Final part of the course

Statistical data Analysis

A.A. 2023-2024

Prof. Alexis Pompili (UniBA) / alexis.pompili@ba.infn.it

Content of this part: - Least Square Fit. Minimum χ^2 and its connection with ML fit. Goodness-of-fit.
- Extraction of a physical signal. Neyman-Pearson Lemma and Likelihood ratio.

Minimum χ^2 fit, its connection with ML fit & Goodness-of-fit

Least Squared Method – I

In the **Least Squares Method** (Metodo dei Minimi Quadrati) consider n measurements (of the type $y_i \pm \sigma_i$) corresponding to values x_i of the variable x. Assume we have a **model** for the dependence of y on the variable x given by a **function**:

$y = f(x; \vec{\theta})$ where $\vec{\theta} = (\theta_1, \dots, \theta_m)$ is a set of unknown parameters [see an example in next slide]

IF the measurements y_i are, each, **distributed** around the value $y = f(x_i; \vec{\theta})$ **according to a Gaussian with st. dev.** σ_i, the **likelihood function** for this problem can be written as a product of n Gaussian PDFs:

$$L(\vec{y}; \vec{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\left(\frac{y_i - f(x_i;\vec{\theta})}{\sigma_i}\right)^2}$$

residuals ≡ r

Maximizing $L(\vec{y}; \vec{\theta})$ is equivalent to minimize $-2\ln L(\vec{y}; \vec{\theta})$:

$$-2lnL(\vec{y}; \vec{\theta}) = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i; \vec{\theta})}{\sigma_i}\right)^2 + \sum_{i=1}^{n} ln2\pi\sigma_i^2$$

it is a χ^2 variable

just a constant if σ_i are **known & fixed**: **can be dropped** when minimizing

Thus, the quantity $-2\ln L(\vec{y}; \vec{\theta}) = \chi^2(\vec{\theta})$ is minimized.

NOTE: **this minimization is called Least Squares method** !

Least Squared Method – II

An example of fit performed with the **minimum χ^2 method** (within **ROOT**) can be borrowed by L.Lista's book (Fig. 5.5):

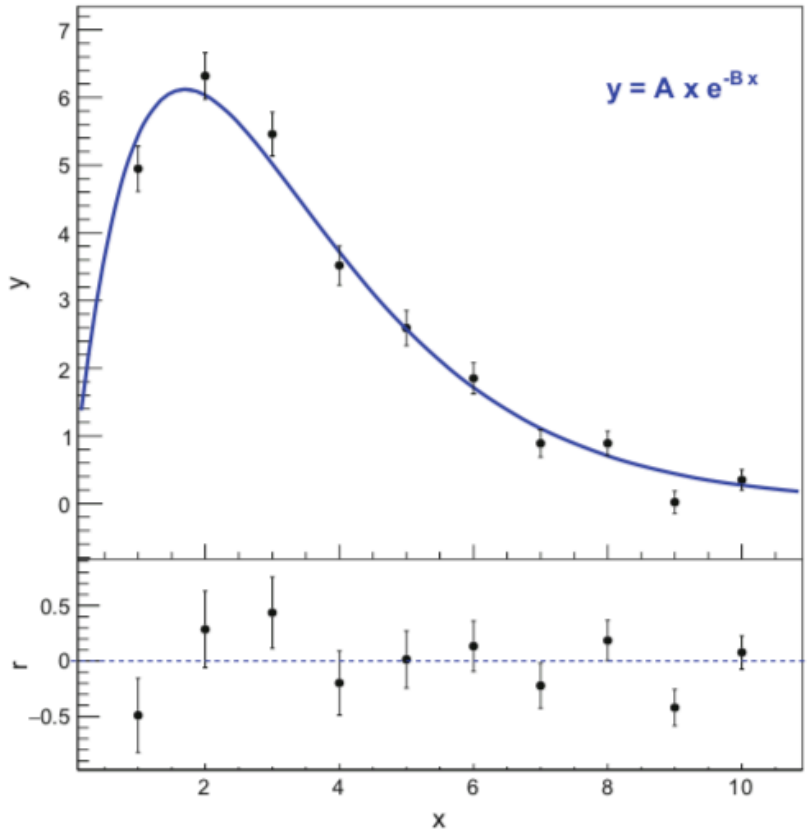


Fig. 5.5 Example of minimum χ^2 fit of a computer-generated dataset. The points with the error bars are used to fit a function model of the type $y = f(x) = A\,x\,e^{-Bx}$, where A and B are unknown parameters determined by the fit. The fit curve is superimposed as *solid blue line*. Residuals are shown in the bottom section of the plot

Residuals are randomly distributed around zero IF the data are distributed according to the assumed model $y = \mathrm{f(x;A,B)} = \mathrm{Ax}e^{-Bx}$

NOTE: in the simplest case of a **linear function** $y = \mathrm{A} + \mathrm{Bx}$ the minimum χ^2 problem can be solved analitically (L.Lista's book, section 5.12.1) [**linear regression**]

Minimum χ^2 method for Binned Data (histograms)

The situation just considered has wide similarities with the case of **binning a distribution of a random variable** when a large number of repeated measurements of this r.v. is available.
In this case the binning choice is natural because computing an unbinned likelihood function may become unpractical (since intensive computing power is needed and machine precision may also become an issue).
By **binning the distribution of the r.v. of interest** and **taking care to choose a number of bins N much smaller than the number of measurements n_i $(i = 1, \dots, N)$ for each i-bin**, in order to ensure an enough large n_i and thus a good Gaussian approximation for the Poisson distribution that would in principle describe the number of entries in a bin, …
… we are in the case in which, dropping again the constant term(s), we can write $-2\ln L$ as:

$$-2lnL(n_i; \mu_i(\vec{\theta})) = \sum_{i=1}^{n} \frac{\left(n_i - \mu_i(\vec{\theta})\right)^2}{\cancel{\mu_i} n_i} \equiv \chi^2_{\cancel{Pearson}\ Neyman}$$

… by having substituted, in the previous expression $-2lnL(\vec{y}; \vec{\theta}) = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i; \vec{\theta})}{\sigma_i}\right)^2$,

- y_i with the observed # of entries n_i
- $f(x_i; \vec{\theta})$ with the expected # of entries $\mu_i(\vec{\theta})$
- the variance σ_i^2 with the ~~expected~~ observed # of entries ~~μ_i~~ n_i

(Poisson)

exchange possible for large n_i

Note that $\mu_i(\vec{\theta}) = \int_{x_i^{LOW}}^{x_i^{UP}} f(x; \vec{\theta}) dx$

… and that **IF** the binning is enough fine: $\mu_i(\vec{\theta}) \cong f(x_i; \vec{\theta}) \delta x_i$ … with

- $x_i = \frac{x_i^{UP} - x_i^{LOW}}{2}$
- $\delta x_i = x_i^{UP} - x_i^{LOW}$

χ^2 and Goodness of Fit - I

One advantage of the **minimum χ^2 method** is that the expected distribution of the minimum χ^2 value (denoted by $\hat{\chi}^2$) is known and is given by the χ^2 distribution with a # of degrees of freedom equal to the # of measurements n minus the # of fit parameters m.

This is a general property. In the "histogram case" that we are considering, …

… the # of degrees of freedom is equal to the # of bins N minus the # of fit parameters k: **n.d.o.f. $= N - k$**

The minimum χ^2 value (denoted by $\hat{\chi}^2$) can be used as a **measurement/quantification of the *goodness of fit* (GOF)**.

Let's argue this important statement by introducing first the concept of **p-value**.

$p-$value : **probability that a χ^2 greater or equal to the minimum value $\hat{\chi}^2$ is obtained from a random fit according to the assumed model**

If data follow the assumed Gaussian distributions, the $p-$value is expected to be a r.v. uniformly distributed from 0 to 1!

This comes from a general property of cumulative distributions (see next slide [#]).

Obtaining a small $p-$value of the fit can be a symptom of a poor description of the assumed theoretical model in the fit.

For this reason, the minimum χ^2 value ($\hat{\chi}^2$) can be used as a measurement of the goodness of fit.

Practically it is a matter of setting a threshold to determine whether or not a fit can be considered acceptable or not; for istance a $p-\text{value}>0.05$ will discard on average 5% of the cases (due to the possibility of statitical fluctuations)

χ^2 and Goodness of Fit - II

[#] Given a PDF, $f(x)$, its cumulative distribution is defined as: $\mathrm{F}(x) = \int_{-\infty}^{x} f(x')dx'$

In particular: $\lim_{x\to-\infty} \mathrm{F}(x) = \int_{-\infty}^{-\infty} f(x')dx' = 0$ & $\lim_{x\to+\infty} \mathrm{F}(x) = \int_{-\infty}^{+\infty} f(x')dx' = 1$

It can be easily demonstrated that the PDF $\mathrm{P(y)}$ of the transformed variable $\mathrm{y} \equiv \mathrm{F}(x)$ is uniform between 0 and 1:

$$\frac{d\mathrm{P(y)}}{dy} = \frac{d\mathrm{P(y)}}{dy}\frac{dx}{dx} = \frac{d\mathrm{P(y)}}{dx}\frac{dx}{dy} \equiv \frac{d\mathrm{P(y)}}{dx}\frac{dx}{dF(x)} = \frac{d\mathrm{P(y)}}{dx}\frac{1}{\frac{dF(x)}{dx}} = f(x)\frac{1}{f(x)} = 1 \quad \Rightarrow \quad \mathrm{P(y)} = \mathrm{cost}$$

Binned Poissonian Fits (histograms with small # of entries) – I

If the Gaussian approximation for the Poisson distribution does not hold because, in many of the N bins, n_i is not large enough … we are obliged to use a Poissonian model, that is of course valid for small # of entries.
In this case the negative log likelihood $-2\ln L$ can be written,

…instead of $-2lnL\left(n_i;\mu_i(\vec{\theta})\right) = -2ln\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi n_i}}e^{-\frac{1}{2}\left(\frac{n_i-\mu_i(\vec{\theta})}{\sqrt{n_i}}\right)^2}$ … as: $-2lnL\left(n_i;\mu_i(\vec{\theta})\right) = -2ln\prod_{i=1}^{N}\frac{e^{-\mu_i(\vec{\theta})}\mu_i(\vec{\theta})^{n_i}}{n_i!}$

Using the **approach proposed by Baker-Cousins** the likelihood can be divided by its maximum value which does not depend on the unknown parameters (rather it is based on their best estimates) and does not change the fit result.
In this way we deal with a negative log likelihood **ratio** (that we denote with λ):

$$\lambda = -2ln\frac{L\left(n_i;\mu_i(\vec{\theta})\right)}{\hat{L}(n_i;\mu_i)}$$

… that we can rewrite, with good approximation, exchanging μ_i with n_i : $\lambda = -2ln\frac{L\left(n_i;\mu_i(\vec{\theta})\right)}{L(n_i;n_i)}$

… and with the usual algebra:

$$\lambda = -2ln\frac{L\left(n_i;\mu_i(\vec{\theta})\right)}{L(n_i;n_i)} = -2ln\prod_{i=1}^{N}\frac{e^{-\mu_i(\vec{\theta})}\mu_i(\vec{\theta})^{n_i}}{n_i!}\cdot\frac{n_i!}{e^{-n_i}n_i^{n_i}} = -2\sum_{i=1}^{N}ln\frac{e^{-\mu_i(\vec{\theta})}\mu_i(\vec{\theta})^{n_i}}{\cancel{n_i!}}\cdot\frac{\cancel{n_i!}}{e^{-n_i}n_i^{n_i}}$$

$$= -2\sum_{i=1}^{N}ln\left[\frac{e^{-\mu_i(\vec{\theta})}}{e^{-n_i}}\cdot\left(\frac{\mu_i(\vec{\theta})}{n_i}\right)^{n_i}\right] = +2\sum_{i=1}^{N}\left[-ln(e^{-\mu_i(\vec{\theta})+n_i}) - n_i ln\left(\frac{\mu_i(\vec{\theta})}{n_i}\right)\right] = 2\sum_{i=1}^{N}\left[\mu_i(\vec{\theta}) - n_i + n_i ln\left(\frac{n_i}{\mu_i(\vec{\theta})}\right)\right]$$

Binned Poissonian Fits (histograms with small # of entries) – II

- Now, the important result derives from the **Wilks' theorem** (it deals with likelihood ratios; will be discussed in the in-depth part):

  If the model is correct … **the distribution of the minimum value of λ can be asymptotically approximated by a χ^2 distribution with a n.d.o.f. = (# bins - # fit parameters)!**

  For this reason, the neg-log-likelihood ratio λ is denoted as χ^2_λ in (5.71-5.72) equations of L.Lista's book.

  - The χ^2_λ can be used to determine a p-value that provides a measure of the goodness-of-fit, as previously discussed for the situation in which Gaussian approximation is valid and the neg-log-likelihood is a "genuine" χ^2.

- However, the asymptotic approximation will not hold **if** the # of measurements is **not** sufficiently large and, consequently, **the distribution of χ^2_λ will deviate from a χ^2 distribution**.

  - In this cases **the distribution can still be determined by generating a sufficiently number of Monte Carlo pseudo-experiments** that reproduce the theoretical PDF (MC toys), and thus the p-value can be computed accordingly.

χ^2 and Goodness of Fit - III / comparison with ML fits

It must be noted that :

Unlike minimum χ^2 fits, in general, for Maximum Likelihood fits the value of $-2\ln L$ for which the likelihood function is maximized does not provide a measurement of the goodness of the fit.

What can be still done in this case ?

1) IF the ML fit is **binned** it is possible to calculate both an overall **normalized χ^2** and **bin-by-bin pulls**

2) IF the ML fit is genuinely **unbinned** you can still bin a posteriori (after the fit) and proceed as in **(1)**

3a) you can still distinguish which model (implemented in the PDF) is the best among several by considering the minimum among the minimum values of $-2\ln L$ for each model

3b) It is possible to obtain in some cases a goodness-of-fit measurement by finding the ratio of the likelihood functions evaluated in two different hypotheses since Wilks' theorem ensures that … a likelihood ratio, under some conditions that hold in particular circumstances, is asymptotically distributed as a χ^2 for a large number of repeated measurements.

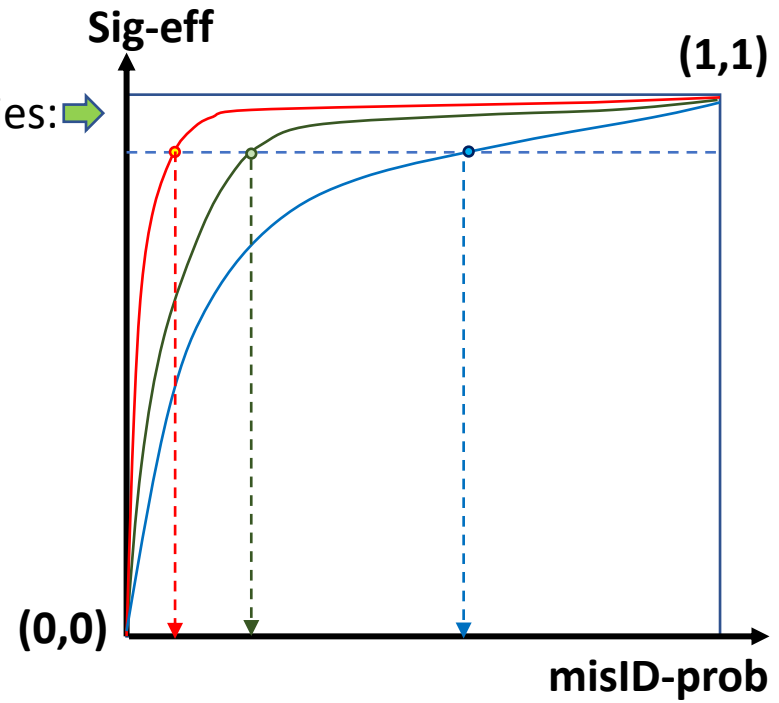
EXTRACTION of a physical SIGNAL

Neyman-Pearson Lemma & Likelihood Ratio - I

As we have discussed in a previous part of the course devoted to hypothesis testing and the ROC curve, the extraction of a physical signal is obtained by applying a **set of selection criteria** based on some variables/observables (a single variable - or a combination of variables - is called ***test statistic***) to select signal while rejecting background(s). The selection algorithm based upon a specific set can be represented by a ROC curve in the plane representing the **signal efficiency** against the **contamination level** (**misidentification probability** or probability of background's survival).

The performance of a selection criterion can be considered ***optimal*** if it achieves the smallest misidentification probability for a desired/target value of the selection efficiency.

Suppose having different test statistics and the corresponding different ROC curves; for a given signal efficiency the curves provide different misidentification probabilities:



According to the **Neyman-Pearson lemma**:

the **optimal test statitistic** is given by the **ratio of the likelihood functions** $L(\vec{x}|H_1)$ and $L(\vec{x}|H_0)$ evaluated for the observed data sample $\vec{x}$ under the two hypotheses H_1 & H_0 :

$$\lambda(\vec{x}) = \frac{L(\vec{x}|H_1)}{L(\vec{x}|H_0)}$$

In what sense it is optimal?

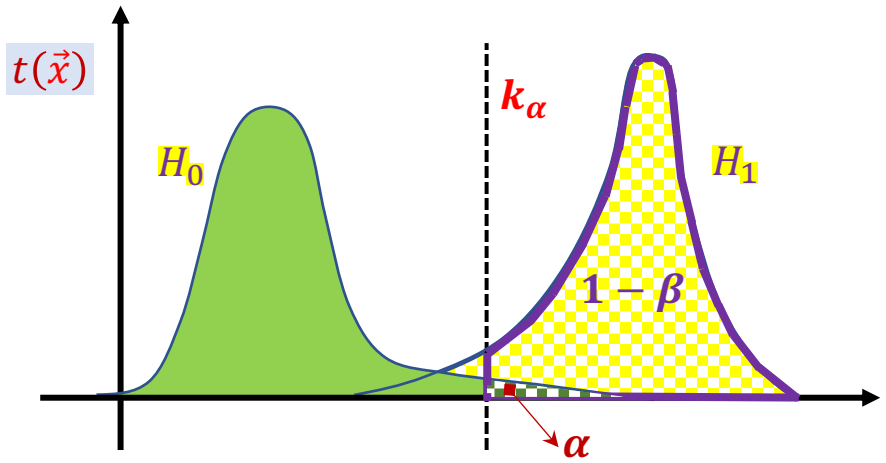
Neyman-Pearson Lemma & Likelihood Ratio - II

The likelihood ratio as test statistic is ***optimal*** in the sense that… **for a fixed background misidentification probability α, the selection that corresponds to the largest possible signal selection efficiency 1-β is given by:**

$$\lambda(\vec{x}) = \frac{L(\vec{x}|H_1)}{L(\vec{x}|H_0)} \geq k_\alpha$$

… where, by varying the "cut_value" k_α the required/targeted value of α may be achieved.

In other words… the likelihood ratio is the test statistic that ***optimally minimizes the overlap*** **between the two PDFs for the background and the signal hypotheses (H_0 and H_1).**



This Lemma provides the selection that achieves the optimal performances only if the joint multi-dimensional PDFs characterizing our problem are known! However in many realsitic cases it is not easy to determine the correct model and approximated solution are adopted, like numerical methods and Machine-Learning algorithms (Neural Networks or Boosted Decision Trees) that may find selections with performances close to the optimal limit given by the Lemma.

Neyman-Pearson Lemma & Likelihood Ratio - III

IF the variables $x_1, \dots, x_n$ that characterize our problem **are independent, the likelihood function can be factorized into the product of one-dimentional marginal PDFs:**

$$\lambda(\vec{x}) = \frac{L(\vec{x}|H_1)}{L(\vec{x}|H_0)} = \frac{\prod_{i=1}^{n} f_i(x_i|H_1)}{\prod_{i=1}^{n} f_i(x_i|H_0)}$$

In this case (namely **factorization holds), optimal selection performances are achieved, according to the Lemma !**

In real analysis life the variables we deal with are not independent (remember that uncorrelated variables are not necessarly independent!**), but still** the factorized expression can be used as discriminant even if performances will not be optimal anymore.

The quantity to be used as test statistic is the so-called **Projective Likelihood Ratio**:

$$\lambda(x_1, \dots, x_n) = \frac{\prod_{i=1}^{n} f_i(x_i|H_1)}{\prod_{i=1}^{n} f_i(x_i|H_0)}$$

Note: the marginalized PDFs can be obtained using Monte Carlo training samples
to produce distributions corresponding with enough good approximation to the marginal PDFs.

Multivariate-based selection with Machine Learning

- **The Neyman-Pearson Lemma sets up an *upper limit* to the performance of any possible selection, from those classical cut-based selections to the most recently introduced Machine-Learning algorithms which can often go close to the performance of an ideal selection based on the likelihood ratio.**

The most powerful approximate methods, implemented by means of computer algorithms, are organized as follows.
The algorithm receives as input a set of discriminant variables, each of which individually does not allow to reach an optimal selection power, and computes an output that combines the input variables.
The computation of the output value is based on an often very large set of parameters.
The discriminant output is taken as test statistic and is adopted to select the signal with the desired efficiency by means of a single cut on the value of the output.

An optimal choice of the parameters can allow to achieve the best possible performances.
The usual strategy consists in tuning the discriminant parameters providing as input to the algorithm large datasets distributed according either the H_0 and the H_1 hypotheses. Typically distributions according to background hypothesis are taken from real data, often using control samples, while signal-like distributions according to the H_1 hypothesis are derived by simulated data (Monte Carlo). By comparing the discirminant output to the true origin of the dataset the parameters are modified. This process is called *training* and the algorithms that use such kind of training samples are called supervised machine learning algorithms.
The typical problem of this process is called *overtraining* and it is depicted in Fig. 9.6 of L.Lista's book**.**