united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

SMR: 1343/2

EU ADVANCED COURSE IN COMPUTATIONAL NEUROSCIENCE

An IBRO Neuroscience School

(30 July - 24 August 2001)

"Advanced Statistics"

presented by:

Martin STETTER
Siemens AG, Corporate Technology
CT IC 4, Otto-Hahn-Ring 6
81739 München
GERMANY

Statistics – Short Intro

Martin Stetter

Siemens AG, Corporate Technology
CT IC 4, Otto-Hahn-Ring 6
81739 München, Germany

martin.stetter@mchp.siemens.de

Why Statistics in CNS?

- Analysis of biological neural data.
- Statistical characterization of sensory data.
- Modeling the brain as a statistical inference machine...
  - Statistical characterization of optimal codes.
  - Description of brain dynamics at a population level.
  - Statistical biophysical models (reaction kinetics).
- You name it...

Topics to be Addressed

- Introduction and Definitions.
- Moments and Cumulants.
- Some Basic Concepts of Information Theory.
- Point Processes.
- Statistical Modeling – Basic Concepts.
- Statistical Modeling – Examples.

Introduction and Definitions

One Discrete Random Variable

Consider a discrete random variable X.
If sampled ("trial"), it randomly assumes one of the values
$X = x(1), x(2), \ldots, x(M)$.

Probability for event $X = x(i)$: $\mathrm{Prob}(X = x(i)) =: p_i$.

Properties:

- $0 \leq p_i \leq 1$. $p_i = 0 : x(i)$ does'nt ever occur; $p_i = 1 : x(i)$ occurs in every trial ($\Rightarrow$ no other $x(j)$ ever observed).
- $\sum_{i=1}^{N} p_i = 1$. After all, anything must happen in a trial.

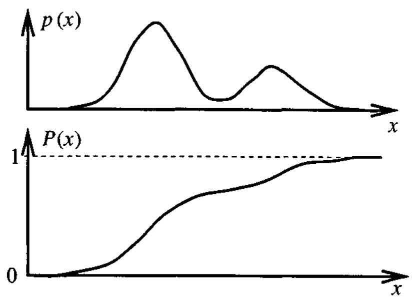
One Continuous Random Variable

A continuous random variable X can take any real or complex value x (possibly within an interval). We assume $x \in \mathbb{R}$
Need different probability concept (because $\mathrm{Prob}(X = x) = 0$).

- $P(x) := \mathrm{Prob}(X \leq x)$ Cumulative distribution function.
- $p(x)dx := \mathrm{Prob}(x \leq X \leq x + dx)$. $p(x)$ = probability density function (pdf), if dx small (infinitesimal).

Properties:

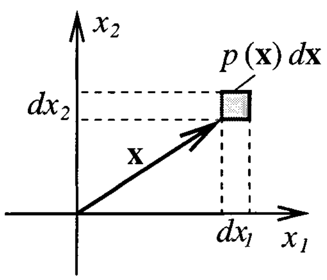
- $p(x) \geq 0$.
- $\int_{-\infty}^{\infty} p(x')\, dx' = 1$.
- $P(x) = \int_{-\infty}^{x} p(x')\, dx'$. $0 \leq P \leq 1$; $P(x)$ monotonically increasing.

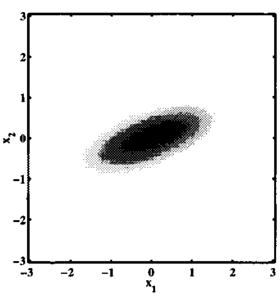


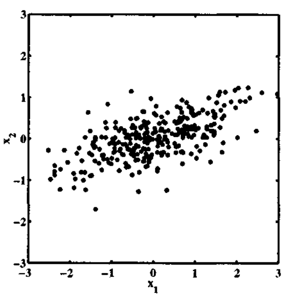
Several Random Variables, Random Vectors

Often, several random variables $X_1, \ldots, X_d$ have to be considered at the same time (multivariate data).

- $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ Random vector. Assumes values $\mathbf{x} \in \mathbb{R}^d$.
- $p(\mathbf{x}) = p(x_1, x_2, ..., x_d)$ Joint probability density function.







Gray-level image of a 2D-Gaussian pdf $p(\mathbf{x})$

300 vectors drawn from p

Example: Multivariate Gaussian Distribution

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}(\det \mathbf{G})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{G}^{-1}(\mathbf{x}-\mu)\right) =: \phi(\mathbf{x}|\mu, \mathbf{G})$$

μ = mean; $\mathbf{G}$ = covariance matrix (symmetric, positive definite, see later).

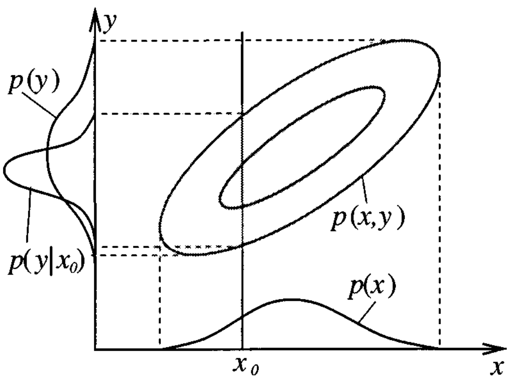
Interpretations of Probabilities

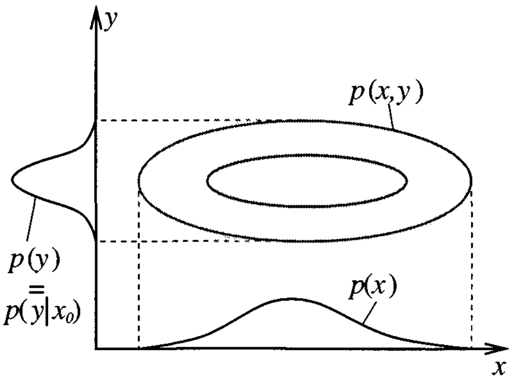
- Relative frequency of finding a value x_i of a random variable after many trials ("frequentist philosophy").
- Our belief, that one of several possibilities (labelled by x_i) will happen in the near future ("Bayesian philosophy").

Some Probability Definitions and Probability Laws

Consider two random variables X and Y.

- Conditional Probability: $p(y|x)dy$ Probability for finding $Y \in [y, y+dy]$, if we already know that $X \in [x, x+dx]$.
- $p(y,x) = p(y|x)p(x) = p(x|y)p(y)$.
- Marginalization: $p(y) = \int p(y,x)dx = \int p(y|x)p(x)dx$.
- X and Y are independent $\Longleftrightarrow p(x,y) = p(x)p(y) \Leftrightarrow p(y|x) = p(y)$.





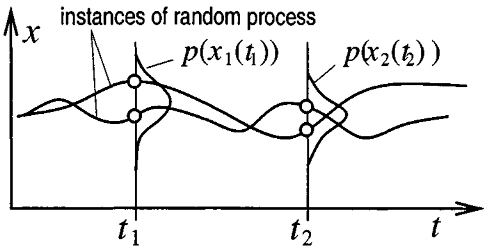
Bayes' Law: $p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y')p(y')dy'}$

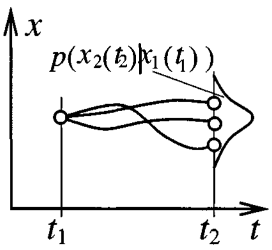
Consider higher-dimensional random vectors $\mathbf{X} = (X_1, \ldots, X_d)$:

- Bayes for Subsets $\mathbf{X}$, $\mathbf{Y}$: $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})/p(\mathbf{x})$.
- Decomposition: $p(x_1, ..., x_d) =$ $p_d(x_d|x_{d-1}, ..., x_1)p_{d-1}(x_{d-1}|x_{d-2}, ..., x_1) \ldots p_2(x_2|x_1)p_1(x_1)$
- Special case: Independence: $p(x_1, ..., x_d) = p_d(x_d)\ p_{d-1}(x_{d-1}) \ldots p_1(x_1) = \prod_{k=1}^{d} p_k(x_k)$
- Special case: 1st order Markov chain: $p(x_1, ..., x_d) = p_d(x_d|x_{d-1})\ p_{d-1}(x_{d-1}|x_{d-2}) \ldots p_2(x_2|x_1)\ p_1(x_1)$

Random Processes and Random Fields

- Random process = random time series (e.g. LFP data, Spike train etc...).
- Formally: Set of random variables $X(t)$ labelled by a (discrete) time index t. It assumes random values $x(t)$. $p(x(t))$ is the probability for observing the value x in the small time interval around t.
- Full characterization by joint pdf: $p(x_1(t_1), x_2(t_2), \ldots, x_d(t_d)) \equiv p(x)$
- Stationary random process: $p(x_1(t_1), x_2(t_2), \ldots, x_d(t_d)) = p(x_1(t_1 + \tau), \ldots, x_d(t_d + \tau))\ \ \forall \tau$. In particular: $p_k(x(t_k)) = p_k(x(t_k + \tau))$.
- 1st order Markov process: $p(x_1(t_1), x_2(t_2), \ldots, x_d(t_d)) =$ $p_d(x_d(t_d)|x_{d-1}(t_{d-1})) \ldots p_2(x_2(t_2)|x_1(t_1))p_1(x_1(t_1))$





Gaussian random process: z

$$p(x) = \frac{1}{N} \exp\left(-\frac{1}{2}\sum_{t,t'}(x(t) - \mu(t))G(t,t')(x(t') - \mu(t'))\right).$$

- Random field = random image (e.g. instances are a set of camera images).
- Formally: Set of random variables $X(\mathbf{r})$ labelled by a (discrete) spatial index $\mathbf{r}$, characterized by $p(x_1(\mathbf{r}_1), x_2(\mathbf{r}_2), \ldots, x_d(\mathbf{r}_d)) \equiv p(x(\mathbf{r}))$.
- 1st order Markov random field: $p(x_1(\mathbf{r}_1), x_2(\mathbf{r}_2), \ldots, x_d(\mathbf{r}_d)) = \prod_k p_k(x(\mathbf{r}_k)|x(\mathbf{r}_k + (\Delta r, 0)), x(\mathbf{r}_k + (0, \Delta r))$.

Moments and Cumulants

The Mean

Consider random vector $\mathbf{X} = (X_1, \ldots, X_d)$, distributed according to $p(\mathbf{x})$. Consider a function $f(\mathbf{X})$.

Def. Mean:

$$\langle f \rangle := \int p(\mathbf{x})\ f(\mathbf{x})\ d\mathbf{x}$$

- $p(\mathbf{x})$ known $\Rightarrow$ Means of arbitrary functions available.
- Means for all functions f known $\Rightarrow$ $p(\mathbf{x})$ can be determined $\Rightarrow$ Statistics of $\mathbf{X}$ completely known.

Practically, set of all functions f not accessible.
$\Rightarrow$ Look for a “clever” set of functions.

Moments

Def.: n-th order moment of random vector $\mathbf{X}$:

$$\langle X_{i1} X_{i2} \dots X_{in} \rangle = \int p(\mathbf{x})\, x_{i1}\, x_{i2}, \dots x_{in}\, d\mathbf{x}$$
$$= \int p(x_{i1}, \dots x_{in})\, x_{i1}, \dots x_{in}\, dx_{i1}, \dots dx_{in}\ .$$

Examples:

- 1st order: $\mu_i = \langle X_i \rangle = \int p(x_i)\, x_i\, dx_i$ = mean value; $\mu = (\mu_1, \dots, \mu_d)$.
- 2nd order: $\langle X_1^2 \rangle$, $\langle X_2 X_3 \rangle$.
- 3rd order: $\langle X_1 X_2 X_3 \rangle$, $\langle X_1^3 \rangle$, $\langle X_2 X_3^2 \rangle$

All moments known $\Rightarrow$ Mean of every Taylor-expansible function known.

Cumulants – Motivation

Goal: Construct functions of random variables, which characterize $p(\mathbf{x})$ efficiently and have additional desirable properties.

Example – Covariance:

$$G_{i,j} = \langle X_i X_j \rangle - \mu_i \mu_j = \int dx_i \int dx_j\ p(x_i, x_j) x_i x_j - \mu_i \mu_j .$$

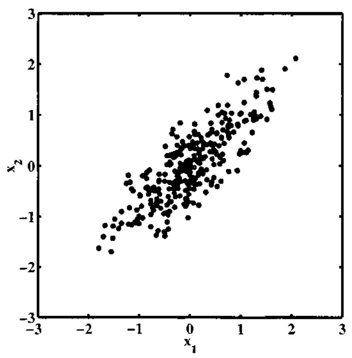
$G_{i,i} = \sigma_i^2$ = variance of X_i.

Covariance matrix of a random vector $\mathbf{X}$:

$\mathbf{G} = (G_{i,j}) = \langle \mathbf{X}\ \mathbf{X}^T \rangle - \mu\ \mu^T$.

If $\mathbf{X}$ has independent components $\Rightarrow$ $G_{i,j} = \delta_{i,j} \sigma_i^2$.

Plot: Gaussian data, $G_{i,i} = 1$; $G_{1,2} = 0.8$;



Cumulants – Definition

Def: Cumulant generating function: Log of (inv.) fourier transform of $p(\mathbf{x})$:

$$\Phi(\mathbf{s}) := \log \int \exp(i\mathbf{s}^T\mathbf{x})\ p(\mathbf{x})\ d\mathbf{x} = \log\langle\exp(i\mathbf{s}^T\mathbf{x})\rangle$$

Def: n-th order cumulant of random vector $\mathbf{X}$:

$$\langle\langle X_{i1},\ \dots, X_{in}\rangle\rangle = (-i)^n\ \frac{\partial\Phi(s_1, s_2\ \dots\ s_d)}{\partial s_{i1}\ \partial s_{i2},\ \dots\ \partial s_{in}}\bigg|_{\mathbf{s}=0}$$

Cumulants are expansion coefficients of Φ around 0 (existence assumed). Hence:

- All cumulants known
  $\Rightarrow$ we know $\Phi(\mathbf{s})$
  $\Rightarrow$ we know $p(\mathbf{x}) = \frac{1}{2\pi^d} \int \exp(-i\mathbf{x}^T\mathbf{s})\ \exp(\Phi(\mathbf{s}))\ d\mathbf{s}$.

But even more interestingly ...

Cumulants – Some Properties

- $p(\mathbf{x})$ Gaussian $\Rightarrow$ all higher than 2. order cumulants vanish.

  Why?
  – The Fourier transform of a Gaussian is again a Gaussian.
  – Its logarithm, $\Phi(\mathbf{s})$ is a square function.
  – All higher than second order derivatives vanish...
  $\Rightarrow$ Use higher order cumulants for detection of non-Gaussianity.

- $\mathbf{X}$ has independent components $\Leftrightarrow$ All cross-cumulants vanish.
  Only the cumulants $\langle\langle X_i^n\rangle\rangle, i = 1, \dots, d$ can be non-zero.

  Why?
  – $p(\mathbf{x})$ factorizes $\Leftrightarrow$ we find $\Phi(\mathbf{s}) = \sum_i \phi_i(s_i)$,
  $\Leftrightarrow$ Mixed derivatives vanish.

Cumulants – Examples

- **1. order:**
  $\boxed{\langle\langle X_1\rangle\rangle = \langle X_1\rangle = \mu_1 \quad \text{Mean}}$
- **2. order:** $\langle\langle X_1, X_2\rangle\rangle = \langle X_1X_2\rangle - \mu_1\mu_2 = G_{1,2}$ (covariance).
  $\boxed{\langle\langle X_1, X_1\rangle\rangle = \sigma_1^2 \text{ variance}}$
- **3. order:** $\langle\langle X_1, X_2, X_3\rangle\rangle = \langle X_1X_2X_3\rangle$
  $-\langle X_1X_2\rangle\mu_3 - \langle X_2X_3\rangle\mu_1 - \langle X_3X_1\rangle\mu_2 + \mu_1\mu_2\mu_3$;
  $\boxed{\langle\langle X_1, X_1, X_1\rangle\rangle = \langle X_1^3\rangle - 3\langle X_1^2\rangle\mu_1 + \mu_1^3 = S \quad \text{skewness}}$
- **4. order:** (... Gulp ...) but for zero-mean data (μ = 0):
  $\langle\langle X_1, X_2, X_3, X_4\rangle\rangle = \langle X_1X_2X_3X_4\rangle$
  $-\langle X_1X_2\rangle\langle X_3X_4\rangle - \langle X_1X_3\rangle\langle X_2X_4\rangle - \langle X_1X_4\rangle\langle X_2X_3\rangle$.
  $\boxed{\langle\langle X_1, X_1, X_1, X_1\rangle\rangle = \langle X_1^4\rangle - 3\sigma_1^4 = K \quad \text{kurtosis}}$

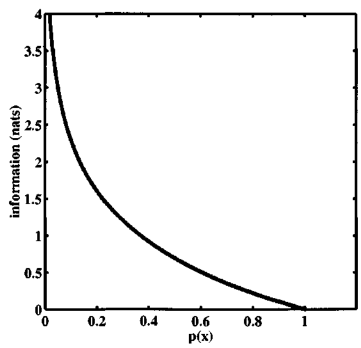
Some Basic Concepts of Information Theory

Shannon Information and Information Entropy

- Consider a discrete random variable X, which assumes values $x(i),\ i = 1,\ \dots, M$ with probabilities p_i.
- Find a quantity "information" such that
  – it describes our "surprise" to see a particular value $x(i)$ happen.
  – the information of two statistically independent events adds up linearly.

Def: Information of symbol $x(i)$:

$$J(x(i)) = \ln\frac{1}{p_i} = -\ln p_i$$



Def: Shannon entropy:

$$H(X) = \langle J \rangle = -\sum_i p_i \ \ln p_i$$

- Average information gained by sampling X once.
- Average length of message (nats) needed at least to describe one observation.

For continuous random variables, we can define the *differential entropy* (should be scored against a reference value, e.g. for $p_i = const.$)

$$H(X) = -\int p(x) \ \ln p(x) \ dx.$$

The (differential) entropy for a random vector is

$$H(\mathbf{X}) = -\int p(\mathbf{x}) \ \ln p(\mathbf{x}) \ d\mathbf{x}.$$

Some Properties

- Among all random variables with bounded values (discrete or cont.) $H(X) = \max \iff p = const.$ (uniform distribution).
- Among all random variables with the same mean μ and variance σ^2 $H(X) = \max \iff p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(x-\mu)^2/2\sigma^2\right)$ (Gaussian distribution).
- $\mathbf{X} = (X_1, \ldots, X_d)$ random vector distributed according to $p(\mathbf{x})$. X_i are independent, $p(\mathbf{x}) = \prod_i p_i(x_i) \iff H(\mathbf{X}) = \sum_{i=1}^d H(X_i)$ $(H(X_i) = -\int p_i(x_i) \ln p_i(x_i) \ dx_i$ marginal or pixel entropies)
- Generally: $H(\mathbf{X}) \leq \sum_{i=1}^d H(X_i)$

Information Capacity C

Def. Information capacity C: Maximum amount of information that can be carried by the random vector $\mathbf{X}$ (is a function of $p(\mathbf{x})$).

Can be achieved, if

- $\mathbf{X}$ has independent components: $H(\mathbf{X}) = \sum_i H(X_i)$, and
- each marginal entropy $H(X_i)$ is maximal

Example: Discrete random vector, M values for each of the d components.

- Maximum marginal entropies: $\Rightarrow H(X_i) = -\sum_{k=1}^{M} \frac{1}{M} \ln \frac{1}{M} = \ln M$.
- Independence $\Rightarrow C = H_{\max}(\mathbf{X}) = d\, H(X_i) = d \ln M$.

Interpretation of capacity: Maximum description length we can expect for the random variable $\mathbf{X}$.

Redundancy

In presence of redundancy, we need less than the capacity to describe the statistics.

Def. Redundancy of $\mathbf{X}$:

$$R = 1 - \frac{H(\mathbf{X})}{C}$$

or

$$R = \underbrace{\frac{1}{C}\left(C - \sum_{i=1}^{d} H(X_i)\right)}_{\text{(1) Due to non-uniform dist.}} + \underbrace{\frac{1}{C}\left(\sum_{i=1}^{d} H(X_i) - H(\mathbf{X})\right)}_{\text{(2) Due to mutual dependencies}} .$$

Mutual Information

Term (2) of the redundancy is proportional to the *Mutual Information* I between the components X_i:

$$I(\mathbf{X}) = \int p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{\prod_{i=1}^{d} p_i(x_i)}.$$

$I(\mathbf{x})$ measures, how much $p(\mathbf{x})$ differs from factorization.

Kullback-Leibler Divergence

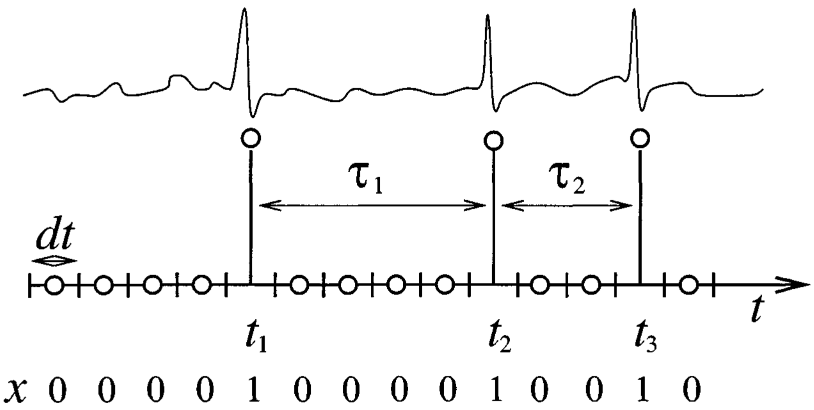
Distance measure between two pdf's $p(\mathbf{x})$ and $q(\mathbf{x})$:

$$K(p||q) = -\int d\mathbf{x}\, p(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})}$$

Observation: $I(\mathbf{X}) = K(p(\mathbf{x})||\prod_i p_i(x_i))$.

Stochastic Point Processes

Definition



– Binary random process.

– Resembles an "idealized spike train" ...

– Defined by set $\{t_i\}$ of "1"-events or by t_1 plus $\{\tau_i\}$ of interspike-intervals.

Statistical Characterization

Consider process of length T, divided into N small bins dt: $T = N\ dt$.

- Full characterization: Joint density for all configurations of events:

$$p(x_1, \dots, x_N)\ d^N t \quad x_i \in \{0, 1\} \qquad (2^N \text{ numbers}).$$

- Special case: Instantaneous Rate: (Biology: PSTH)

$$R(t_n) = R(n\ dt) = p(x_n = 1) = \frac{\text{Prob(Spike in } [t_n, t_n + dt])}{dt}.$$

  Stationary process: $R(t) = R$.

- Special case: Autointensity function: (Biology: Autocorrelogram)

$$C(t_m, t_n) = p(x_m | x_n) = \frac{\text{Prob(Spike in } [t_n, t_n + dt] \mid \text{Spike in } [t_m, t_m + dt])}{dt}$$

  Stationary process: $C(t_1, t_2) = C(t_2 - t_1, 0) \equiv C(t_2 - t_1)$.

Poisson-Process

Point process with independent events x_n: $p(x_1, \dots, x_N) = \prod_n^N p(x_n)$.

Some Properties:

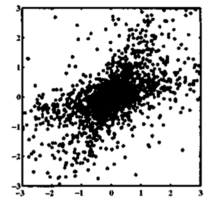
- Fully characterized by the rate $R(t)$. Prob(Spike in $[t, t + dt]$) = $R(t)\ dt$.
- Homogeneous Poisson process: $R(t) \equiv R$.
- Interval density: $p(\tau) = R \exp(-R\tau)$.
- Interspike intervals are statistically independent.
- Spike count Z in time interval T is Poisson-distributed:
  $P(Z) = \exp(-L) L^Z / Z!$; $L = \int_0^T R(t) dt$ (Homog: $L = RT$).
  $\Longrightarrow \langle Z \rangle = \sigma_Z^2 = RT$.

General Aspects of Statistical Data Modeling

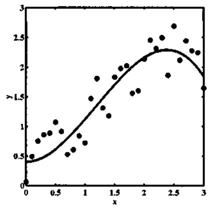
Goals of Statistical Modeling

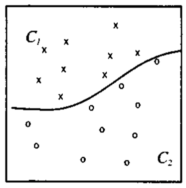
Realistic Situation:
Data sample $\mathbf{x}^{(\alpha)}$, $\alpha = 1, \ldots, P$, drawn from an unknown pdf $p(\mathbf{x})$.
Goal: Extraction of statistical structure from the data. Important techniques:

- **Density estimation.**
  Estimate the pdf underlying the data. Characterize redundancies (structure) therein.



- **Function approximation.**
  Regression: characterize functional relationships between pairs of data.
  Classification: characterize underlying prob. of class-membership.

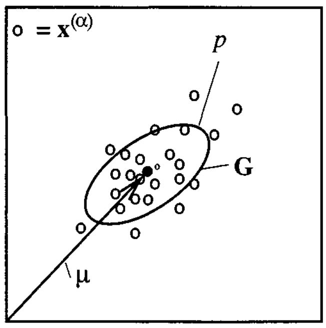




Parametric Density Estimation

- Model the pdf of random vector $\mathbf{X}$ as a parameterized function: $p(\mathbf{x}|\mathbf{w})$.
  (We introduce a "Generative Model" for the data).
- $\mathbf{w}$ is the set of parameters.
  $\mathbf{w}$ is optimized such that data are optimally described by $p(\mathbf{x}|\mathbf{w})$.
- Example: Gaussian: $p(\mathbf{x}|\mathbf{w}) = \phi(\mathbf{x}|\mu, \mathbf{G})$.

Advantage: Few parameters to be estimated.
Disadvantage: Who knows, if the model is correct?

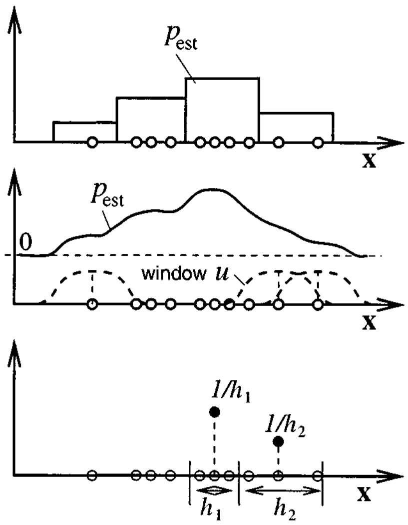


Non-Parametric Density Estimation

Functional form of pdf $p(\mathbf{x})$ is not specified in advance (est. from data only).

Examples:

- Histogram method.
  Divide data space into intervals of width h. Calculate relative frequencies.



- Kernel density estimator.
  "Smoothing" of data cloud:
  $\hat{p}(\mathbf{x}) = \sum_{\alpha=1}^{P} u((\mathbf{x} - \mathbf{x}^{(\alpha)})/h)$,
  with $u(\mathbf{x}) \geq 0$, $\int u(\mathbf{x})\, d\mathbf{x} = 1$.
- K-nearest neighbors.
  Average over K adjacent data points, no fixed h.

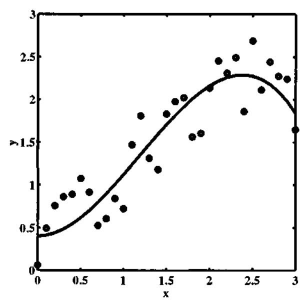
Parameters h or K have to be suitably chosen (nontrivial).

Regression

Formulate model for underlying deterministic structure in input-output pairs, $(\mathbf{x}^{(\alpha)}, \mathbf{y}^{(\alpha)})$, $\alpha = 1, \ldots, P$ of data:

$$\mathbf{y} = \mathbf{f}(\mathbf{w}; \mathbf{x}) + \mathbf{n}$$

$\mathbf{f}$ = regression function, parameterized by $\mathbf{w}$.
$\mathbf{n}$ = random noise vector.



dots: data. line: $\mathbf{f}$.

- Parameter set $\mathbf{w}$ is adjusted for an optimal description of data.
- Link to density estimation:
  – Estimate joint density $p(\mathbf{x}, \mathbf{y})$.
  – Take regression function as conditional average:
  $\mathbf{f}(\mathbf{x}) = \hat{\mathbf{y}}(\mathbf{x}) = \int \mathbf{y}\; p(\mathbf{y}|\mathbf{x})\; d\mathbf{y}$.

Maximum Likelihood Parameter Estimation

Goal: Optimize params $\mathbf{w}$ of parametric models given the data $\{\mathbf{x}^{(\alpha)}\}$.
Principle: Adjust $\mathbf{w}$ as to maximize the likelihood, that the observed data have been generated by the model:

$\mathbf{w}^{\mathrm{ML}} = \mathrm{argmax}_{\mathbf{w}} L(\mathbf{w})$ with

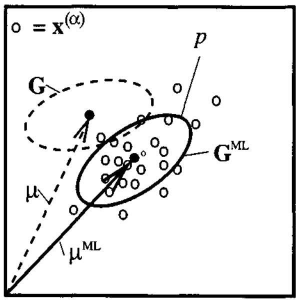
$$L(\mathbf{w}) := p(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(P)}|\mathbf{w}) = p(\{\mathbf{x}^{(\alpha)}\}|\mathbf{w}).$$

Equivalently: Minimize negative log likelihood:
$\mathbf{w}^{\mathrm{ML}} = \mathrm{argmin}_{\mathbf{w}} E(\mathbf{w})$ with

$$E(\mathbf{w}) = -\ln L(\mathbf{w}) = -\ln p(\{\mathbf{x}^{(\alpha)}\}|\mathbf{w}).$$

For independently drawn data points:

$$E(\mathbf{w}) = -\ln \prod_\alpha p(\mathbf{x}^{(\alpha)}|\mathbf{w}) = -\sum_\alpha \ln p(\mathbf{x}^{(\alpha)}|\mathbf{w}).$$



Gauss: Optimize μ, $\mathbf{G}$

Maximum Likelihood for Regression

Consider regression via estimation of the joint density $p(\mathbf{y}, \mathbf{x})$.
ML for data $(\mathbf{x}^{(\alpha)}, \mathbf{y}^{(\alpha)})$, $\alpha = 1, \ldots, P$:

$$L(\mathbf{w}) = \prod_\alpha p(\mathbf{y}^{(\alpha)}, \mathbf{x}^{(\alpha)}|\mathbf{w}) = \prod_\alpha p(\mathbf{y}^{(\alpha)}|\mathbf{x}^{(\alpha)}, \mathbf{w})\, p(\mathbf{x}^{(\alpha)})$$

Leaving constants away:

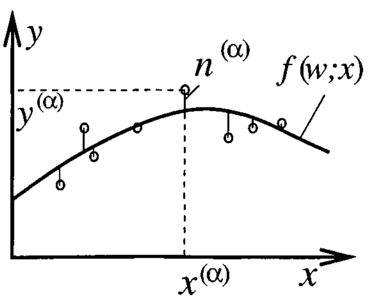
$$E(\mathbf{w}) = -\sum_\alpha \ln p(\mathbf{y}^{(\alpha)}|\mathbf{x}^{(\alpha)}, \mathbf{w}).$$

With $\mathbf{y}^{(\alpha)} = \mathbf{f}(\mathbf{w}; \mathbf{x}^{(\alpha)}) + \mathbf{n}^{(\alpha)}$ and $\mathbf{n}$ distrib. according to $p_n(\mathbf{n})$:

$$p(\mathbf{y}^{(\alpha)}|\mathbf{x}^{(\alpha)}, \mathbf{w}) = p_n(\mathbf{y}^{(\alpha)} - \mathbf{f}(\mathbf{w}; \mathbf{x}^{(\alpha)})).$$

For Gaussian noise, $p_n = \phi$, ML = least squares:

$$E(\mathbf{w}) = -\frac{1}{2\sigma^2} \sum_\alpha \left(\mathbf{y}^{(\alpha)} - \mathbf{f}(\mathbf{w}; \mathbf{x}^{(\alpha)})\right)^2.$$



Bayesian Inference

Goal: Specify whole pdf of model parameters $\mathbf{w}$ given
– the known data set $\{\mathbf{x}^{(\alpha)}\} =: \chi$ and
– prior knowledge (the amount of "blind" belief in the models).
Principle: Use Bayes' law as follows:

$$\underbrace{p(\mathbf{w}|\chi)}_{\text{posterior}} = \frac{1}{p(\chi)} \underbrace{p(\chi|\mathbf{w})}_{\text{likelihood}} \underbrace{p(\mathbf{w})}_{\text{prior}}$$

Bayesian density estimation: Use average over all models:

$$p(\mathbf{x}|\chi) = \int p(\mathbf{x}|\mathbf{w})\, p(\mathbf{w}|\chi)\, d\mathbf{w}$$

Special Case: Maximum a Posteriori (MAP) parameter estimation (reduces to maximum likelihood for flat priors).

$$\mathbf{w}^{\mathrm{MAP}} = \mathrm{argmax}_{\mathbf{w}} p(\mathbf{w}|\chi)$$

Statistical Data Modeling – Examples

Density Estimation by Gaussian Mixture Models

General mixture model:

$$p(\mathbf{x}) = \sum_{j=1}^{M} p(\mathbf{x}|j)P(j)$$

$p(\mathbf{x}|j)$ = component density.
$P(j)$ = mixing parameter. Specifies probability that the data point is generated by component j.

Gaussian mixture model:

$$p(\mathbf{x}) = \sum_{j=1}^{M} \phi(\mathbf{x}|\mu_j, \mathbf{G}_j)P(j)$$

Optimize μ_j, $\mathbf{G}_j$, and $P(j)$, $j = 1, ..., M$, e.g. by Maximum Likelihood.

Regression by Linear Models

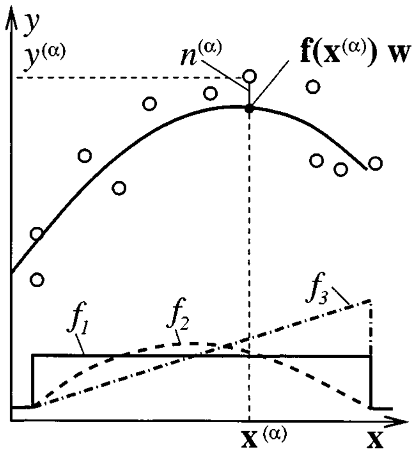
Special case: Regression function is linear in $\mathbf{w}$: $f(\mathbf{w};\mathbf{x}) \equiv \mathbf{f}(\mathbf{x})\mathbf{w}$.
$\Longrightarrow$ Output is linear combination of prototype functions.

For a data set $\mathbf{x}^{(\alpha)}$, $y^{(\alpha)}$, $\alpha = 1, \ldots, P$:

$$y^{(\alpha)}(\mathbf{x}^{(\alpha)}) = \sum_{j=1}^{M} f_j(\mathbf{x}^{(\alpha)}) w_j + n^{(\alpha)}$$

$$\mathbf{y} = \mathbf{F}\mathbf{w} + \mathbf{n}, \quad \mathbf{y} = (y^{(\alpha)})$$

$$\mathbf{F} = \begin{pmatrix} f_1(\mathbf{x}^{(1)}) & f_2(\mathbf{x}^{(1)}) & \cdots & f_M(\mathbf{x}^{(1)}) \\ f_1(\mathbf{x}^{(2)}) & f_2(\mathbf{x}^{(2)}) & \cdots & f_M(\mathbf{x}^{(2)}) \\ \vdots & \vdots & & \vdots \\ f_1(\mathbf{x}^{(P)}) & f_2(\mathbf{x}^{(P)}) & \cdots & f_M(\mathbf{x}^{(P)}) \end{pmatrix}$$



design matrix

Maximum Likelihood Solution for Linear Models

Gaussian noise: ML provides least squares solution:

$$\mathbf{w}^{\mathrm{ML}} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{y}$$

Estimate of noise variance:

$$\hat{\sigma}_n^2 = \frac{(\mathbf{R}\mathbf{x})^T(\mathbf{R}\mathbf{x})}{\mathrm{trace}(\mathbf{R}\mathbf{x})},$$

$\mathbf{R} = \mathbf{I} - \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}$ = residual generating matrix.
Variance of parameter w_j:

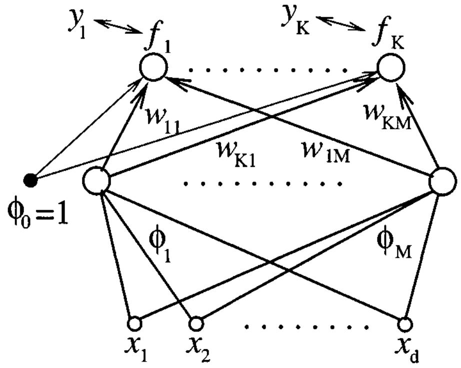
$$\sigma_j^2 = \left(\hat{\sigma}_n^2(\mathbf{F}^T\mathbf{F})^{-1}\right)_{j,j}$$

Radial Basis Functions for Regression

Generalization/unification of linear models and gaussian mixture models.
– Approximate regression function by linear superpositions of Gaussians:

$$f_k(\mathbf{x}) = \sum_{j=0}^{M} w_{kj}\ \phi_j(\mathbf{x}|\mu_j, \mathbf{G}_j), \quad \phi_0 \equiv 1, \quad k = 0, \ldots, K$$
$$\mathbf{f}(\mathbf{x}) = \mathbf{W}\ \phi(\mathbf{x}), \quad \phi = (\phi_0, \phi_1, \ldots, \phi_M) \equiv (\phi_j)$$

- Optimize parameters $\mu_j, \mathbf{G}_j$ of Gaussians by use of input values $\mathbf{x}^{(\alpha)}$ only (like Mixture of Gaussians).
- Use optimal Gaussians as model functions of a linear model:
  $\mathbf{W}^T = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{Y}$.
  $\Phi_{\alpha j} = \phi_j(x^{(\alpha)})$; $\mathbf{Y}_{\alpha k} = y_k^{(\alpha)}$.
  $j = 0, \ldots, M$; $k = 1, \ldots, K$.

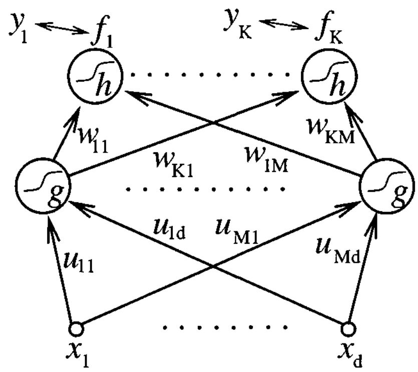


RBF as a Neural Network.

Multilayer Perceptrons (MLP)

Can be viewed as a generalization of RBF Networks
– to several layers,
– to nonlinear transfer functions,
– to arbitrary weights.

Dynamics for two layers:



Two-layer MLP (bias omitted).

$a_j = \sum_{i=1}^{d} u_{ji}x_i, \quad z_j = g(a_j)$ Synaptic input and output of hidden node j.
$a_k = \sum_{j=0}^{M} w_{kj}z_j, \quad f_k = h(a_k)$. Synaptic input and output of output node k.

Total:

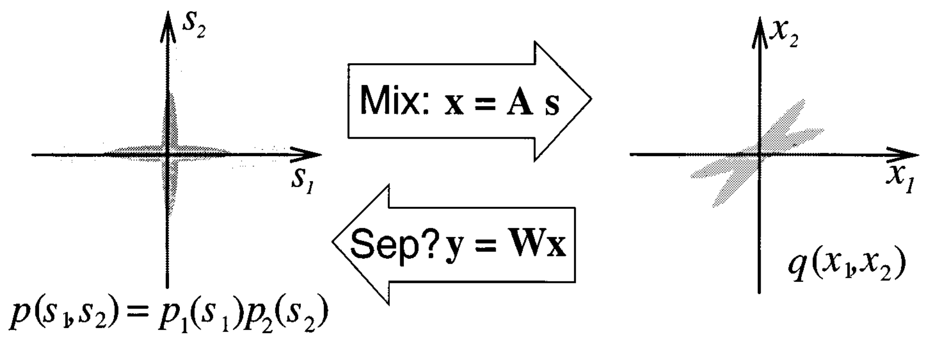
$$f_k(\mathbf{W}, \mathbf{U}; \mathbf{x}) = h\left(\sum_{j=0}^{M} w_{kj}\ g\left(\sum_{i=0}^{d} u_{ji}\ x_i\right)\right) = h(\sum_{j=0}^{M} w_{kj}g(a_j)) = h(a_k).$$

Error-Backpropagation:

Recipe from minimization of cost function E (e.g. $E = \sum_k (y_k - f_k)^2$):

- Apply input $\mathbf{x}^{(\alpha)}$, calculate PSPs a_j and activities z_j, f_k (forward propagation).
- Calculate output errors $\delta_k = \partial E / \partial a_k$ (e.g. $\delta_k = h'(a_k)(f_k - y_k)$).
- Propagate errors back trough the net: $\delta_j = g'(a_j) \sum_k w_{kj} \delta_k$.
- Modify weights according to $\Delta w_{kj} = -\delta_k z_j$.

Independent Component Analysis (ICA)



- **Problem:** Measure $\mathbf{x} = \mathbf{A}\ \mathbf{s}$. Mixing matrix $\mathbf{A}$ and sources $\mathbf{s}$ unknown.
- **Goal:** Estimate both $\mathbf{A}$ and $\mathbf{s}$ from the fact (or assumption), that $\mathbf{s}$ has independent components.

Solution (sketch):

- Formulate generative model for the pdf of the observed data:
  $q(\mathbf{x}) = \det \mathbf{W}\; p(\mathbf{W}\mathbf{x}) = \det\mathbf{W}\; \prod_i p_i((\mathbf{W}\;\mathbf{x})_i)$.
  (Can be viewed as a special case of density estimation).

- Maximum likelihood:
  $E(\mathbf{W}) = -\frac{1}{P}\log\prod_\alpha q(\mathbf{x}^{(\alpha)}) = -\frac{1}{P}\sum_\alpha\sum_i \log p_i((\mathbf{W}\mathbf{x}^{(\alpha)})_i) - \log\det\mathbf{W}$
  = min.

- Resulting update rule (nat. gradient):
  $\Delta\mathbf{W} = (\mathbf{I} - \Theta(\mathbf{y})\mathbf{y}^T)\mathbf{W}; \quad \mathbf{y} = \mathbf{W}\mathbf{x}; \quad \theta_i = \frac{p_i'}{p_i}$

Mixture 1

Mixture 2

Estim. source 1

Estim. source 2

Further Reading

- C. M. Bishop, Neural Networks for Pattern Recognition. Clarendon Press, Oxford, 1995.

- C. W. Gardiner, Handbook of Sochastic Methods. Springer, Berlin, 1983.

- W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck and D. Warland, Spikes. MIT Press, Cambridge, 1998.