mathbf{I} + \Delta\Omega\mathbf{I}\mathbf{I} + O(\Delta\Omega^2))\bar{\mathbf{R}} \tag{14}$$

$$\bar{\mathbf{R}} + \Delta\mathbf{R} = \bar{\mathbf{R}} + \Delta\Omega\mathbf{I}\bar{\mathbf{R}} + O(\Delta\Omega^2) \tag{15}$$

To a first approximation:

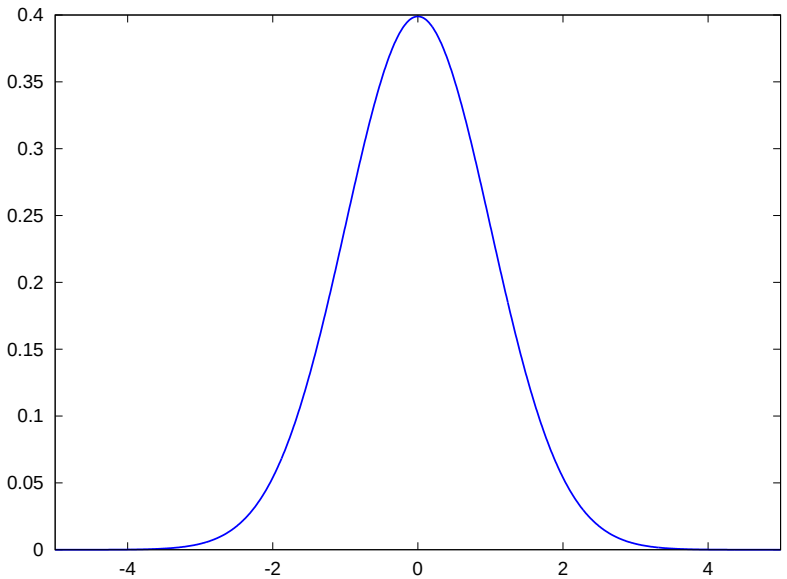
$$\Delta\mathbf{R} = \Delta\Omega\mathbf{I}\bar{\mathbf{R}} \tag{16}$$

The covariance matrix of the rotation can then be defined as:

$$V[\mathbf{R}] = E[\Delta\Omega^2\mathbf{I}\mathbf{I}^\top] \tag{17}$$

The eigenvector with largest associated eigenvalue then approximates the vector around which the rotation is most likely

The Gaussian Distribution



$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \qquad (18)$$

where we can check that

$$E[\mathbf{x}] = \boldsymbol{\mu} \qquad V[\mathbf{x}] = \mathbf{\Sigma} \qquad (19)$$

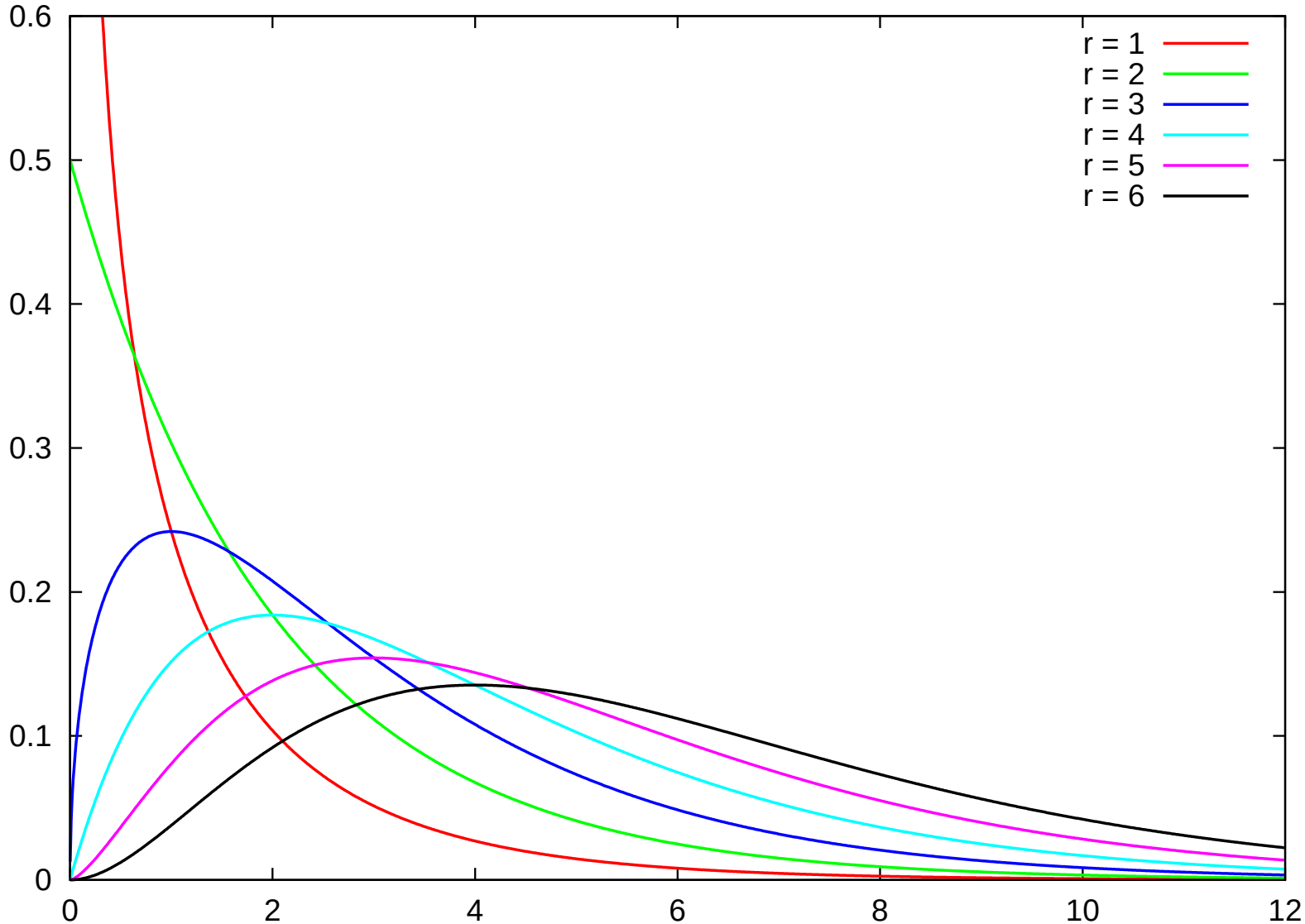
Interesting properties

- The quantity

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{20}$$

  is the squared *Mahalanobis distance* of $\mathbf{x}$ from the mean.
- Uncorrelated normally distributed variables are always independent
- If $\boldsymbol{\Sigma}$ is not full rank, we define the Gaussian distribution in the space spanned by the data
- The *central limit theorem* states that the sum of a sufficiently large number of independent random variables with finite variance converges to a normal distribution.

The χ^2 Distribution



The χ^2 Distribution

If $x_1, \ldots, x_r$ are r independent samples from $\mathcal{N}(0, 1)$, then

$$R = x_1^2 + \cdots + x_r^2 \tag{21}$$

has a χ^2 distribution.
Some nice properties:

- $E[R] = r$
- $V[R] = 2r$
- The mode of the distribution is at $R = r - 2$
- The sum of independent χ^2 variables is χ^2 distributed

Properties of the χ^2 distribution

- For a multivariate random variable $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ of rank r, the quadratic sum

$$R = \mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x} \tag{22}$$

  is χ^2 distributed with r degrees of freedom

χ^2 test

A frequentist approach to reject hypotheses:

- Construct random variables $R = x_1^2 + \cdots + x_r^2$ such that each x_i has zero mean if the hypothesis holds
- If the variables do not have zero mean, $E[R]$ becomes larger
- The hypothesis is rejected with *significance level* a (*confidence level* $(1 - a)$) if R falls in the region $(\chi^2_{r,a}, \infty)$

Note that this allows you to reject hypotheses, not to accept them!

Maximum Likelihood

Find the parameter by maximising the *likelihood*

$$l(\boldsymbol{\theta}) \triangleq p(\{\mathbf{y}\}|\boldsymbol{\theta}) \tag{23}$$

$$= \prod_i p(\mathbf{y}_i|\boldsymbol{\theta}) \tag{24}$$

Maximum a Posteriori

The major problem with ML estimation is *overfitting*; learning the structure of the data extremely well, but performing poorly on new examples.
If we have prior knowledge, we can encode this in the model in the form of a prior probability distribution over the parameters, and update these with the observed data:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{25}$$

where the *marginal probability density* $p(\mathbf{y})$ is a constant. The MAP estimate is obtained by maximising $p(\boldsymbol{\theta}|\mathbf{y})$.
This reduces overfitting, but this is not Bayesian inference.

Bayesian Inference

In Bayesian inference, we learn a distribution over parameters. We consider that all parameters we are not interested in are "nuisance parameters". To obtain the distribution over the quantity of interest, θ_i, we *marginalise out* the other parameters:

$$p(\theta_i|\mathbf{y}) \propto \int_{\theta_{\neg i}} p(\mathbf{y}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}_{\neg i} \tag{26}$$

"Statistical Estimation"

If we have a model $\mathbf{y} = \mathbf{A}\mathbf{x} + \epsilon$, *where* $\mathbf{A}$ *is known and* ϵ *has a Gaussian distribution with known parameters, how do we find* $\mathbf{x}$*?*

Find the parameter by maximising the *likelihood*:

$$\ell(\boldsymbol{\theta}) \triangleq p(\mathbf{y}|\boldsymbol{\theta}), \text{ where } \boldsymbol{\theta} = \mathbf{x} \tag{27}$$

$$= \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp -\tfrac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}) \tag{28}$$

Ignoring constants and since $\exp(\cdot)$ is a monotonically increasing function, this is equivalent to minimising the Mahalanobis distance, resulting in:

$$\hat{\mathbf{x}} = (\mathbf{A}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{29}$$

and the error has a χ^2 distribution

Kalman Filter

Linear dynamical system defined as:

$$\mathbf{x}_{t+1} = \mathbf{A}_t \mathbf{x}_t + \mathbf{B}_t \mathbf{v}_t \tag{30}$$

$$\mathbf{y}_t = \mathbf{C}_t \mathbf{x}_t + \mathbf{w}_t \tag{31}$$

where $\mathbf{A}_t$, $\mathbf{B}_t$ and $\mathbf{C}_t$ are fixed matrices; $\mathbf{v}_t$ and $\mathbf{w}_t$ are normally distributed with known means and covariances.

The filter updates its estimators $\hat{\mathbf{x}}_t$ and $V[\hat{\mathbf{x}}_t]$ at each time step by first estimating them given past observations, and updating it with the current observation.

The Kalman "smoother" additionally performs a backward pass to include information from the future in the estimate of $\mathbf{x}_t$

Fisher Information Matrix

Score:

$$\mathbf{l} = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}) \tag{32}$$

Fisher Information Matrix

$$\mathbf{J} = E[\mathbf{l}\mathbf{l}^{\top}] \tag{33}$$

can be written, if the log-likelihood is twice differentiable, as

$$\mathbf{J} = E[-\nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta})] \tag{34}$$

Intuitively: the more peaked the log-likelihood, the more informative the distribution
Cramér-Rao Lower Bound The Fisher Information Matrix provides a lower bound on the variance of an estimator of a parameter. If the bound is attained, the estimator is said to be *efficient*

Akaike Information Criterion

The Akaike Information Criterion

$$AIC = 2m' - 2\sum_i \log p(\mathbf{x}_i; \hat{\theta}) \tag{35}$$

where m' is the rank of the fisher information matrix $\mathbf{J}$

- This penalises models that are too flexible, and optimises the expected likelihood of future data.
- Other information criteria are also commonly used, such as the BIC, which penalise complex models slightly differently. The BIC penalises a high number of parameters less as more data becomes available.