Statistical Data Analysis: Lecture 6

1	Probability, Bayes’ theorem
2	Random variables and probability densities
3	Expectation values, error propagation
4	Catalogue of pdfs
5	The Monte Carlo method
→ 6	**Statistical tests: general concepts**
7	Test statistics, multivariate methods
8	Goodness-of-fit tests
9	Parameter estimation, maximum likelihood
10	More maximum likelihood
11	Method of least squares
12	Interval estimation, setting limits
13	Nuisance parameters, systematic uncertainties
14	Examples of Bayesian approach

Statistical tests (in a particle physics context)

Suppose the result of a measurement for an individual event is a collection of numbers $\vec{x} = (x_1, \ldots, x_n)$

x_1 = number of muons,

x_2 = mean p_t of jets,

x_3 = missing energy, ...

$\vec{x}$ follows some n-dimensional joint pdf, which depends on the type of event produced, i.e., was it

$$\mathrm{pp} \rightarrow t\bar{t}\ , \qquad \mathrm{pp} \rightarrow \tilde{g}\tilde{g}\ , \ldots$$

For each reaction we consider we will have a hypothesis for the pdf of $\vec{x}$, e.g., $f(\vec{x}|H_0)$, $f(\vec{x}|H_1)$, etc.

Often call H_0 the signal hypothesis (the event type we want); H_1, H_2, ... are background hypotheses.

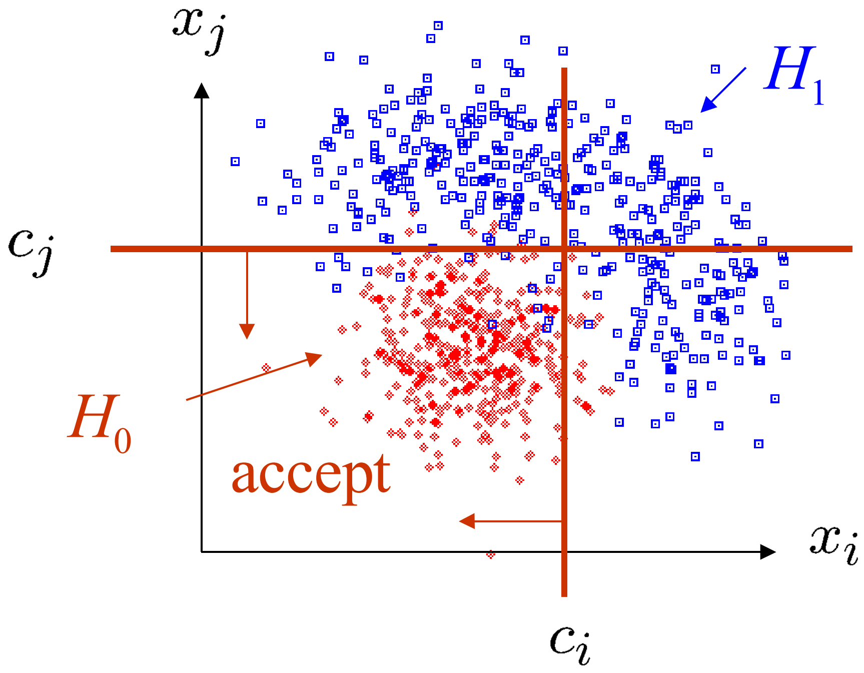
Selecting events

Suppose we have a data sample with two kinds of events, corresponding to hypotheses H_0 and H_1 and we want to select those of type H_0.

Each event is a point in $\vec{x}$ space. What 'decision boundary' should we use to accept/reject events as belonging to event type H_0?

Perhaps select events with 'cuts':

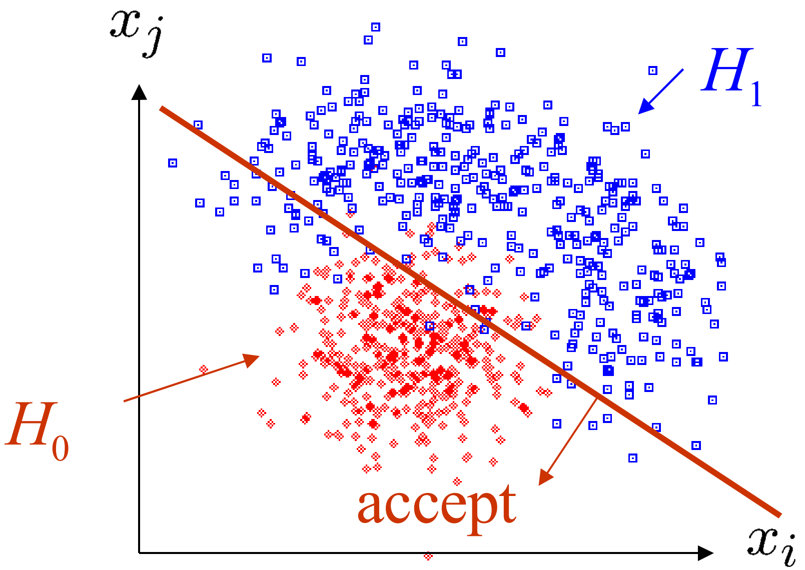
$$x_i < c_i$$
$$x_j < c_j$$



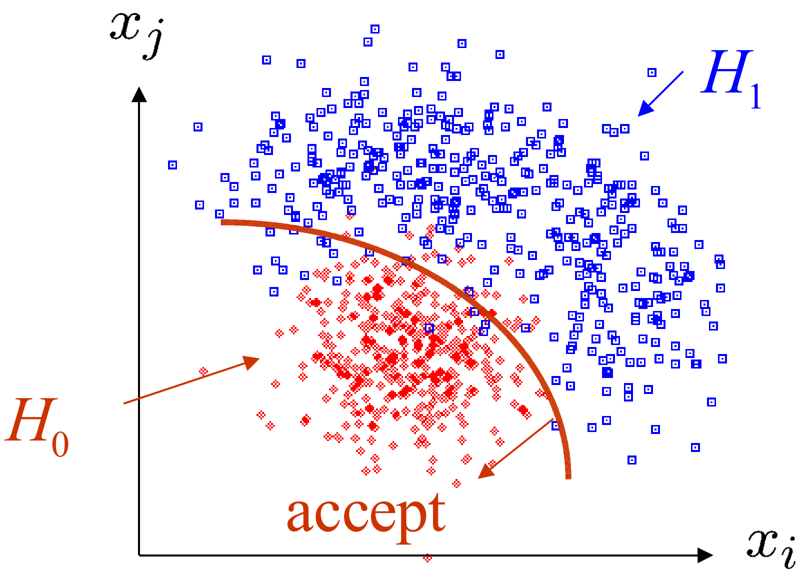
Other ways to select events

Or maybe use some other sort of decision boundary:

linear



or nonlinear



How can we do this in an ‘optimal’ way?

What are the difficulties in a high-dimensional space?

Test statistics

Construct a 'test statistic' of lower dimension (e.g. scalar)

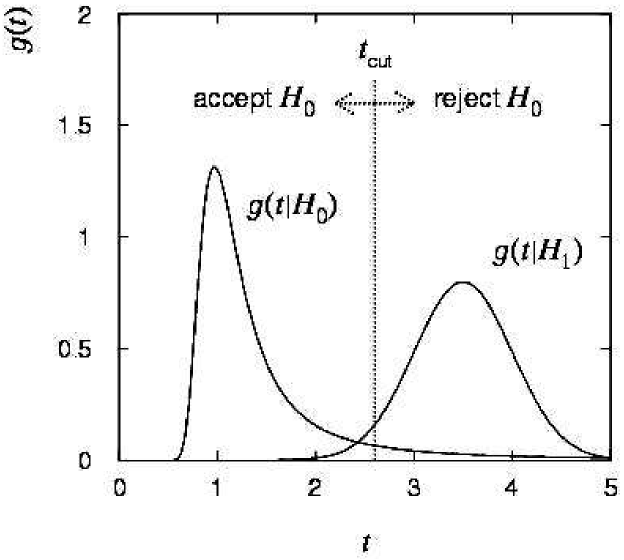
$$t(x_1, \ldots, x_n)$$

Goal is to compactify data without losing ability to discriminate between hypotheses.

We can work out the pdfs $g(t|H_0),\; g(t|H_1),\; \ldots$

Decision boundary is now a single 'cut' on *t*.

This effectively divides the sample space into two regions, where we accept or reject H_0.

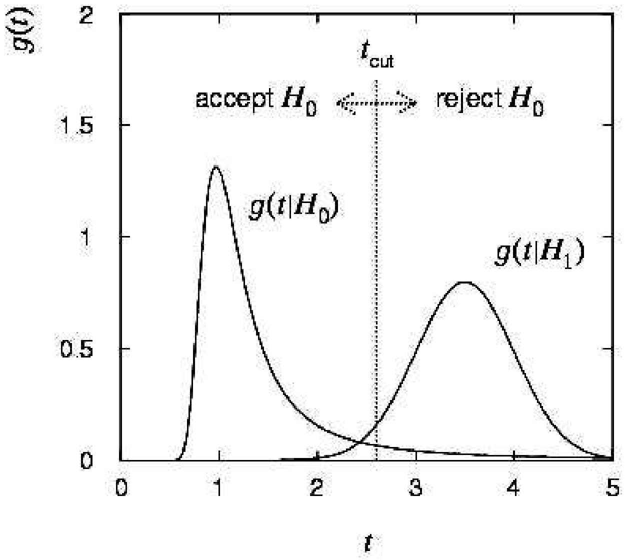


Significance level and power of a test

Probability to reject H_0 if it is true (error of the 1st kind):

$$\alpha = \int_{t_{\rm cut}}^{\infty} g(t|H_0)\,dt$$

(significance level)



Probability to accept H_0 if H_1 is true (error of the 2nd kind):

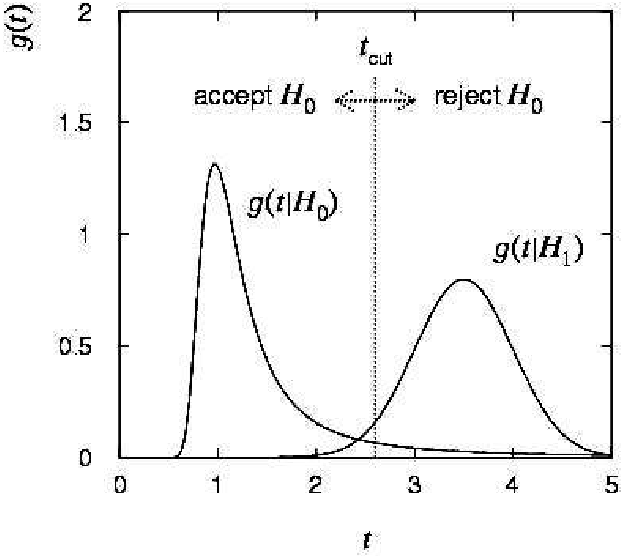
$$\beta = \int_{-\infty}^{t_{\rm cut}} g(t|H_1)\,dt$$

($1 - \beta =$ power)

Efficiency of event selection

Probability to accept an event which is signal (signal efficiency):

$$\varepsilon_{\rm s} = \int_{-\infty}^{t_{\rm cut}} g(t|{\rm s})\, dt = 1 - \alpha$$



Probability to accept an event which is background (background efficiency):

$$\varepsilon_{\rm b} = \int_{-\infty}^{t_{\rm cut}} g(t|{\rm b})\, dt = \beta$$

Purity of event selection

Suppose only one background type b; overall fractions of signal and background events are π_s and π_b (prior probabilities).

Suppose we select events with $t < t_{cut}$. What is the 'purity' of our selected sample?

Here purity means the probability to be signal given that the event was accepted. Using Bayes' theorem we find:

$$
\begin{aligned}
P(\mathrm{s}|t < t_{\mathrm{cut}}) &= \frac{P(t < t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}}}{P(t < t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}} + P(t < t_{\mathrm{cut}}|\mathrm{b})\pi_{\mathrm{b}}} \\
&= \frac{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}}}{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}} + \varepsilon_{\mathrm{b}}\pi_{\mathrm{b}}}
\end{aligned}
$$

So the purity depends on the prior probabilities as well as on the signal and background efficiencies.

Constructing a test statistic

How can we select events in an 'optimal way'?

Neyman-Pearson lemma (proof in Brandt Ch. 8) states:

To get the lowest ε_b for a given ε_s (highest power for a given significance level), choose acceptance region such that

$$\frac{f(\vec{x}|\mathrm{s})}{f(\vec{x}|\mathrm{b})} > c$$

where c is a constant which determines ε_s.

Equivalently, optimal scalar test statistic is

$$\boxed{t(\vec{x}) = \frac{f(\vec{x}|\mathrm{s})}{f(\vec{x}|\mathrm{b})}}$$

N.B. any monotonic function of this is just as good.

Purity vs. efficiency — optimal trade-off

Consider selecting n events:

expected numbers s from signal, b from background;

→ $n \sim$ Poisson $(s + b)$

Suppose b is known and goal is to estimate s with minimum relative statistical error.

Take as estimator: $\hat{s} = n - b$.

Variance of Poisson variable equals its mean, therefore

$$V[\hat{s}] = V[n - b] = V[n] = s + b \quad \rightarrow \quad \frac{\sigma_{\hat{s}}}{s} = \frac{\sqrt{s+b}}{s}$$

So we should maximize $\dfrac{s}{\sqrt{s+b}}$ (or $\varepsilon_s/\sqrt{b}$ if $s \ll b$),

equivalent to maximizing product of signal efficiency × purity.

Why Neyman-Pearson doesn’t always help

The problem is that we usually don’t have explicit formulae for the pdfs $f(\vec{x}|\mathsf{s}), \; f(\vec{x}|\mathsf{b})$.

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data, and enter each event into an *n*-dimensional histogram.

Use e.g. *M* bins for each of the *n* dimensions, total of M^n cells.

But *n* is potentially large, → prohibitively large number of cells to populate with Monte Carlo data.

Compromise: make Ansatz for form of test statistic $t(\vec{x})$ with fewer parameters; determine them (e.g. using MC) to give best discrimination between signal and background.

Multivariate methods

Many new (and some old) methods:

- Fisher discriminant
- Neural networks
- Kernel density methods
- Support Vector Machines
- Decision trees
  - Boosting
  - Bagging

New software for HEP, e.g.,

`TMVA` , Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

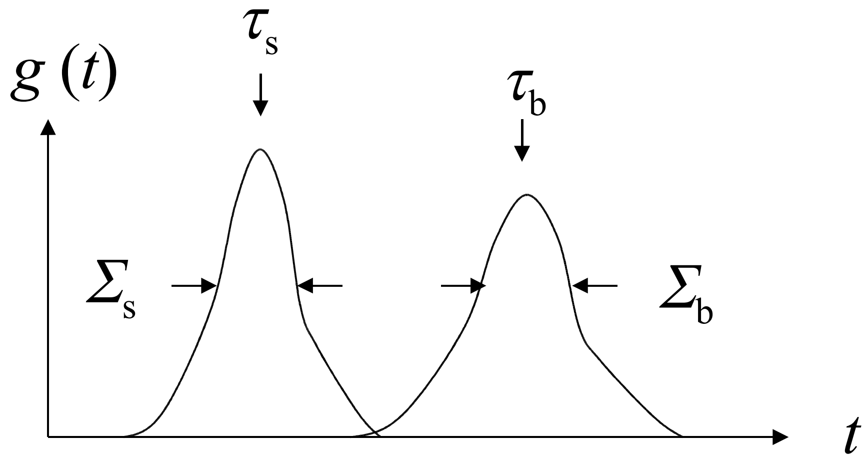
`StatPatternRecognition`, I. Narsky, physics/0507143

Linear test statistic

Ansatz: $$t(\vec{x}) = \sum_{i=1}^{n} a_i x_i$$

Choose the parameters $a_1, \ldots, a_n$ so that the pdfs $g(t|\mathrm{s}),\ g(t|\mathrm{b})$ have maximum 'separation'. We want:

large distance between mean values, small widths



→ Fisher: maximize $$J(\vec{a}) = \frac{(\tau_{\mathrm{s}} - \tau_{\mathrm{b}})^2}{\Sigma_{\mathrm{s}}^2 + \Sigma_{\mathrm{b}}^2}$$

Determining coefficients for maximum separation

We have

$$(\mu_k)_i = \int x_i f(\vec{x}|H_k)\, d\vec{x}$$

$$(V_k)_{ij} = \int (x-\mu_k)_i (x-\mu_k)_j f(\vec{x}|H_k)\, d\vec{x}$$

where

$$k = 0, 1 \quad \text{(hypothesis)}$$

$$i, j = 1, \ldots, n \quad \text{(component of } \vec{x}\text{)}.$$

In terms of mean and variance of $t(\vec{x})$ this becomes

$$\tau_k = \int t(\vec{x}) f(\vec{x}|H_k) d\vec{x} = \vec{a}^T \vec{\mu}_k \, ,$$

$$\Sigma_k^2 = \int (t(\vec{x}) - \tau_k)^2 f(\vec{x}|H_k)\, d\vec{x} = \vec{a}^T V_k \vec{a} \, .$$

Determining the coefficients (2)

The numerator of *J*(***a***) is

$$(\tau_0 - \tau_1)^2 = \sum_{i,j=1}^{n} a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j$$

‘between’ classes

$$= \sum_{i,j=1}^{n} a_i a_j B_{ij} = \vec{a}^T B \vec{a} \; ,$$

‘within’ classes

and the denominator is

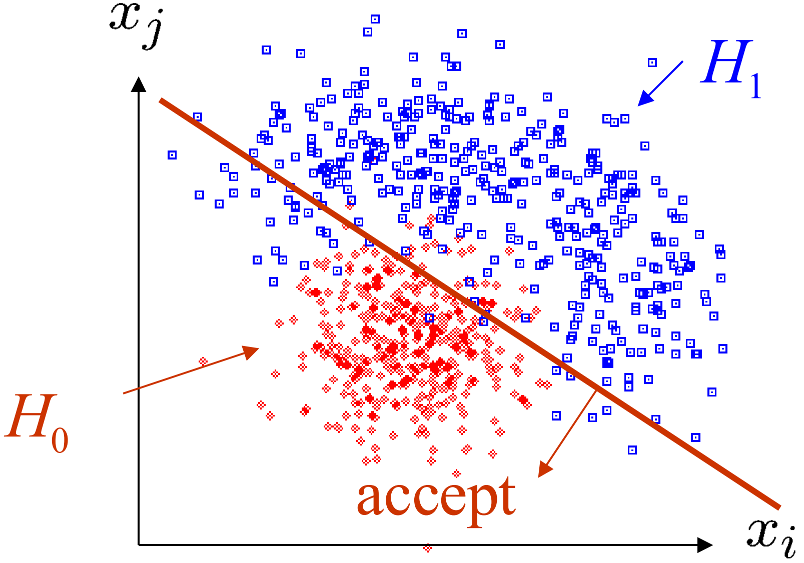
$$\Sigma_0^2 + \Sigma_1^2 = \sum_{i,j=1}^{n} a_i a_j (V_0 + V_1)_{ij} = \vec{a}^T W \vec{a}$$

→ maximize $J(\vec{a}) = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}} = \frac{\text{separation between classes}}{\text{separation within classes}}$

Fisher discriminant

Setting $\frac{\partial J}{\partial a_i} = 0$ gives Fisher's linear discriminant function:

$$\boxed{t(\vec{x}) = \vec{a}^T \vec{x}, \qquad \text{with } \vec{a} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)}$$



Corresponds to a linear decision boundary.

Fisher discriminant: comment on least squares

We obtain equivalent separation between hypotheses if we multiply the a_i by a common scale factor and add an arbitrary offset a_0:

$$t(\vec{x}) = a_0 + \sum_{i=1}^{n} a_i x_i$$

Thus we can fix the mean values τ_0 and τ_1 under the null and alternative hypotheses to arbitrary values, e.g., 0 and 1.

Then maximizing $J(\vec{a}) = (\tau_0 - \tau_1)^2/(\Sigma_0^2 + \Sigma_1^2)$
is equivalent to minimizing

$$\Sigma_0^2 + \Sigma_1^2 = E_0[(t - \tau_0)^2] + E_1[(t - \tau_1)^2]$$

Maximizing
Fisher's $J(\boldsymbol{a})$
→ 'least squares'

In practice, expectation values replaced by averages using samples of training data, e.g., from Monte Carlo models.

Fisher discriminant for Gaussian data

Suppose $f(\vec{x}|H_k)$ is multivariate Gaussian with mean values

$$E_0[\vec{x}] = \vec{\mu}_0 \text{ for } H_0\ , \qquad E_1[\vec{x}] = \vec{\mu}_1 \text{ for } H_1\ ,$$

and covariance matrices $V_0 = V_1 = V$ for both. We can write the Fisher discriminant (with an offset) as

$$t(\vec{x}) = a_0 + (\vec{\mu}_0 - \vec{\mu}_1)^T V^{-1} \vec{x}\ .$$

Then the likelihood ratio becomes

$$\begin{aligned} r &= \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} \\ &= \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})_0^T V^{-1}(\vec{x}-\vec{\mu}_0) + \frac{1}{2}(\vec{x}-\vec{\mu})_1^T V^{-1}(\vec{x}-\vec{\mu}_1) \right] \\ &\propto e^t \end{aligned}$$

Fisher discriminant for Gaussian data (2)

That is, $t \propto \ln r + \text{const.}$ (monotonic) so for this case, the Fisher discriminant is equivalent to using the likelihood ratio, and thus gives maximum purity for a given efficiency.

For non-Gaussian data this no longer holds, but linear discriminant function may be simplest practical solution.

Often try to transform data so as to better approximate Gaussian before constructing Fisher discrimimant.

Fisher discriminant and Gaussian data (3)

Multivariate Gaussian data with equal covariance matrices also gives a simple expression for posterior probabilities, e.g.,

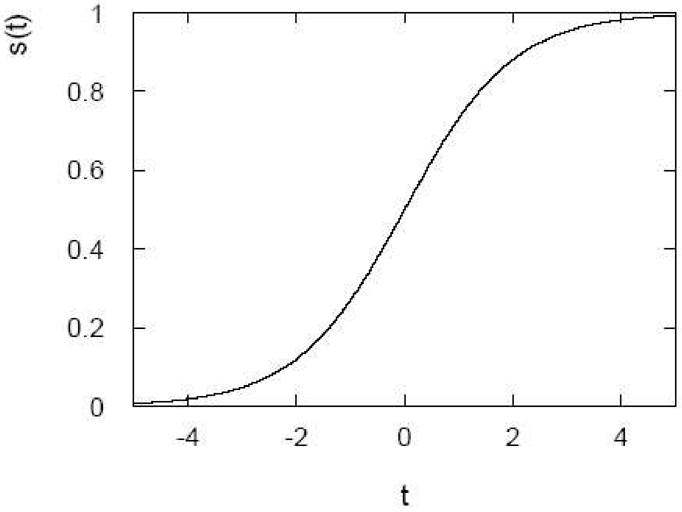
$$P(H_0|\vec{x}) = \frac{f(\vec{x}|H_0)\pi_0}{f(\vec{x}|H_0)\pi_0 + f(\vec{x}|H_1)\pi_1} = \frac{1}{1 + \frac{\pi_1}{\pi_0 r}} \ .$$

For a particular choice of the offset a_0 this can be written:

$$P(H_0|\vec{x}) = \frac{1}{1 + e^{-t}} \equiv s(t) \ ,$$

which is the logistic sigmoid function:

(We will use this later in connection with Neural Networks.)



Wrapping up lecture 6

We looked at statistical tests and related issues:

discriminate between event types (hypotheses),
determine selection efficiency, sample purity, etc.

We discussed a method to construct a test statistic using a linear function of the data:

Fisher discriminant

Next we will discuss nonlinear test variables such as

neural networks