Probabilistic and Bayesian Modelling in Machine Learning and Artificial Intelligence

Manfred Opper

Background reading

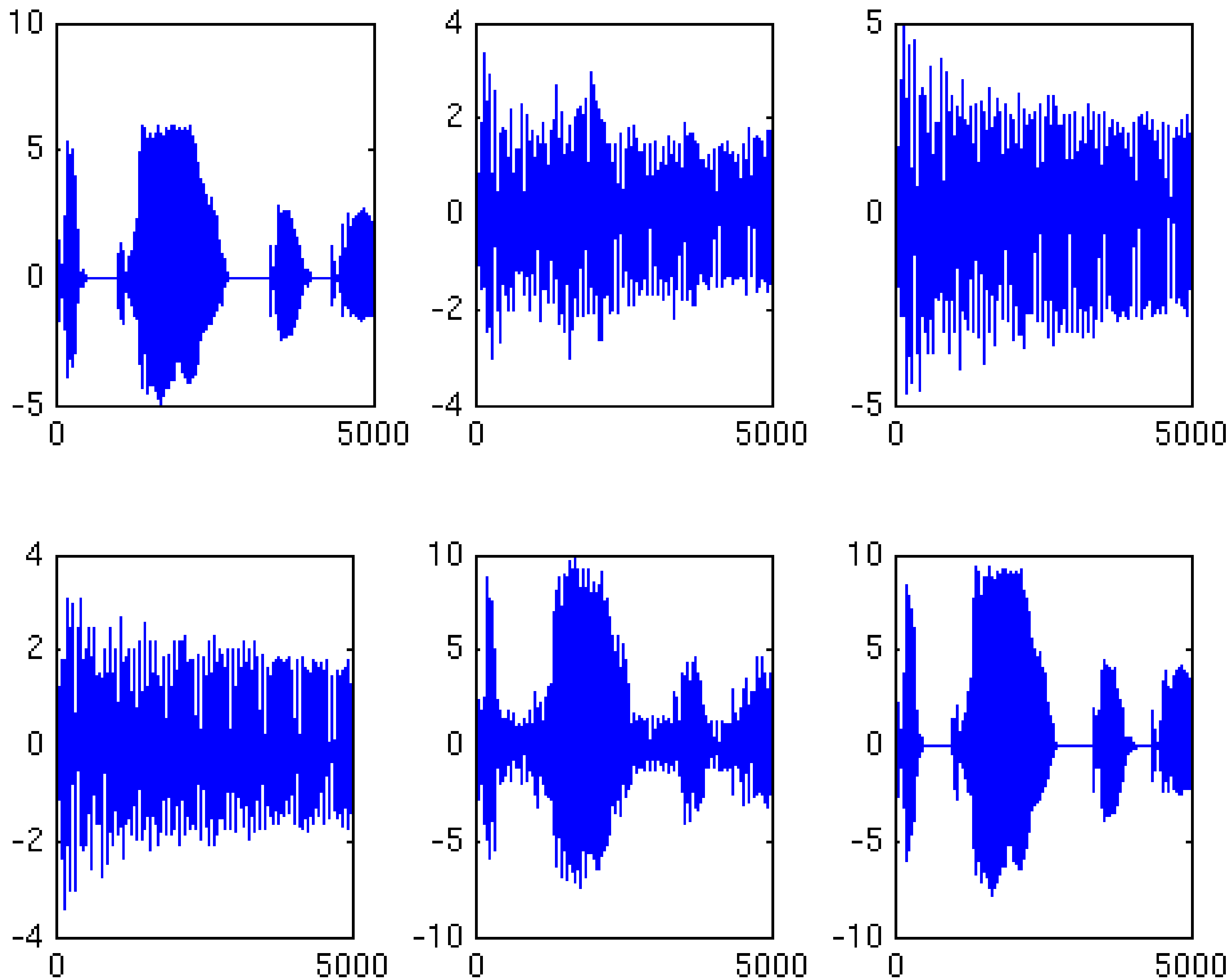
Pattern Recognition and Machine Learning, Christopher M. Bishop, Springer, 2006.

Information Theory, Inference, and Learning Algorithms, David J C MacKay, Cambridge University Press, 2003.

Bayesian Reasoning and Machine Learning, David Barber, Cambridge University Press, 2012.

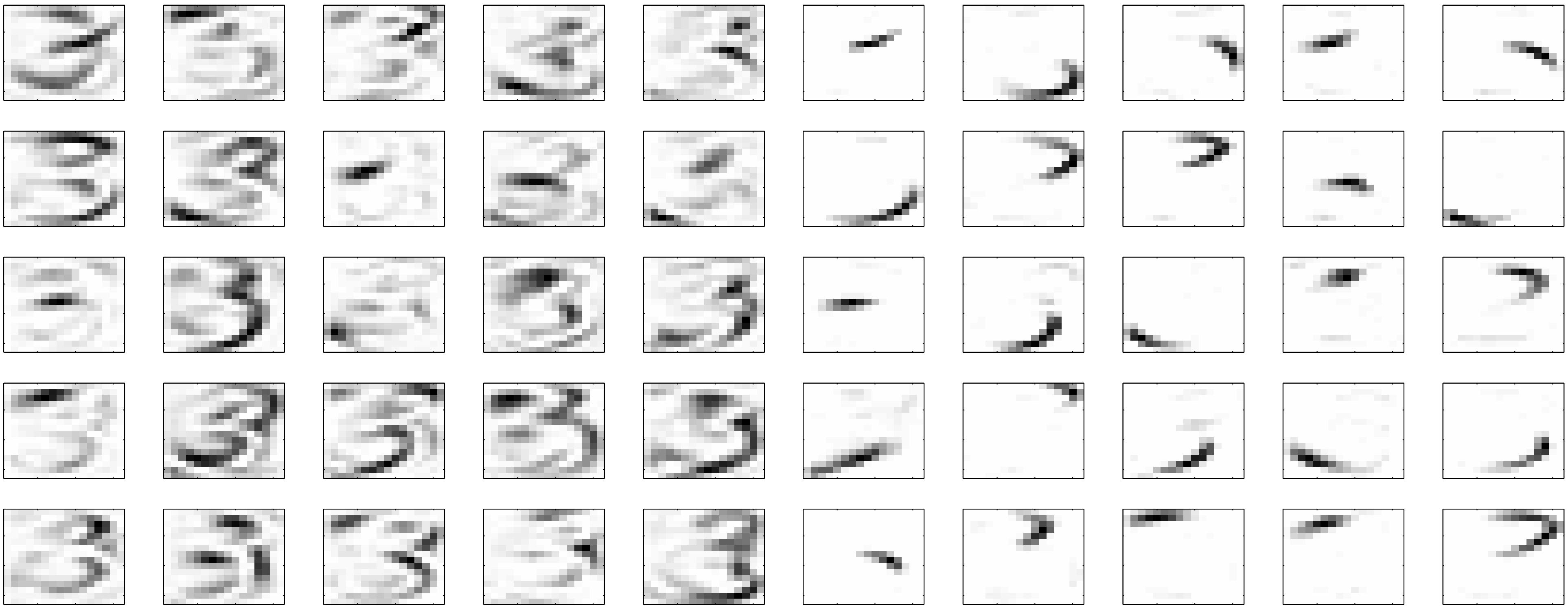
Machine Learning - A probabilistic Perspective, Kevin P. Murphy, The MIT Press, 2012.

Advanced Mean Field Methods, M Opper and D Saad (eds.), The MIT Press, 2001.



10
5
0
-5
0
5000
4
2
0
-2
-4
0
5000
5
0
-5
0
5000
4
2
0
-2
-4
0
5000
10
5
0
-5
-10
0
5000
10
5
0
-5
-10
0
5000

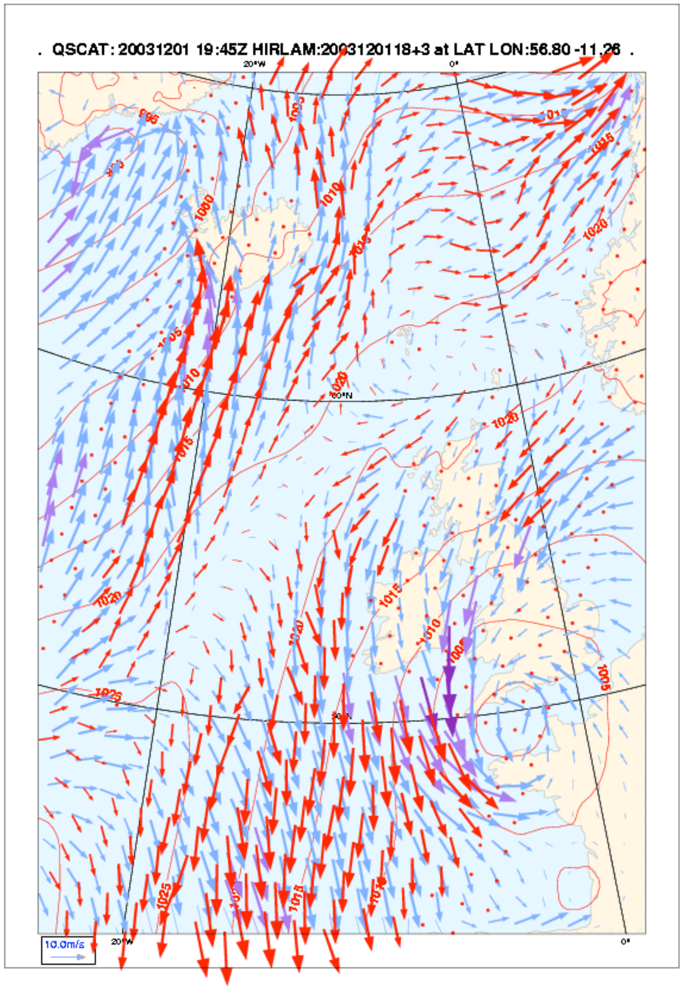
ICA for feature extraction



left: unconstrained **right:** constrained (positive) mixing matrix $\mathbf{A}$.

$x_i(t) =$ sequence of 500 images (handwritten '3's).$p(s) = e^{-s}$, $s \geq 0$.
Shown are the $m = 25$ columns $A_{\bullet j}$ of the matrix $\mathbf{A}$.

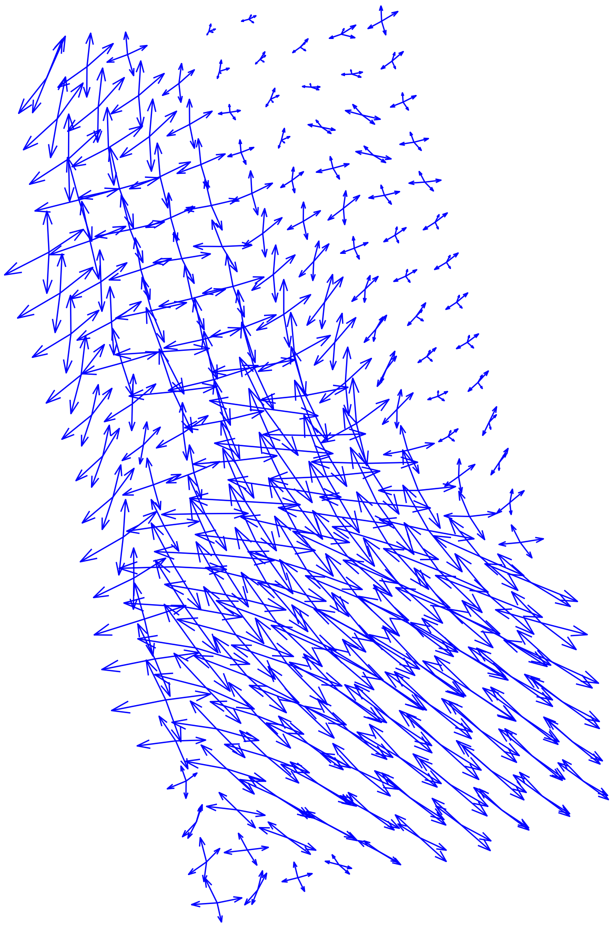
Measuring Windfields

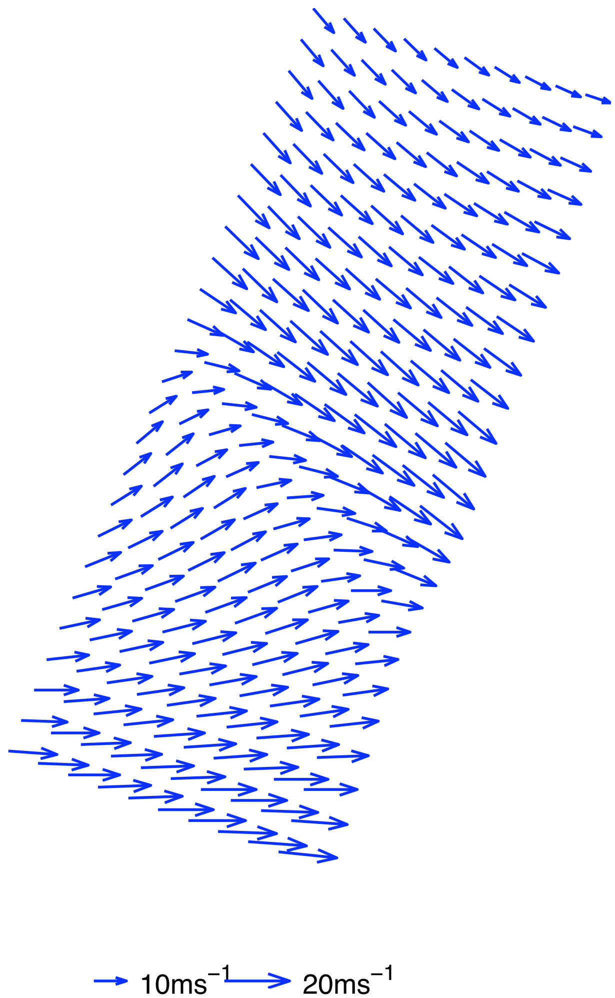


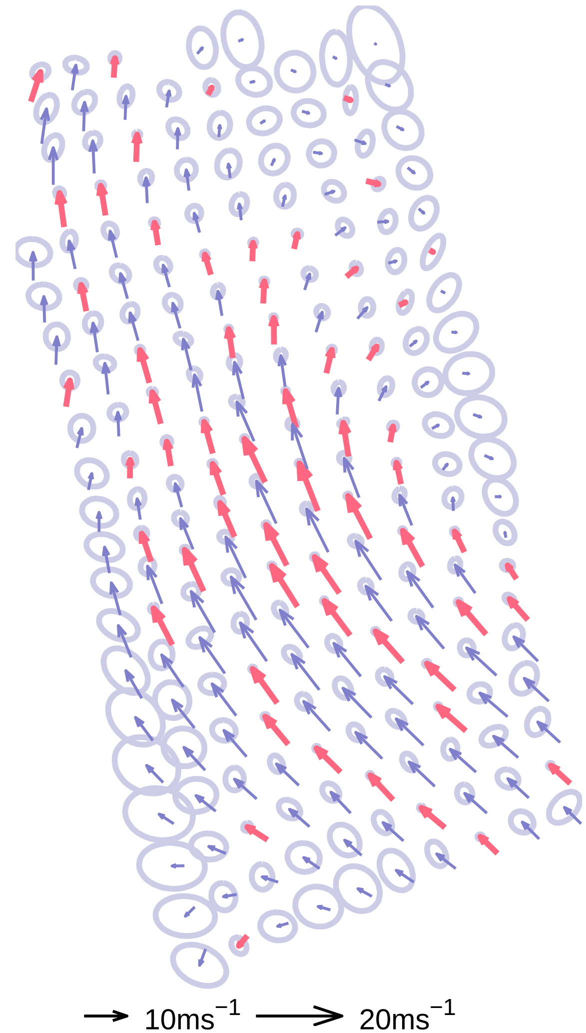
(Ad Stoffelen/KNMI)

Scatterometry: Measuring windfields using radar backscattering on waterwaves (from satellites).

Ambiguities and prior knowledge



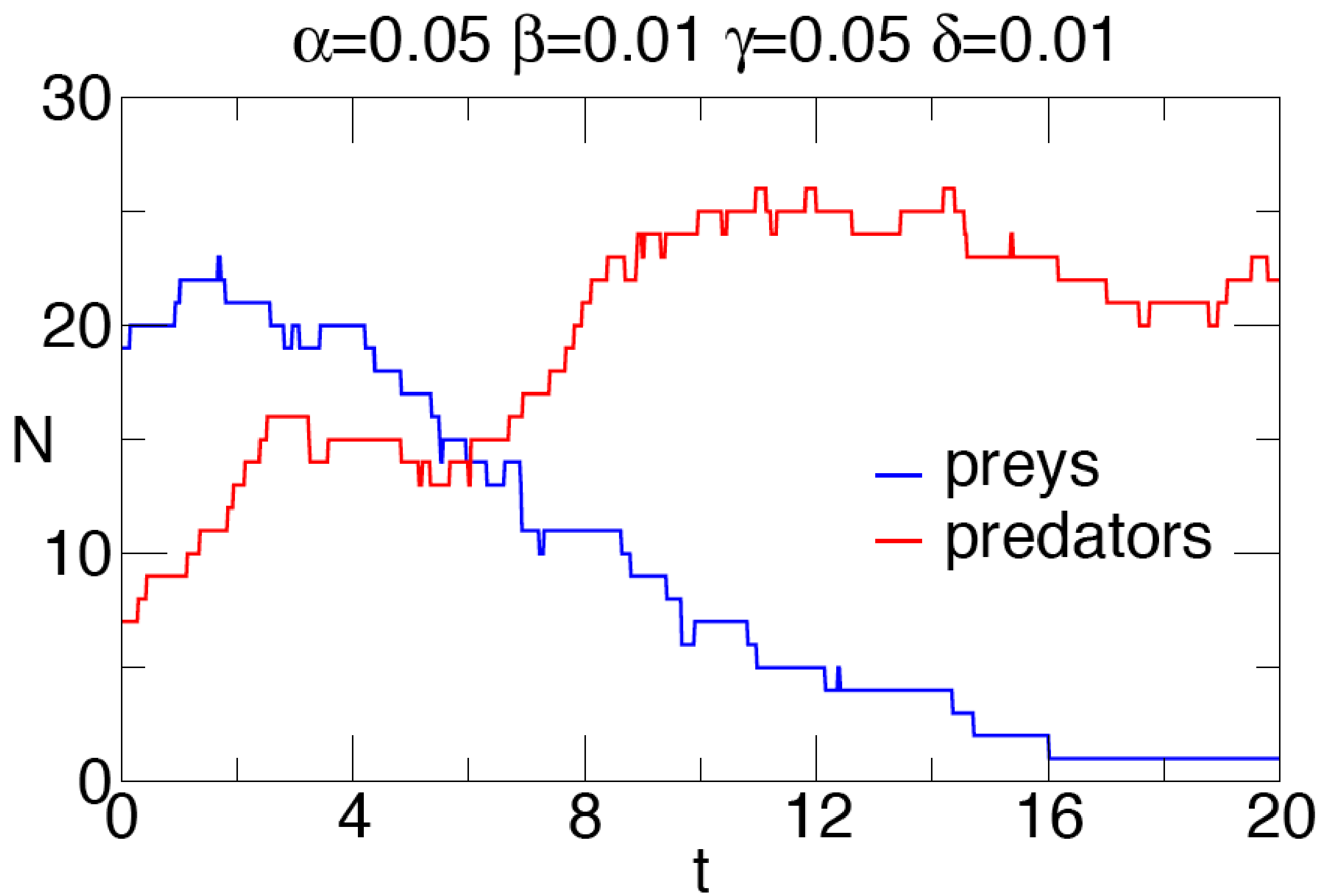




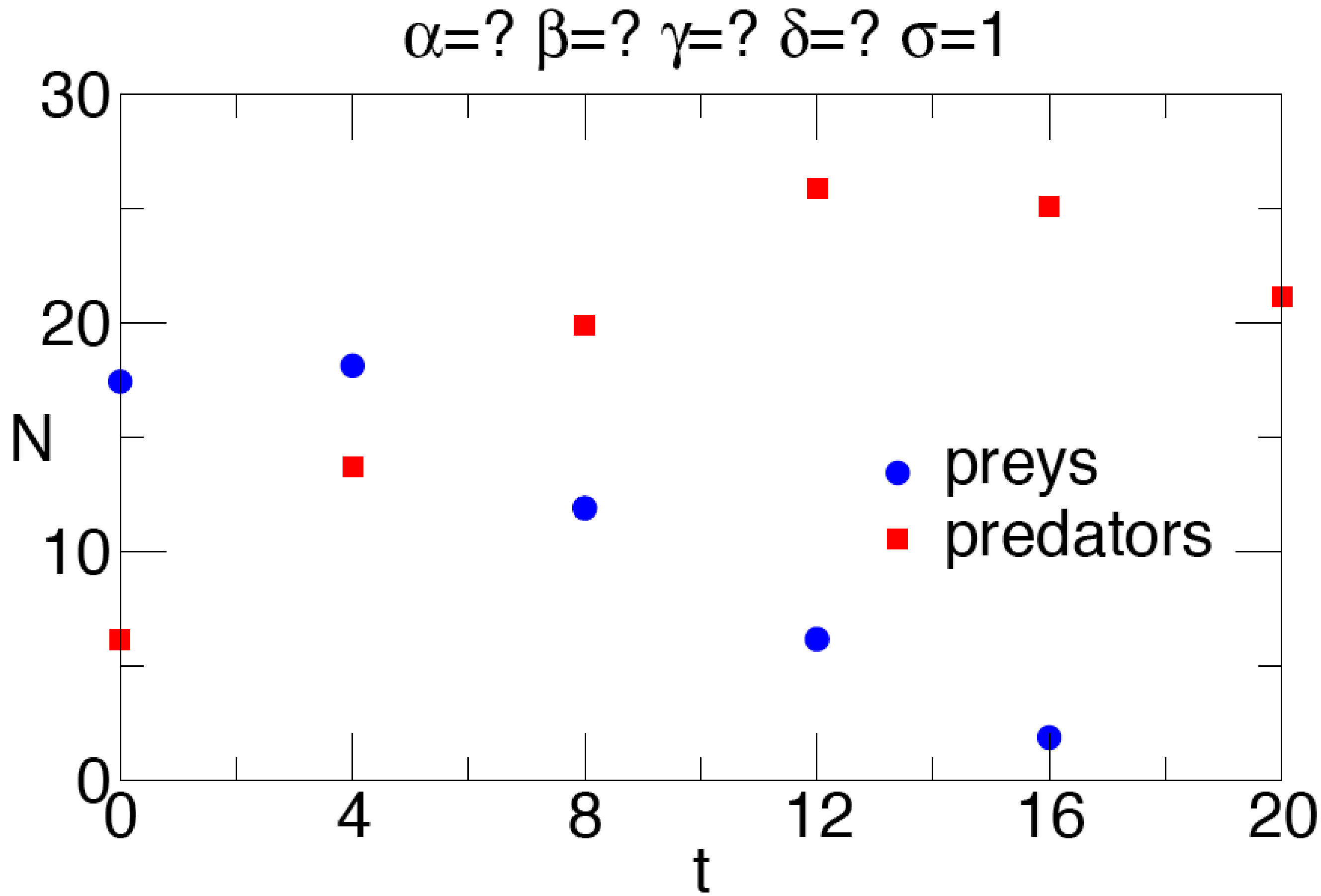
Likelihood typical a priori sample mean prediction.

Stochastic Lotka Volterra Model

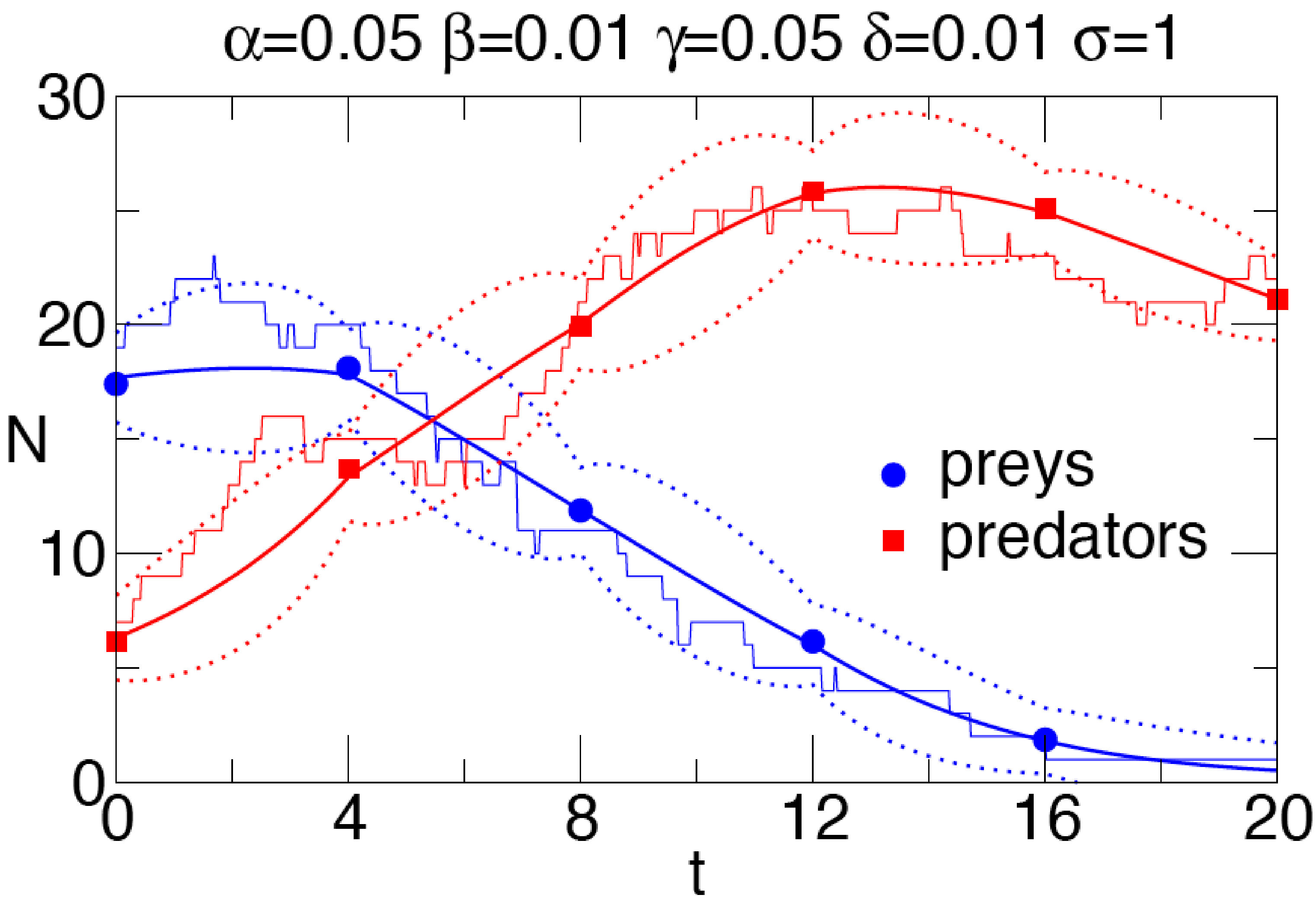
$$\begin{aligned}
\text{Prey} \to 2\,\text{Prey} &\quad \text{with Rate} \quad \alpha X_{\text{Prey}} \\
\text{Prey} \to \varnothing &\quad \text{with Rate} \quad \beta X_{\text{Prey}} X_{\text{Pred}} \\
\text{Predator} \to 2\,\text{Predator} &\quad \text{with Rate} \quad \delta X_{\text{Prey}} X_{\text{Pred}} \\
\text{Pred} \to \varnothing &\quad \text{with Rate} \quad \gamma X_{\text{Pred}}
\end{aligned}$$



The actual time series and the reaction constants



Discrete observations from a continuous time series



α=0.05 β=0.01 γ=0.05 δ=0.01 σ=1
30
20
10
0
N
0
4
8
12
16
20
t
preys
predators

Some probability essentials

Definitions

Sample Space Ω: Space of possible outcomes ω of a random experiment.

Events: (measurable) subsets of Ω.

Probabilities: Number $P(A)$ assigned to events A.

We have $0 \leq P(A) \leq 1$, $P(\emptyset) = 0$ and $P(\Omega) = 1$.

Addition Rule: If $A \cap B = \emptyset$ Then $P(A \cup B) = P(A) + P(B)$ (extends to countable sequence of disjoint events).

Random Variables are functions of outcomes $X(\omega)$.

For *discrete* rvs we define *the probability mass function* $P_X(x) = P(X = x)$. Often we speak (sloppily) about *the distribution* of X.

Joint distribution of two random variables:

$$P_{X,Y}(x,y) = P(X = x, Y = y) \ .$$

Marginal distributions: $P_X(x) = \sum_y P_{X,Y}(x,y)$ and $P_Y(y) = \sum_x P_{X,Y}(x,y)$.

For continuous random variables we define a *probability density* $p_X(x)$ by $\int_a^b p_X(x)\, dx = P(a < X < b)$.

A *joint density* can be defined for two (and more) variables:

$$\int\int_S p_{X,Y}(x,y)\, dxdy = P((X,Y) \in S)$$

for a set $S \in R^2$. *.

Marginal densities are obtained e.g. as $p(x) = \int_{-\infty}^{\infty} p(x,y)dy$

*Note: When it is clear which random variables are involved, I often write simply $p(x)$ instead of $p_X(x)$.

Transformation of random variables and their densities:

Let $y = f(x)$ be an invertible transformation and let the density of x be $p(x)$. We are interested in the density $q(y)$ of the random variable y.

Using $p(x)dx = q(y)dy$, we get

$$q(y) = p(x(y)) \left| \frac{dx}{dy} \right| = p(x(y)) \frac{1}{\left| \frac{dy}{dx} \right|}$$

Conditional Probabilities

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ and similarly for conditional distributions: $P(x|y) = \frac{P(x,y)}{P(y)}$ and *conditional densities* $p(x|y) = \frac{p(x,y)}{p(y)}$.

Bayes Rule!!!

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(x)}{\sum_{x'} P(y|x')P(x')}.$$

Expectations

The expectation of X is defined as

$E(X) = \sum_x P(x)\ x$ (discrete case) or $E(X) = \int p(x)\ x\ dx$ (continuous case). For a function g of the rva X, we can show that

$E(g(X)) = \sum_x P(x)\ g(x)$ (discrete) or $E(g(X)) = \int p(x)\ g(x)\ dx$ (continuous).

Mean: $\mu = E[X]$

Variance: $Var(X) = E((X - \mu)^2) = E(X^2) - (E(X))^2$.

Linearity

$E(aX + bY) = aE(X) + bE(Y)$

Conditional Expectation

$E(Y|X = x)$ or $E(Y|x)$:

$E(g(Y)|X = x) = \sum_y g(y)\ P(y|x)$ (discrete case) and $E(g(Y)|X = x) = \int g(y)\ p(y|x)\ dy$ (continuous case).

Independence

(*Multiplication rule*):

A family of events $A_1, A_2, \ldots$ are called *independent* if for any subset $\{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$ $P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$.

A family of random variables $X_1, X_2, \ldots$ are called *independent* if for any subset $\{X_{i_1}, X_{i_2}, \ldots, X_{i_k}\}$ $P(X_{i_1}, X_{i_2}, \ldots, X_{i_k}) = P(X_{i_1})P(X_{i_2}) \cdots P(X_{i_k}) = \prod_{j=1}^{k} P(x_{i_j})$ (with an analogous definition for densities). Hence, if X and Y independent then $P(x|y) = \frac{P(x,y)}{P(y)} = P(x)$.

Some properties of independent random variables $X_1, X_2, \ldots, X_N$:

- $E(X_1 \cdot X_2 \cdots X_N) = \prod_{i=1}^{N} E(X_i)$.

- $\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \mathrm{Var}(X_i)$.

- <u>Law of large numbers</u>

  Let $X_1, X_2, \ldots, X_N$, i.i.d. with finite variance σ^2 and $S_N = \frac{1}{N}\sum_{i=1}^{N} X_i$, then one can show that

  $\lim_{N\to\infty} P(|S_N - E(X)| > \varepsilon) = 0.$

  Hence, when N large, with high probability we have $\frac{1}{N}\sum_{i=1}^{N} X_i \approx E(X)$.

  The proof uses addititivity of VAR and *Markov's* inequality.

Reminder of Gaussian densities

1-D Gaussian density

The density of a one dimensional Gaussian random variable $x \sim \mathcal{N}(\mu, \sigma^2)$ with *mean* $E(x) = \mu$ and variance $\sigma^2 = E(x-\mu)^2$ is given by

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The d-dimensional Gaussian distribution

Let $\mathbf{x} = (x_1, \ldots, x_d)^T$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$

The Gaussian density for $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

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

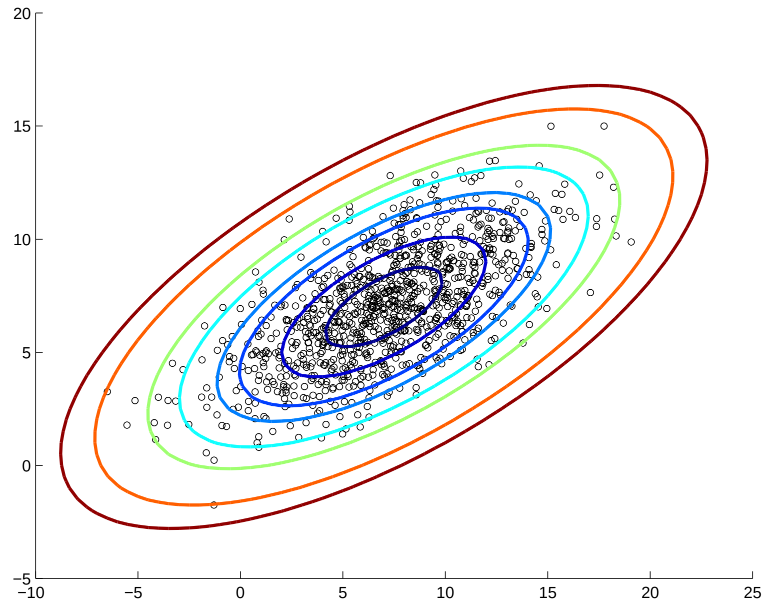
$\boldsymbol{\mu} = E[\mathbf{x}]$ is **mean** vector and $\boldsymbol{\Sigma}$ is a $d \times d$ *covariance* matrix. One can show that

$$\Sigma_{ij} = E(x_i - \mu_i)(x_j - \mu_j)$$

or $\Sigma = E(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T$.

Example:

Lines of constant density and random data for a two dimensional Gaussian. The mean is $\boldsymbol{\mu} = (7,7)^T$ and the covariance matrix is $\boldsymbol{\Sigma} =$



$$\begin{pmatrix} 16.6 & 6.8 \\ 6.8 & 6.4 \end{pmatrix}$$

Eigenvalue problem for $\boldsymbol{\Sigma}$

To understand the properties of this density, we need to make a little detour and consider

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{2}$$

with an eigenvector $\mathbf{u}_i$ and eigenvalue λ_i, where $i = 1, \ldots, d$. $\boldsymbol{\Sigma}$ is a real symmetric matrix with orthonormal eigenvectors $\mathbf{u}_i \cdot \mathbf{u}_j = \mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$. With the $d \times d$ *orthogonal* matrix formed by the d column eigenvectors

$$\mathbf{U} = (\mathbf{u}_1 \mathbf{u}_2 \cdots \mathbf{u}_d). \tag{3}$$

we have $\mathbf{U}^T\mathbf{U} = \mathbf{I}$.

Using (3) and the diagonal matrix $\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$ we can rewrite the eigenvalue equations (2) as $\mathbf{\Sigma U} = \mathbf{U \Lambda}$ or

$$\mathbf{\Sigma} = \mathbf{U \Lambda U}^{-1} = \mathbf{U \Lambda U}^T \tag{4}$$

and

$$\mathbf{\Sigma}^{-1} = \mathbf{U \Lambda}^{-1} \mathbf{U}^{-1} = \mathbf{U \Lambda}^{-1} \mathbf{U}^T \tag{5}$$

$\mathbf{U}$ defines an *orthogonal* transformation by $\mathbf{y} = \mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu})$, or $\mathbf{x} = \boldsymbol{\mu} + \mathbf{U y}$. This transformation preserves inner products, i.e. we have for two vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ that $\mathbf{y}_1^T \mathbf{y}_2 = (\mathbf{x}_1 - \boldsymbol{\mu})^T (\mathbf{x}_2 - \boldsymbol{\mu})$. It can be understood as a transformation to a new coordinate system given by a combination of a *shift* and a *rotation*. We also get

$$(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{y}^T \mathbf{\Lambda}^{-1} \mathbf{y} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \ldots + \frac{y_d^2}{\lambda_d}$$

Using the new coordinate system, we see that

- surfaces of constant probability density for the Gaussian density $p(\mathbf{x})$, eq. (1) are *ellipsoids*.

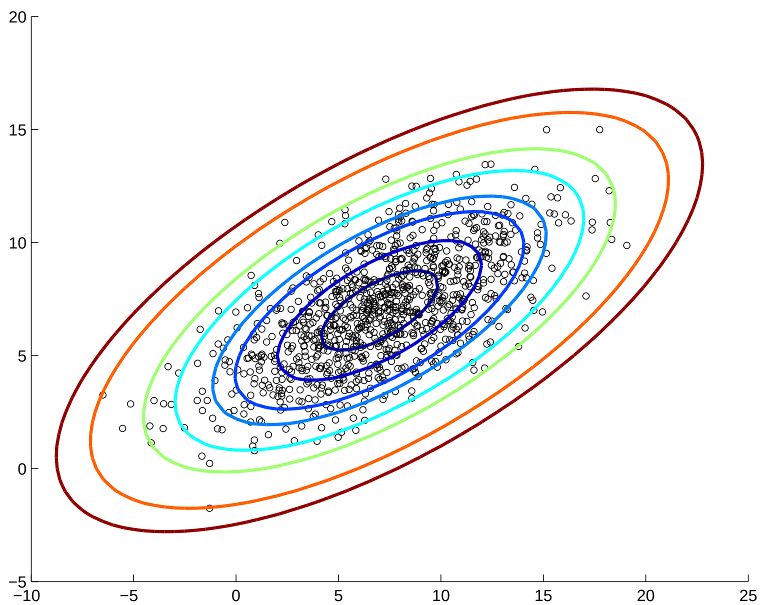
- the random variables defined by y coordinates $\mathbf{Y} = \mathrm{U}^T(\mathbf{X} - \boldsymbol{\mu})$ are *independent*, ie.

  $p(\mathbf{y}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\lambda_i}} e^{-\frac{y_i^2}{2\lambda_i}}$

- We see that Σ is indeed the matrix of covariances, i.e

  $\Sigma_{ij} = E(x_i - \mu_i)(x_j - \mu_j)$, i.e. $\Sigma = E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T$.

Back to the example:



The covariance matrix is $\mathbf{\Sigma} = \begin{pmatrix} 16.6 & 6.8 \\ 6.8 & 6.4 \end{pmatrix}$. The eigenvalues are $\lambda_1 = 20$ and $\lambda_2 = 3$ with eigenvectors $\mathbf{u}_1 = \frac{1}{\sqrt{5}}(2, 1)^T$, and $\mathbf{u}_2 = \frac{1}{\sqrt{5}}(1, -2)^T$.

- Generate Gaussian distributed random vectors $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ from vectors $\mathbf{z}$ with *indepedent* normal components $E(z_i z_j) = \delta_{ij}$ by the transformation $\mathbf{x} = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{z} + \boldsymbol{\mu}$.

  Alternative method: Perform *Cholesky decomposition* $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$. Then set $\mathbf{x} = \mathbf{A}\mathbf{z}$.

- Sums of jointly Gaussian random variables are Gaussian. Marginal & conditional densities of jointly Gaussian random variables are Gaussian.

- Central limit theorems: For i.i.d. x_i with finite variance, the normalised sum $z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - m)$ becomes asymptotically Gaussian distributed.

Some inequalities

<u>Cauchy–Schwarz:</u>

$$\{E(xy)\}^2 \le E(x^2)E(y^2) \; .$$

Equality = if and only if $P(sx = ty) = 1$ for some nonrandom s and t.

<u>Markov:</u>

$$P(x \ge a) \le \frac{E(x)}{a}$$

for $x \ge 0$.

<u>Chebychev:</u>

$$P(|x| \ge a) \le \frac{E(x^2)}{a^2}$$

Follows from *Markov* by substituting $x \to x^2$.

Jensen

For $f(\cdot)$ **convex** (i.e. $f^{''}(x) \geq 0$ for all x) we have

$$E[f(X)] \geq f(E[X])$$

Proof: For fixed (non random y), Use the Taylor expansion

$$f(X) = f(y) + (X - y)f'(y) + \frac{1}{2}(X - y)^2 f^{''}(\xi) \geq f(y) + (X - y)f'(y)$$

where $\xi \in [x, y]$. we have

$$E[f(X)] \geq f(y) + (E[X] - y)f'(y)$$

The result follows by setting $y = E[X]$. If f strictly convex: Equality $=$ if and only if $X = E(X)$ a.e.

The KL divergence

For any two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$, we can show using Jensen's inequality that the **Kullback–Leibler divergence**

$$KL(p, q) = E_p \left[\ln \frac{p(\mathbf{x})}{q(\mathbf{x})} \right] \geq 0$$

where E_p denotes expectation wrt to p. One has equality $= 0$ if and only if $p = q$ almost everywhere. The KL is a asymmetric dissimilarity measure between distributions. It is invariant against transformations of the random variables.