Statistical Methods for Neutrino Physics

VIII Int. Pontecorvo Neutrino Physics School

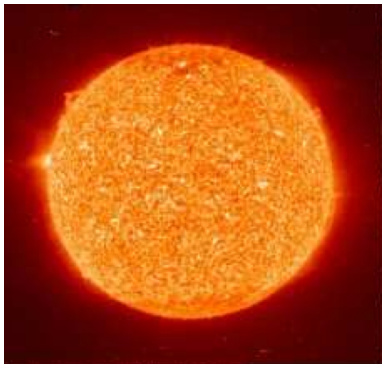
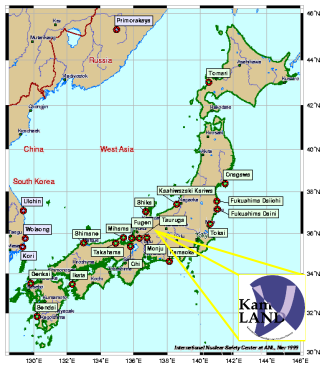
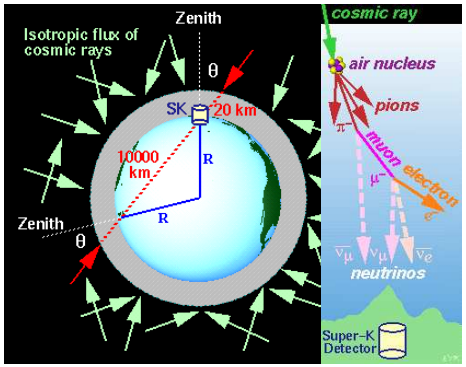
Thomas Schwetz

Karlsruher Institut für Technologie

1–10 Sept. 2019, Sinaia, Romania

Global data on neutrino oscillations

various neutrino sources and vastly different energy and distance scales:

sun	reactors	atmosphere	accelerators
			
Homestake,SAGE,GALLEX SuperK, SNO, Borexino	KamLAND, D-CHOOZ RENO, DayaBay	SuperKamiokande IceCube	MINOS, T2K, NOvA

- global data fits nicely with the 3 neutrinos from the SM
- "anomalies" (at 2-3 σ) which do not fit the 3-flavour picture: LSND, MiniBooNE, reactor anomaly, no LMA MSW up-turn of solar neutrino spectrum

3-flavour neutrino parameters

- 3 masses: Δm^2_{21}, Δm^2_{31}, m_0
- 3 mixing angels: θ_{12}, θ_{13}, θ_{23}
- 3 phases: 1 Dirac (δ), 2 Majorana (α_1, α_2)

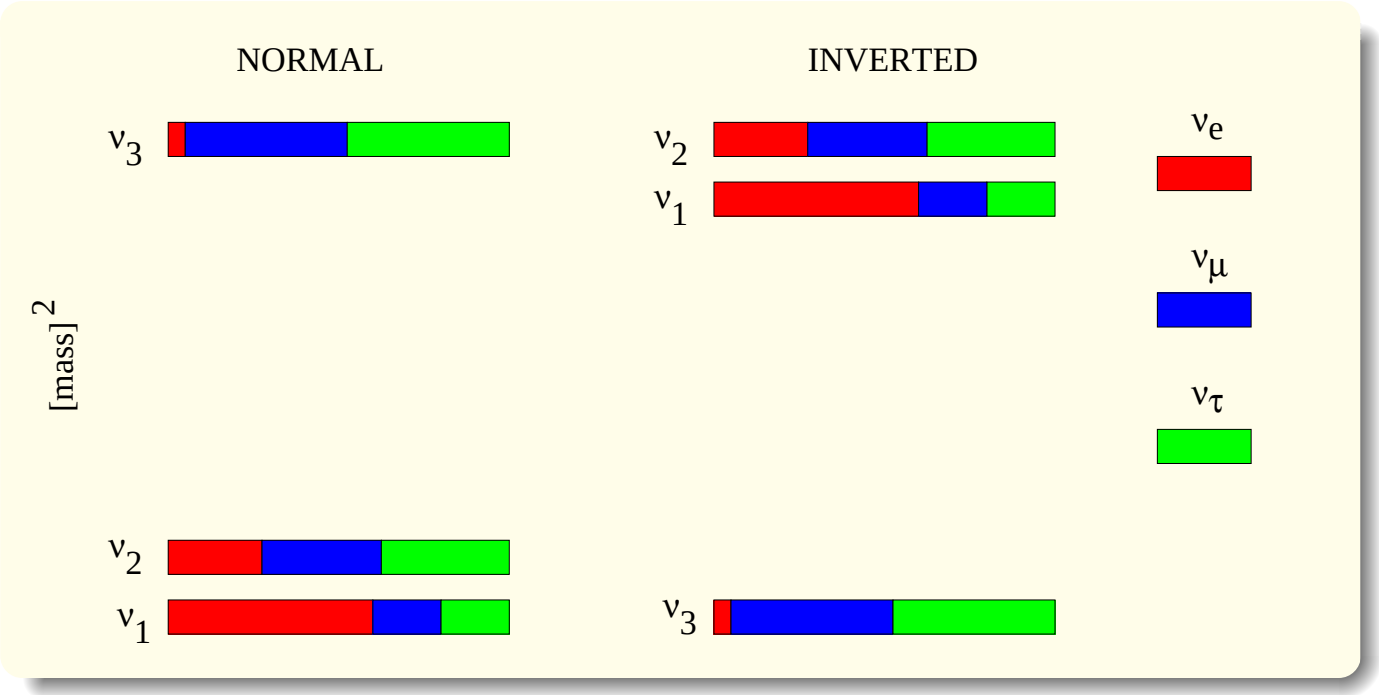
neutrino oscillations
absolute mass observables
lepton-number violation (neutrinoless double-beta decay)

3-flavour neutrino parameters

- 3 masses: Δm^2_{21}, Δm^2_{31}, m_0
- 3 mixing angels: θ_{12}, θ_{13}, θ_{23}
- 3 phases: 1 Dirac (δ), 2 Majorana (α_1, α_2)

neutrino oscillations
absolute mass observables
lepton-number violation (neutrinoless double-beta decay)

Neutrino mass states and mixing



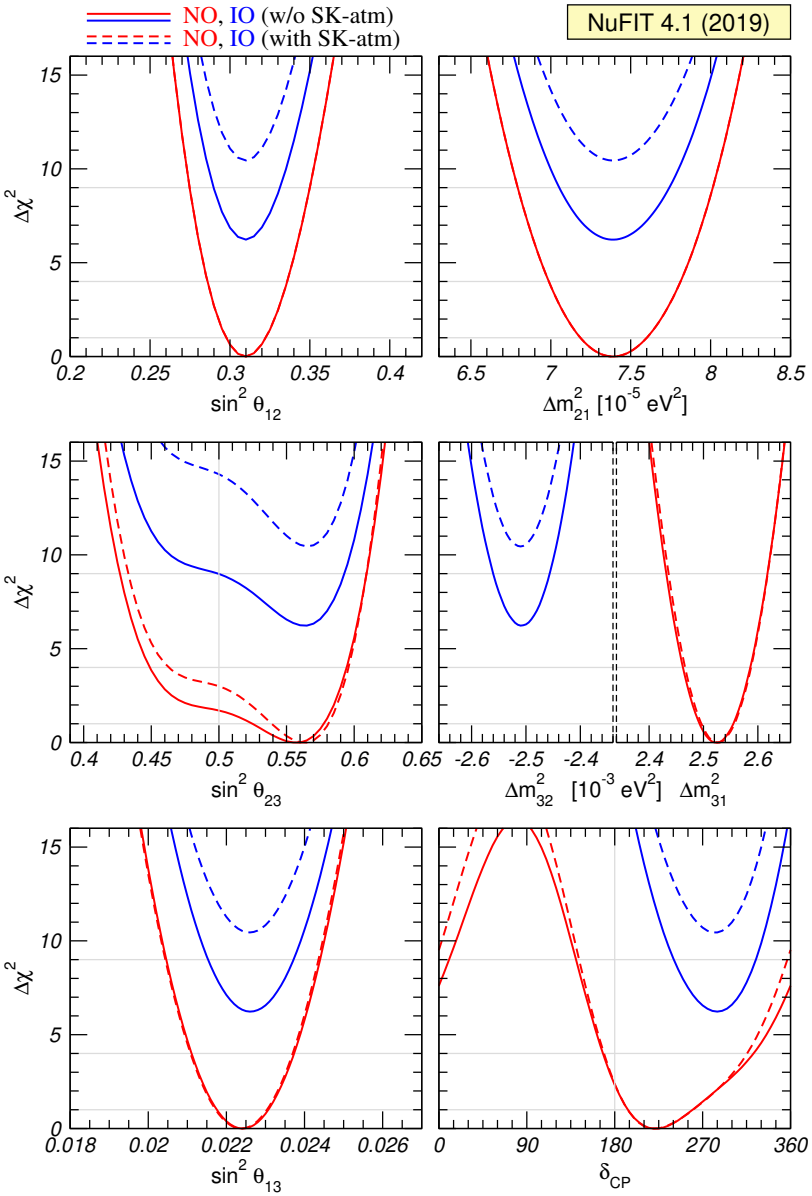
Three-flavour oscillation parameters

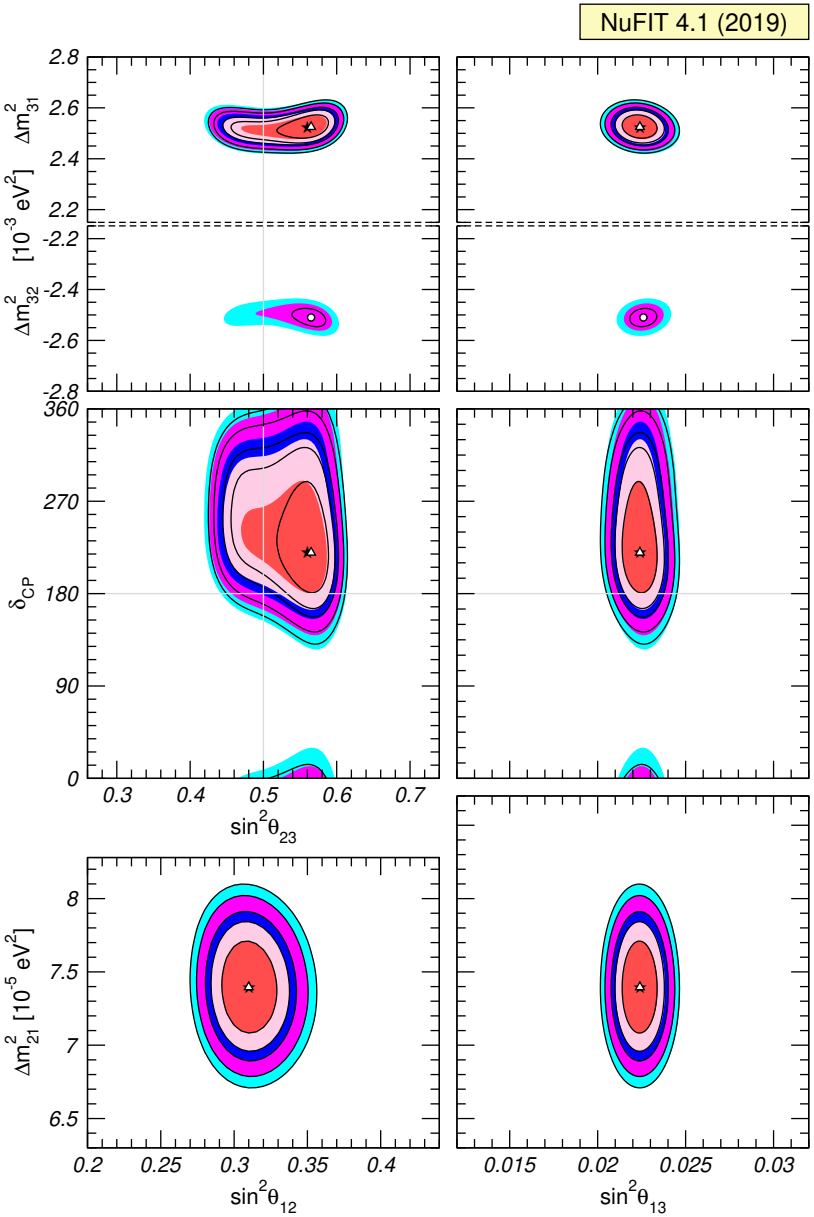
- each oscillation parameter is determined by several (classes of) experiments
- especially true for not-so-well determined parameters
- interplay of different data sets $\Rightarrow$ global analyses

Three-flavour oscillation parameters

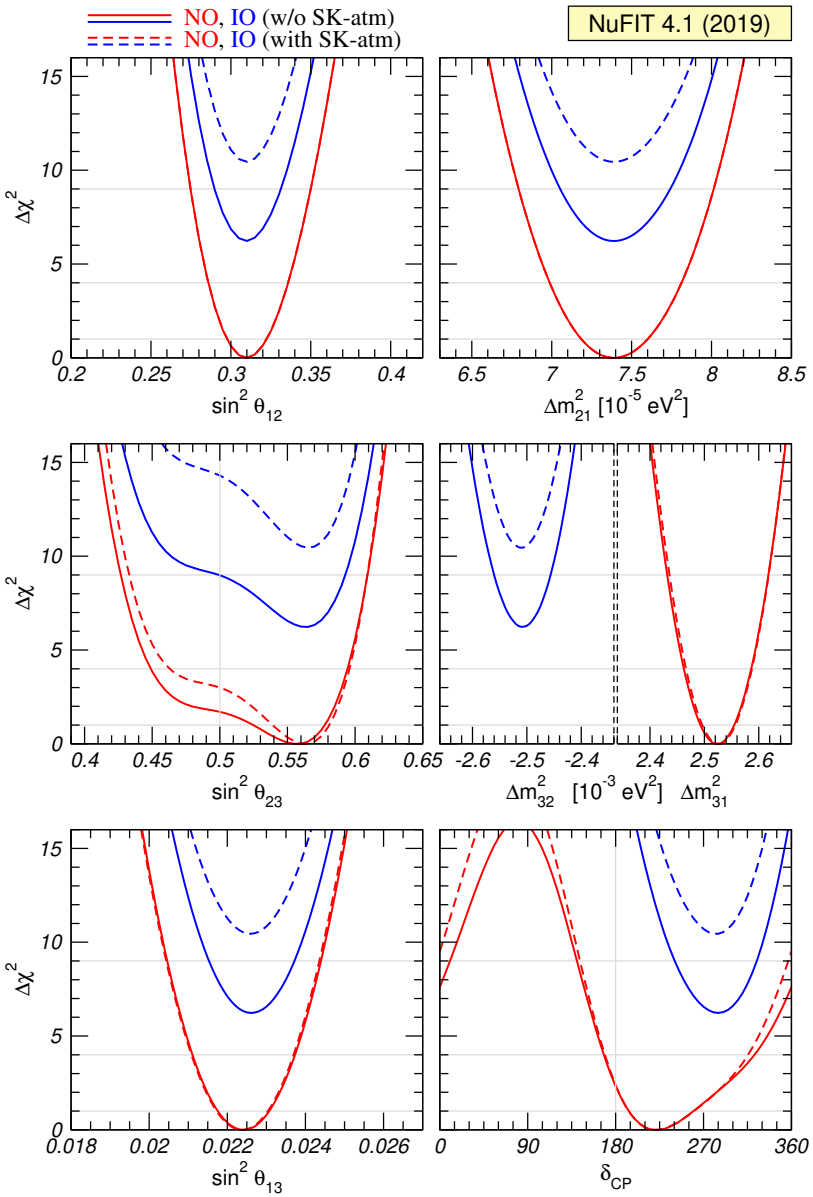
- each oscillation parameter is determined by several (classes of) experiments
- especially true for not-so-well determined parameters
- interplay of different data sets $\Rightarrow$ global analyses
- NuFit collaboration: www.nu-fit.org
  with M.C. Gonzalez-Garcia, M. Maltoni, et al.
- latest paper: Esteban, Gonzalez-Garcia, Hernandez-Cabezudo, Maltoni, Schwetz, JHEP 1901 (2019) 106 [1811.05487]
- latest version: 4.1 (as of July 2019)

Global 3-flavour fit





Global 3-flavour fit

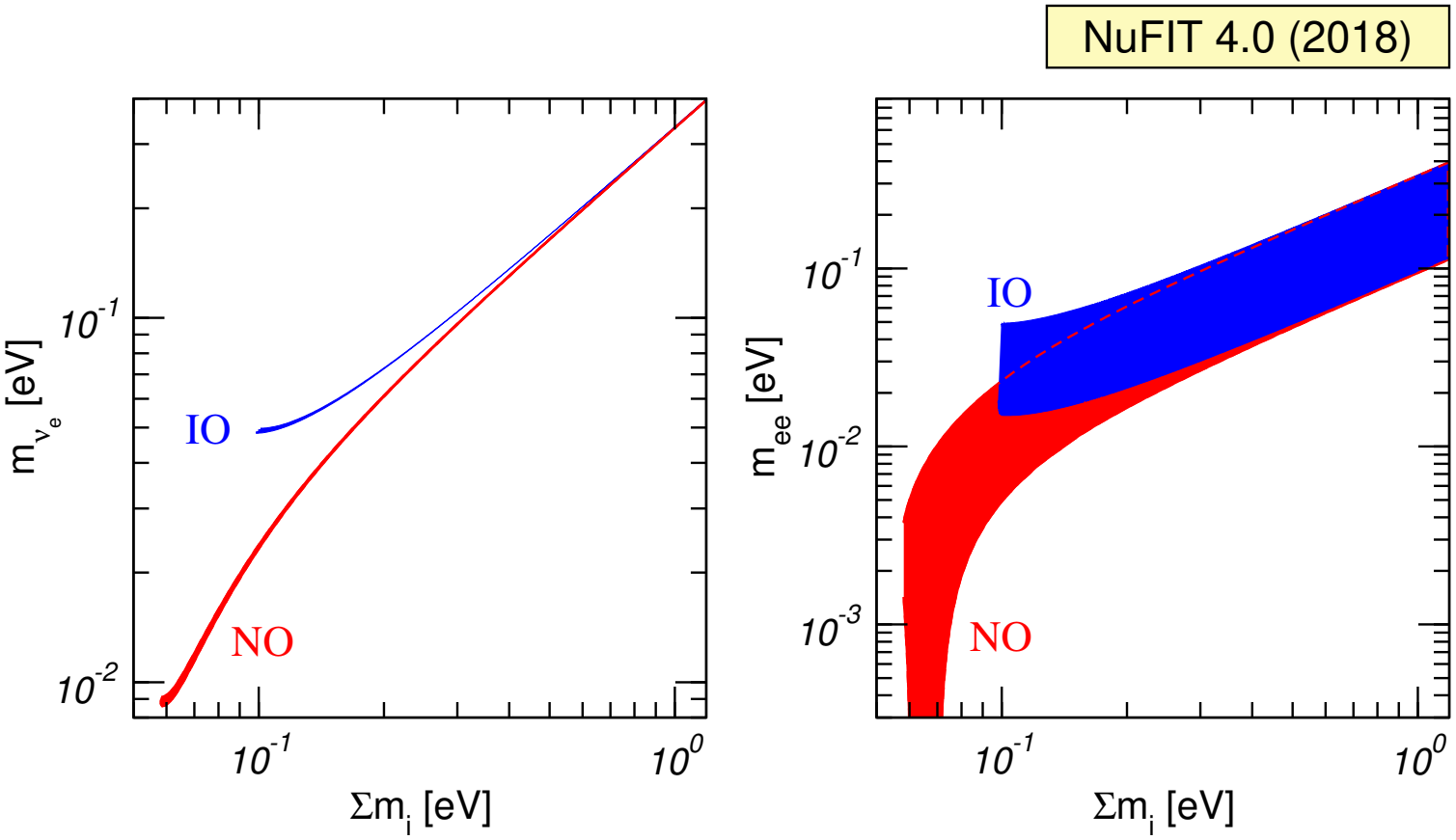


- robust determination (relat. precision at 3σ):

$$\theta_{12}\,(14\%) \quad , \quad \theta_{13}\,(8.9\%)$$
$$\Delta m^2_{21}\,(16\%)\,, \quad |\Delta m^2_{3\ell}|\,(7.6\%)$$

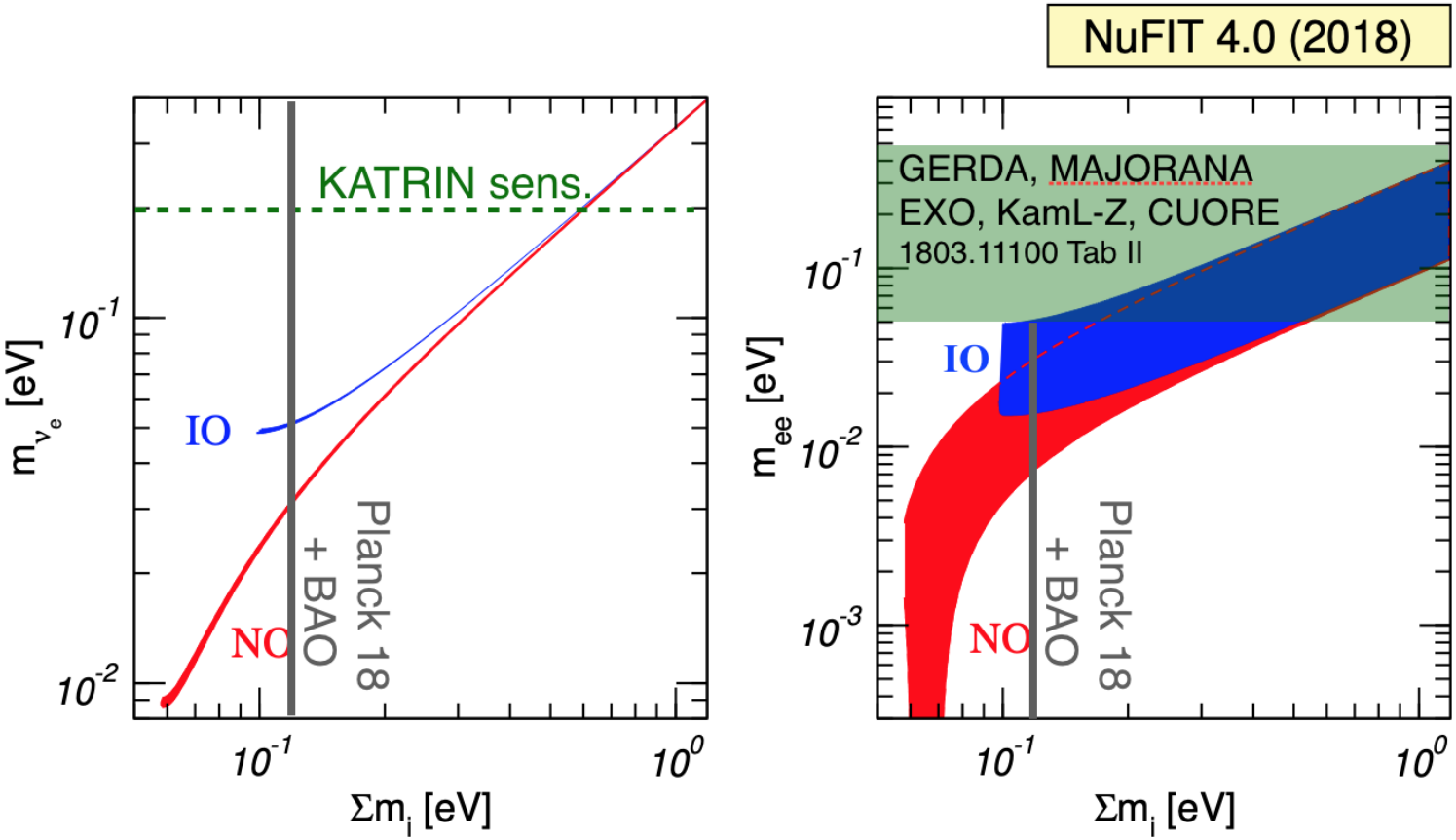
- broad allowed range for $\theta_{23}\,(24\%)$, non-significant indications for non-maximality/octant
- ambiguity in sign of $\Delta m^2_{3\ell} \rightarrow$ mass ordering (3.2σ preference for NO)
- preference for $180° \lesssim \delta_{\mathrm{CP}} \lesssim 360°$

Absolute neutrino mass



- Endpoint of beta spectrum: $m_\beta^2 = \sum_i |U_{ei}^2| m_i^2 = m_\beta^2(\Delta m_{i1}^2, \theta_{1i}, m_0)$
- Cosmology: $\sum = \sum_i m_i = \sum(\Delta m_{i1}^2, m_0)$
- Neutrinoless double beta-decay: $m_{ee} = |\sum_i U_{ei}^2 m_i| = m_{ee}(\Delta m_{i1}^2, \theta_{1i}, m_0, \alpha_i)$

Absolute neutrino mass



- Endpoint of beta spectrum: $m_\beta^2 = \sum_i |U_{ei}^2| m_i^2 = m_\beta^2(\Delta m_{i1}^2, \theta_{1i}, m_0)$
- Cosmology: $\sum = \sum_i m_i = \sum(\Delta m_{i1}^2, m_0)$
- Neutrinoless double beta-decay: $m_{ee} = |\sum_i U_{ei}^2 m_i| = m_{ee}(\Delta m_{i1}^2, \theta_{1i}, m_0, \alpha_i)$

These lectures:

Basic problems in statistics

- Parameter estimation
- Goodness of fit

Confidence intervals

- frequentist
- Bayesian intervals
- Parameter marginalization

Event rates in oscillation experiments

- Reactor experiments
- More complicated situations

Building the χ^2

- Systematical errors in χ^2 analyses

Hypothesis testing

- Frequentist
- Bayesian model selection

Outline

Basic problems in statistics

We have

- a set of observables x_i
- a model (theory) making predictions for those observables,
- and the model may depend on parameters: θ_α.

Now we want to address questions like the following:

- Does the model provide a "good" description of the data? ("model testing" or "goodness-of-fit")
- What are the parameter values $\hat{\theta}_\alpha$ that provide the best description of the data, assuming this model is correct? ("parameter estimation")
- Assuming the model is correct, what is the "acceptable" range for the parameters? ("acceptance regions" or "confidence intervalls")
- Suppose we have two different models (hypotheses), which one of the two gives a "better" description of the data? ("hypothesis testing")

Basic problems in statistics

We have

- a set of observables x_i
- a model (theory) making predictions for those observables,
- and the model may depend on parameters: θ_α.

Now we want to address questions like the following:

- Does the model provide a "good" description of the data? ("model testing" or "goodness-of-fit")
- What are the parameter values $\hat{\theta}_\alpha$ that provide the best description of the data, assuming this model is correct? ("parameter estimation")
- Assuming the model is correct, what is the "acceptable" range for the parameters? ("acceptance regions" or "confidence intervalls")
- Suppose we have two different models (hypotheses), which one of the two gives a "better" description of the data? ("hypothesis testing")

In statistics a "model" predicts the p.d.f. for the observables: $f(\vec{x};\vec{\theta})$
(in physics we often call already the mean value "prediction" and implicitly assume Gaussian or Poisson distribution)

Example

- given set of oscillation parameters θ:
- "predicted number" of ν_e appearance events in T2K $N(\theta)$
- read: the number of events is expected to be Poisson distributed with the Poisson mean given by $N(\theta)$

In statistics a "model" predicts the p.d.f. for the observables: $f(\vec{x};\vec{\theta})$
(in physics we often call already the mean value "prediction" and implicitly assume Gaussian or Poisson distribution)

Example

- given set of oscillation parameters θ:
- "predicted number" of ν_e appearance events in T2K $N(\theta)$
- read: the number of events is expected to be Poisson distributed with the Poisson mean given by $N(\theta)$

Statistic

A "statistic" is any function depending on random variables: $T(x_i)$

- We are free to consider any statistic to address those questions.
- for each of the questions from the previous slide we can use a different statistic
- in practice often "χ^2" (more precisely, a least-squares statistic) is used to address all of them (sometimes this leads to confusion)
- a very important statistic is the likelihood

Frequentist statistics

- the only random (statistical) quantities are data
- there is no way to assign a probablitiy to a model or its parameters
- a model parameter has an unknown but fixed true value

Example: the mass of an apple or the mass of the Higgs

- the apples in a shop will have some distribution in mass, we can assign a p.d.f. for the mass of an apple
- the mass of the Higgs is a fundamental parameter of the SM, it has a fixed (but unknown) value. (The same is true for the electron mass, neutrino mass, mixing angles...)

- in the frequentist approach we can make only probability statements about the outcome of an experiment (if it were repeated many times) under the hypothesis of a model

Frequentist statistics

- the only random (statistical) quantities are data
- there is no way to assign a probablitiy to a model or its parameters
- a model parameter has an unknown but fixed true value

Example: the mass of an apple or the mass of the Higgs

- the apples in a shop will have some distribution in mass, we can assign a p.d.f. for the mass of an apple
- the mass of the Higgs is a fundamental parameter of the SM, it has a fixed (but unknown) value. (The same is true for the electron mass, neutrino mass, mixing angles...)

- in the frequentist approach we can make only probability statements about the outcome of an experiment (if it were repeated many times) under the hypothesis of a model

Frequentist statistics

- the only random (statistical) quantities are data
- there is no way to assign a probablitiy to a model or its parameters
- a model parameter has an unknown but fixed true value

Example: the mass of an apple or the mass of the Higgs

- the apples in a shop will have some distribution in mass, we can assign a p.d.f. for the mass of an apple
- the mass of the Higgs is a fundamental parameter of the SM, it has a fixed (but unknown) value. (The same is true for the electron mass, neutrino mass, mixing angles...)

- in the frequentist approach we can make only probability statements about the outcome of an experiment (if it were repeated many times) under the hypothesis of a model

Bayesian inference

- consider the p.d.f. predicted in a given model as conditional p.d.f. for given parameters $f(\vec{x}|\vec{\theta})$
- we can specify our prior subjective belief on the distribution of the parameters before the experiment is performed: "prior" $\pi(\vec{\theta})$, and use Bayes theorem to obtain a "posterior" p.d.f. for the parameters, given observed data:

$$f(\vec{\theta}|\vec{x}^{\text{obs}}) \propto f(\vec{x}^{\text{obs}}|\vec{\theta})\pi(\vec{\theta})$$

- observations "update" our degree of belief of the parameters
- can also be generalized to statements about the model as a whole (Bayesian model comparison)

Bayesian inference

- consider the p.d.f. predicted in a given model as conditional p.d.f. for given parameters $f(\vec{x}|\vec{\theta})$
- we can specify our prior subjective belief on the distribution of the parameters before the experiment is performed: "prior" $\pi(\vec{\theta})$, and use Bayes theorem to obtain a "posterior" p.d.f. for the parameters, given observed data:

$$f(\vec{\theta}|\vec{x}^{\text{obs}}) \propto f(\vec{x}^{\text{obs}}|\vec{\theta})\pi(\vec{\theta})$$

- observations "update" our degree of belief of the parameters
- can also be generalized to statements about the model as a whole (Bayesian model comparison)

The likelihood

- a "model" predicts the p.d.f. for the observables: $f(\vec{x};\vec{\theta})$
- The likelihood function is the p.d.f. for the observables evaluated at the actuall outcome of an experiment, viewed as a function of the parameters of the model $\mathcal{L}(\vec{\theta}) \equiv f(\vec{x}^{obs};\vec{\theta})$
- If there are n statistically independent measurments x_i and each follows the distribution $f(x;\vec{\theta})$, the joint p.d.f. factorizes and

$$\mathcal{L}(\vec{\theta}) = \prod_{i=1}^{n} f(x_i^{obs};\vec{\theta})$$

- Example: energy spectrum - unbinned LH - histogram, binned LH

Note: the likelihood is not a p.d.f. for the model parameters $\vec{\theta}$

The likelihood

- a "model" predicts the p.d.f. for the observables: $f(\vec{x};\vec{\theta})$
- The likelihood function is the p.d.f. for the observables evaluated at the actuall outcome of an experiment, viewed as a function of the parameters of the model $\mathcal{L}(\vec{\theta}) \equiv f(\vec{x}^{obs};\vec{\theta})$
- If there are n statistically independent measurments x_i and each follows the distribution $f(x;\vec{\theta})$, the joint p.d.f. factorizes and

$$\mathcal{L}(\vec{\theta}) = \prod_{i=1}^{n} f(x_i^{obs};\vec{\theta})$$

- Example: energy spectrum - unbinned LH - histogram, binned LH

Note: the likelihood is not a p.d.f. for the model parameters $\vec{\theta}$

The likelihood

- a "model" predicts the p.d.f. for the observables: $f(\vec{x};\vec{\theta})$
- The likelihood function is the p.d.f. for the observables evaluated at the actuall outcome of an experiment, viewed as a function of the parameters of the model $\mathcal{L}(\vec{\theta}) \equiv f(\vec{x}^{obs};\vec{\theta})$
- If there are n statistically independent measurments x_i and each follows the distribution $f(x;\vec{\theta})$, the joint p.d.f. factorizes and

$$\mathcal{L}(\vec{\theta}) = \prod_{i=1}^{n} f(x_i^{obs};\vec{\theta})$$

- Example: energy spectrum - unbinned LH - histogram, binned LH

Note: the likelihood is not a p.d.f. for the model parameters $\vec{\theta}$

likelihood versus χ^2

- Example: consider observables x_i with multivariate Gaussian distribution:

$$f(\boldsymbol{x}; \boldsymbol{\mu}(\boldsymbol{\theta}), V) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))\right]$$

- The experiment has obtained the measurments $\boldsymbol{x}^{\text{obs}}$, then up to an irrelevant constant the logarithm of the likelihood is

$$\log \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{x}^{\text{obs}} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{x}^{\text{obs}} - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

- the "χ^2" is related to the likelihood by

$$\chi^2(\boldsymbol{\theta}) = -2 \log \mathcal{L}(\boldsymbol{\theta}) = (\boldsymbol{x}^{\text{obs}} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{x}^{\text{obs}} - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

- caveat: not true if $V(\theta)$

likelihood versus χ^2

- Example: consider observables x_i with multivariate Gaussian distribution:

$$f(\boldsymbol{x}; \boldsymbol{\mu}(\boldsymbol{\theta}), V) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))\right]$$

- The experiment has obtained the measurments $\boldsymbol{x}^{\text{obs}}$, then up to an irrelevant constant the logarithm of the likelihood is

$$\log \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{x}^{\text{obs}} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{x}^{\text{obs}} - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

- the "χ^2" is related to the likelihood by

$$\chi^2(\boldsymbol{\theta}) = -2 \log \mathcal{L}(\boldsymbol{\theta}) = (\boldsymbol{x}^{\text{obs}} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{x}^{\text{obs}} - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

- caveat: not true if $V(\theta)$

likelihood versus χ^2

- Example: consider observables x_i with multivariate Gaussian distribution:

$$f(\boldsymbol{x};\boldsymbol{\mu}(\boldsymbol{\theta}),V)=\frac{1}{\sqrt{(2\pi)^n|V|}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1}(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))\right]$$

- The experiment has obtained the measurments $\boldsymbol{x}^{\text{obs}}$, then up to an irrelevant constant the logarithm of the likelihood is

$$\log\mathcal{L}(\boldsymbol{\theta})=-\frac{1}{2}(\boldsymbol{x}^{\text{obs}}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1}(\boldsymbol{x}^{\text{obs}}-\boldsymbol{\mu}(\boldsymbol{\theta}))$$

- the "χ^2" is related to the likelihood by

$$\chi^2(\boldsymbol{\theta})=-2\log\mathcal{L}(\boldsymbol{\theta})=(\boldsymbol{x}^{\text{obs}}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1}(\boldsymbol{x}^{\text{obs}}-\boldsymbol{\mu}(\boldsymbol{\theta}))$$

- caveat: not true if $V(\theta)$

likelihood versus χ^2

The relation

$$\boxed{\chi^2(\boldsymbol{\theta}) = -2 \log \mathcal{L}(\boldsymbol{\theta})}$$

is often used as a general definition of χ^2, also if

- the p.d.f. of $\boldsymbol{x}$ is not Gaussian
- if unbinned data is used for the likelihood
  (a least-square statistic such as "χ^2" requires binned data)

Parameter estimation

The parameters which maximize the likelihood (minimize χ^2) can be used as "estimators" for the (unknown) true values of the parameters:

$$\log \mathcal{L}_{\text{max}} = \log \mathcal{L}(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta})$$

$$\chi^2_{\text{min}} = \chi^2(\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta})$$

- $\hat{\boldsymbol{\theta}}$ is sometimes called the "best fit point", or "maximum likelihood estimator"
- $\hat{\boldsymbol{\theta}}$ is a random variable (a statistic), because it is a function of the data
- in some sense maximum likelihood estimators are "optimal"
  converge towards the true values in the large sample limit

Minimization problem

- if the parameter dependence $\boldsymbol{\mu}(\boldsymbol{\theta})$ is linear (or sufficiently linear): solve the system of equations

$$\frac{\partial \chi^2}{\partial \theta_\alpha} = 0$$

- in non-linear situations the minimization has to be done numerically - can be very non-trivial (multi-dimensional parameter space, local minima, physical boundaries,...)

Goodness of fit

Q: How well does a model explain the data?

Example

- An experiment measures 10 observables x_i.
- The model predicts that x_i should be Gaussian distributed with known mean μ and variance σ (no free parameter in the model).
- If the model is true,

$$X^2 \equiv \sum_{i=1}^{10} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

follows a χ^2-distribution with 10 degrees of freedom: $\chi^2_{10}(X^2)$ (by definition)

Goodness of fit

Q: How well does a model explain the data?

Example

- An experiment measures 10 observables x_i.
- The model predicts that x_i should be Gaussian distributed with known mean μ and variance σ (no free parameter in the model).
- If the model is true,

$$X^2 \equiv \sum_{i=1}^{10} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

follows a χ^2-distribution with 10 degrees of freedom: $\chi^2_{10}(X^2)$ (by definition)

Goodness of fit, example:

- Suppose with the measured x_i we obtain $X^2_{obs} = 25$.
- We calculate the so-called "p-value":

$$p\text{-value} = \int_{25}^{\infty} \chi^2_{10}(z)dz \approx 0.5\%$$

- Interpretation: "If the model is true and the experiment was repeated many times, in 0.5% of the cases we would find $X^2 \geq 25$."
- No probability statement on the model, only on the outcome of the experiment!

Goodness of fit, example:

- Suppose with the measured x_i we obtain $X^2_{obs} = 25$.
- We calculate the so-called "p-value":

$$p\text{-value} = \int_{25}^{\infty} \chi^2_{10}(z)dz \approx 0.5\%$$

- Interpretation: "If the model is true and the experiment was repeated many times, in 0.5% of the cases we would find $X^2 \geq 25$."
- No probability statement on the model, only on the outcome of the experiment!

Goodness of fit including parameter estimation

- An experiment measures n observables x_i
- The model predicts that x_i should be Gaussian distributed with mean μ_i depending on p parameters θ_α:

$$\mu_i(\theta_\alpha) \qquad i = 1, ..., n; \quad \alpha = 1, ..., p$$

  The variances of x_i are σ_i^2.

- If the model is true,

$$X^2_{\text{min}} = \min_{\theta_\alpha} \left[\sum_{i=1}^{n} \left(\frac{x_i - \mu_i(\theta_\alpha)}{\sigma_i} \right)^2 \right]$$

  follows a χ^2-distribution with (n-p) degrees of freedom: $\chi^2_{n-p}(X^2)$ (see Maltoni,Schwetz,hep-ph/0304176 for a pedagogical proof).

Goodness of fit including parameter estimation

- An experiment measures n observables x_i
- The model predicts that x_i should be Gaussian distributed with mean μ_i depending on p parameters θ_α:

$$\mu_i(\theta_\alpha) \qquad i = 1, ..., n; \quad \alpha = 1, ..., p$$

  The variances of x_i are σ_i^2.

- If the model is true,

$$X^2_{\text{min}} = \min_{\theta_\alpha} \left[\sum_{i=1}^{n} \left(\frac{x_i - \mu_i(\theta_\alpha)}{\sigma_i} \right)^2 \right]$$

  follows a χ^2-distribution with (n-p) degrees of freedom: $\chi^2_{n-p}(X^2)$
  (see Maltoni,Schwetz,hep-ph/0304176 for a pedagogical proof).

Goodness of fit including parameter estimation

- given an observed value for X^2_{min} we can calculate the p-value by

$$p\text{-value} = \int_{X^2_{\text{min}}}^{\infty} \chi^2_{n-p}(z)dz$$

- Remember that the χ^2_n distribution has mean n and variance $2n$. Therefore, if the model is true, we expect ("expectation value")

$$X^2_{\text{min}} \approx (n-p) \pm \sqrt{2(n-p)}$$

(sloppy: $X^2_{\text{min}}/\text{d.o.f.} \simeq 1$)

- a "good fit" should have a p-value $\approx 50\%$
- a small p-value $[X^2_{\text{min}} \gg (n-p)]$ indicates an un-likely outcome
- a p-value close to 100% $[X^2_{\text{min}} \ll (n-p)]$ may indicate that errors are estimated too large (model fits data "too good")

Monte Carlo method

- It is not guaranteed that X^2_{min} follows a χ^2 distribution with $n - p$ degrees of freedom.
- In general the distribution has to be calculated by Monte Carlo methods.

the model predicts a p.d.f. for the observables depending on parameters: $f(x_i;\theta_\alpha)$, as well as $\mu_i(\theta_\alpha)$ and $\sigma_i(\theta_\alpha)$.

1. assume certain true values for θ_α and use the random number generator of your computer to generate an "artificial" realisation of the data x_i according to the p.d.f. predicted by the model
2. calculate the least square-statistic for that realisation:

$$X^2_{\text{min}} = \min_{\theta_\alpha} \left[\sum_{i=1}^{n} \left(\frac{x_i - \mu_i(\theta_\alpha)}{\sigma_i(\theta_\alpha)} \right)^2 \right]$$

   and store the value in a histogram
3. repeat those two steps many times
4. calculate the least square-statistic for the real observed data $X^2_{\text{min,obs}}$
5. the p-value is given by the fraction of the artificial data sets for which you have obtained a larger X^2_{min} than the observed one

Monte Carlo method

comments:

- the p-value may depend on the assumed "true values" for θ_α which has been used to generate the artificial data realisations.
  In frequentist statistics we cannot "marginalize" over the true values.
  Need to report the dependence on the true values.

- not restricted to the least-square statistic
  in principle one can use any statistic to evaluate the goodness of fit as long as their distribution can be estimated or calculated by Monte Carlo (though there may be good or bad ones)

Monte Carlo method

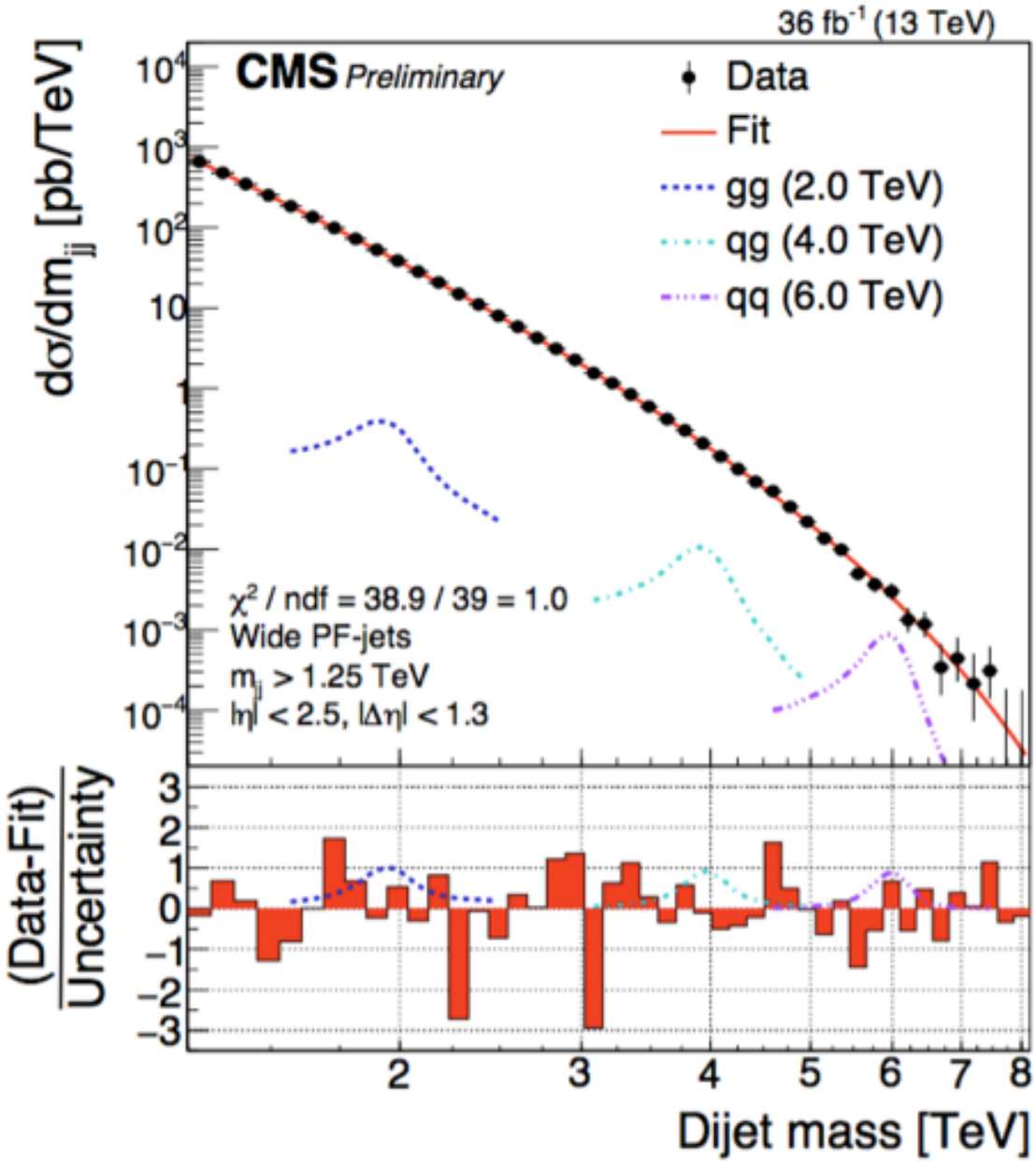
comments:

- the p-value may depend on the assumed "true values" for θ_α which has been used to generate the artificial data realisations.
  In frequentist statistics we cannot "marginalize" over the true values.
  Need to report the dependence on the true values.

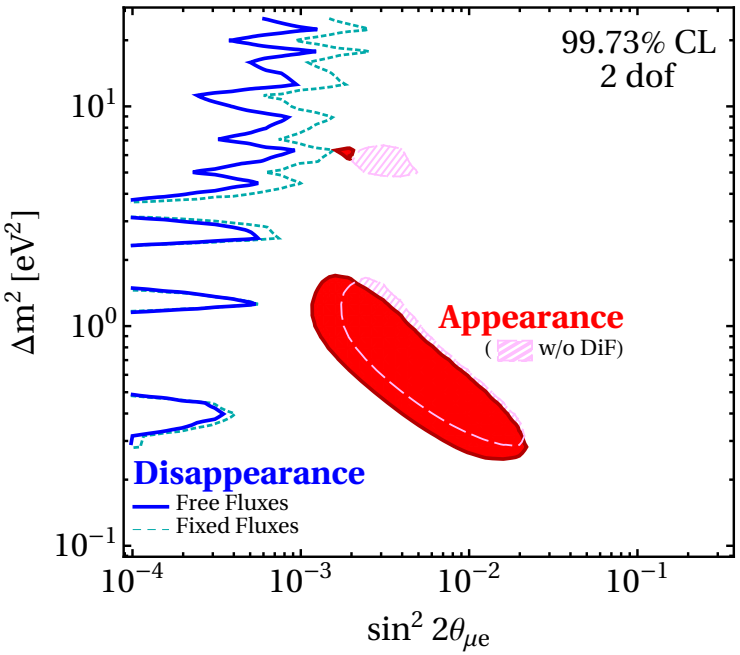
- not restricted to the least-square statistic
  in principle one can use any statistic to evaluate the goodness of fit as long as their distribution can be estimated or calculated by Monte Carlo (though there may be good or bad ones)

Comments on goodness of fit

- While $X^2_{\text{min}}/\text{d.o.f.} \gg 1$ indicates a problem with the fit, $X^2_{\text{min}}/\text{d.o.f.} \approx 1$ does not guarantee that the fit is "good".
- E.g., if the number of d.o.f. is large, there may be many data points which are not very sensitive to the "model" (dilution of the goodness of fit).
- Further diagnostics is recommended.
- E.g., divide data in sub-sets (without looking at the actual outcome of the experiments) and check for consistency
  Maltoni,Schwetz,hep-ph/0304176
- Pull diagram.


36 fb^{-1} (13 TeV)
CMS Preliminary
Data
Fit
gg (2.0 TeV)
qg (4.0 TeV)
qq (6.0 TeV)
dσ/dm$_{jj}$ [pb/TeV]
χ^2 / ndf = 38.9 / 39 = 1.0
Wide PF-jets
m$_{jj}$ > 1.25 TeV
|η| < 2.5, |Δη| < 1.3
(Data-Fit)/Uncertainty
Dijet mass [TeV]

Ex.: global fit with many data points



eV-sterile neutrino oscillations

Dentler et al., 1803.10661

$$\chi^2_{\rm min}/{\rm dof} = 1141/1159$$

p-value $= 64\%$

divide data into appearance and disappearance data, and consider "prize to pay" by the combination [Maltoni,Schwetz,hep-ph/0304176]:

$$\chi^2_{\rm PG} = \chi^2_{\rm min,glob} - \chi^2_{\rm min,app} - \chi^2_{\rm min,dis}$$

χ^2-distribution with P dof, $P = \#$ of params in common to the two sets

$$\Rightarrow \chi^2_{\rm PG}/{\rm dof} = 28.9/2,\ p\text{-value} = 5.3 \times 10^{-7}$$

Outline

- Suppose you have a model depending on some parameters $\boldsymbol{\theta}$
- The goodness of fit of your model is good and you come to the conclusion that the model fits the data well.
- You also obtained best fit values of the parameters $\hat{\boldsymbol{\theta}}$
- Now we want to address the question:
  what is the "acceptable range" for the parameters?

$\rightarrow$ confidence interavals (CI)

Confidence intervals

What is the precise meaning of statements as:

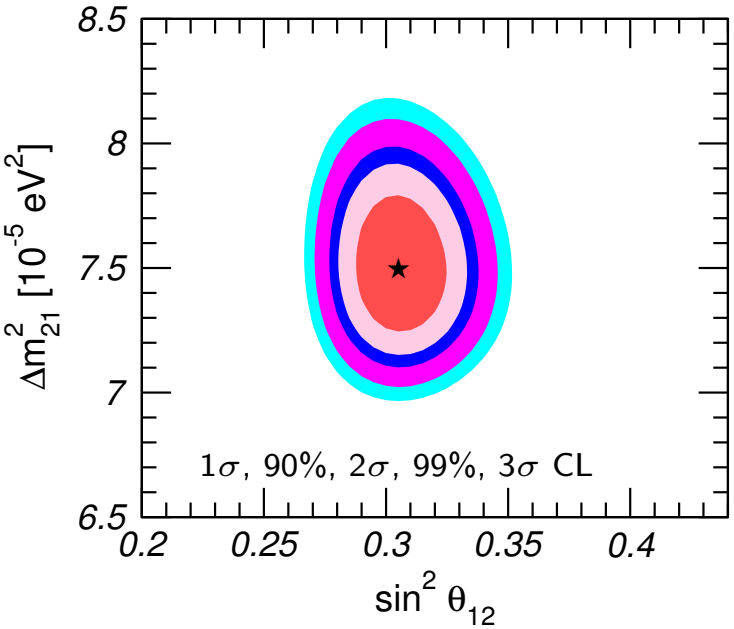
$$m_H = 125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{syst.})\,\text{GeV}$$

Confidence intervals

What is the precise meaning of statements as:

$$m_H = 125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{syst.})\,\text{GeV}$$

or plots like this:



Confidence intervals - frequentist interpretation

meaning of a 90% CL interval (or region):

- Suppose you repeat LHC many times.
- Each time you extract from the data an interval for the Higgs mass using the same procedure as for the one quoted above.
- In each of the many LHC experiments you would obtain a slightly different interval, but in 90% of the cases the interval will contain the true value of the Higgs mass ("coverage").
- Similar for multi-dimensional regions: the CL region (in n-dim space) would cover the true values for the n parameters in 90% of the cases.

Note: The probability statement is on the interval (or region), not on the parameter(s) of interest, which has an unknown fixed value.

Confidence intervals - frequentist interpretation

meaning of a 90% CL interval (or region):

- Suppose you repeat LHC many times.
- Each time you extract from the data an interval for the Higgs mass using the same procedure as for the one quoted above.
- In each of the many LHC experiments you would obtain a slightly different interval, but in 90% of the cases the interval will contain the true value of the Higgs mass ("coverage").
- Similar for multi-dimensional regions: the CL region (in n-dim space) would cover the true values for the n parameters in 90% of the cases.

Note: The probability statement is on the interval (or region), not on the parameter(s) of interest, which has an unknown fixed value.

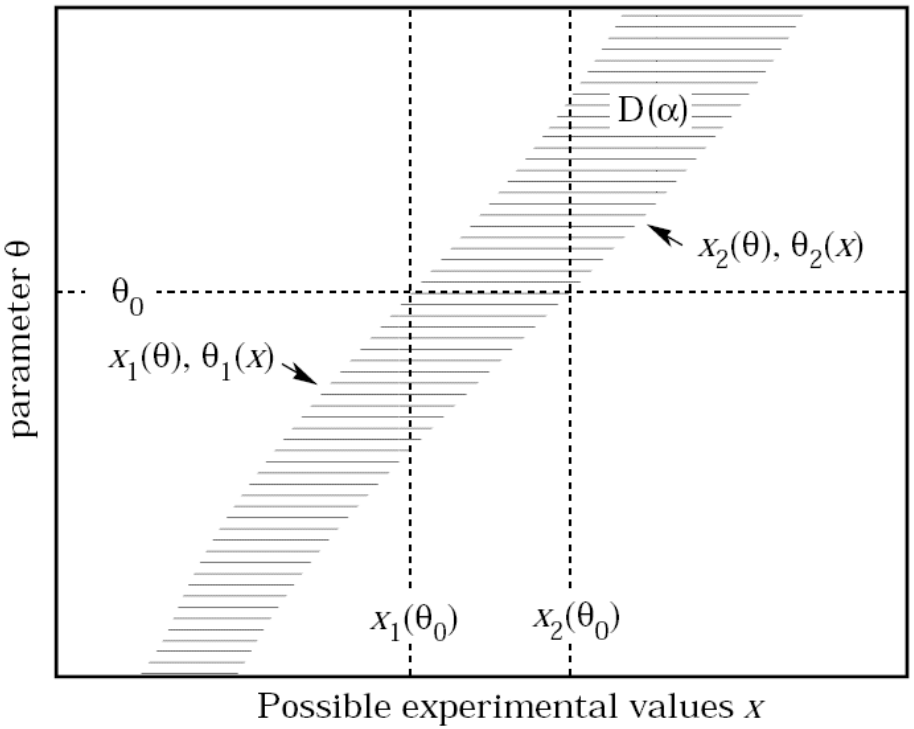
Confidence intervals - frequentist interpretation

meaning of a 90% CL interval (or region):

- Suppose you repeat LHC many times.
- Each time you extract from the data an interval for the Higgs mass using the same procedure as for the one quoted above.
- In each of the many LHC experiments you would obtain a slightly different interval, but in 90% of the cases the interval will contain the true value of the Higgs mass ("coverage").
- Similar for multi-dimensional regions: the CL region (in n-dim space) would cover the true values for the n parameters in 90% of the cases.

Note: The probability statement is on the interval (or region), not on the parameter(s) of interest, which has an unknown fixed value.

Confidence intervals from the confidence belt



$$P[x_1 < x < x_2; \theta] = 1 - \alpha = \int_{x_1}^{x_2} f(x; \theta) dx$$

$$1 - \alpha = P[x_1(\theta) < x < x_2(\theta)] = P[\theta_2(x) < \theta < \theta_1(x)]$$

probability statement on θ_1, θ_2, but not on θ

confidence intervals from the likelihood function

consider the likelihood function $\mathcal{L}(\theta_\alpha) \equiv f(x_i; \theta_\alpha)$, for a model depending on P parameters $\theta_\alpha = (\theta_1, ...\theta_P)$

- $\hat{\theta}_\alpha$: parameter values which maximize the likelihood: $\mathcal{L}_{\max} = \mathcal{L}(\hat{\theta}_\alpha)$ ($\hat{\theta}_\alpha$ are "estimators" of the true values of θ_α)
- under certain conditions (Wilk's theorem),

$$\Delta X^2 \equiv 2 \log \frac{\mathcal{L}_{\max}}{\mathcal{L}(\theta_\alpha)}$$

will be distributed as a χ^2 with P d.o.f. (independent of true values)

- remember $\chi^2 = -2 \log \mathcal{L} \rightarrow$

$$\Delta X^2 = \chi^2(\theta_\alpha) - \chi^2_{\min}$$

confidence intervals from the likelihood function

consider the likelihood function $\mathcal{L}(\theta_\alpha) \equiv f(x_i; \theta_\alpha)$, for a model depending on P parameters $\theta_\alpha = (\theta_1, ...\theta_P)$

- $\hat{\theta}_\alpha$: parameter values which maximize the likelihood: $\mathcal{L}_{\max} = \mathcal{L}(\hat{\theta}_\alpha)$ ($\hat{\theta}_\alpha$ are "estimators" of the true values of θ_α)

- under certain conditions (Wilk's theorem),

$$\boxed{\Delta X^2 \equiv 2 \log \frac{\mathcal{L}_{\max}}{\mathcal{L}(\theta_\alpha)}}$$

  will be distributed as a χ^2 with P d.o.f. (independent of true values)

- remember $\chi^2 = -2 \log \mathcal{L} \rightarrow$

$$\boxed{\Delta X^2 = \chi^2(\theta_\alpha) - \chi^2_{\min}}$$

confidence intervals from the likelihood function

consider the likelihood function $\mathcal{L}(\theta_\alpha) \equiv f(x_i; \theta_\alpha)$, for a model depending on P parameters $\theta_\alpha = (\theta_1, ...\theta_P)$

- $\hat{\theta}_\alpha$: parameter values which maximize the likelihood: $\mathcal{L}_{\max} = \mathcal{L}(\hat{\theta}_\alpha)$ ($\hat{\theta}_\alpha$ are "estimators" of the true values of θ_α)

- under certain conditions (Wilk's theorem),

$$\boxed{\Delta X^2 \equiv 2 \log \frac{\mathcal{L}_{\max}}{\mathcal{L}(\theta_\alpha)}}$$

  will be distributed as a χ^2 with P d.o.f. (independent of true values)

- remember $\chi^2 = -2 \log \mathcal{L} \rightarrow$

$$\boxed{\Delta X^2 = \chi^2(\theta_\alpha) - \chi^2_{\text{min}}}$$

using χ^2

suppose the experiment divides the range of observation into N bins

$$\text{define:} \qquad \chi^2(\theta_\alpha) \equiv \sum_{i=1}^{n} \left(\frac{x_i - \mu_i(\theta_\alpha)}{\sigma_i} \right)^2$$

$\chi^2(\theta_\alpha)$	$=$	$\chi^2_{\text{min}}(\hat{\theta}_\alpha)$	$+$	$\Delta\chi^2(\theta_\alpha)$
N		$N - P$		P
		parameter estimation, goodness of fit		confidence interval

- χ^2_{min} follows a χ^2-distribution with $N - P$ d.o.f. and can be used to evaluate the goodness of fit.
- $\Delta\chi^2$ follows a χ^2-distribution with P d.o.f.

Confidence regions from $\Delta\chi^2$

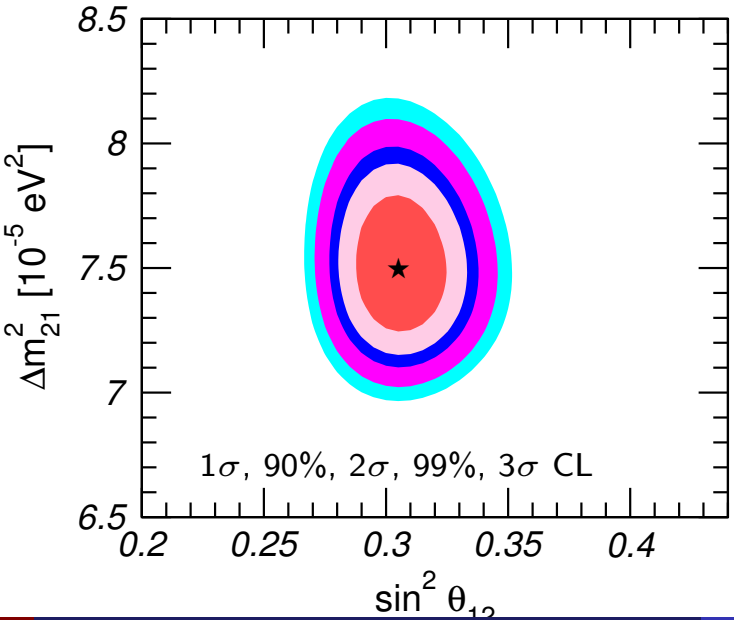
A P-dimensional region in the space θ_α at given CL is obtained by requiring $\Delta\chi^2(\theta_\alpha) < X(\text{CL})$ (contours in $\Delta\chi^2$)

d.o.f. \ CL	68%(1σ)	90%	95%(2σ)	99%	99.73%(3σ)
1	1	2.71	4	6.64	9
2	2.28	4.61	5.99	9.21	11.8
3	3.51	6.25	7.82	11.4	14.2

Confidence regions from $\Delta\chi^2$

A P-dimensional region in the space θ_α at given CL is obtained by requiring $\Delta\chi^2(\theta_\alpha) < X(\text{CL})$ (contours in $\Delta\chi^2$)

d.o.f. \ CL	68%(1σ)	90%	95%(2σ)	99%	99.73%(3σ)
1	1	2.71	4	6.64	9
2	2.28	4.61	5.99	9.21	11.8
3	3.51	6.25	7.82	11.4	14.2



On Wilk's theorem (1938)

When is $\Delta\chi^2(\theta_\alpha)$ really χ^2-distributed?

- Wilks theorem applies if the theoretical predictions $\mu_i(\theta_\alpha)$ span a linear space when θ_α are varied
- this holds when the predictions can be expanded to linear order $\mu_i(\theta_\alpha) \approx A_i + B_{i\alpha}\theta_\alpha$
- this is exact for a linear model
- in non-linear models, this holds in the vicinity of the best fit point and is reliable up to a certain CL, beyond which the non-linear character of the parameter dependence becomes important
- for "powerful" data the linear approximation will hold to high CL, for "weak" data non-linearities may become important already at low CL.

On Wilk's theorem (1938)

important examples where Wilk's theorem does not hold:

- close to a physical boundary of a parameter
  ex.: absolute neutrino mass observables: $m^2_{\nu_e} \geq 0, \sum_i m_i \geq 0$
  ex.: upper limit on sterile neutrino mixing $|U_{\alpha 4}|^2$

- when certain values of the predictions $\mu_i(\theta_\alpha)$ cannot be reached
  ex.: trigonometric dependencies:
  $\sin^2 2\theta_{23} \leq 1$, δ_{CP} dependence

On Wilk's theorem (1938)

important examples where Wilk's theorem does not hold:

- close to a physical boundary of a parameter
  ex.: absolute neutrino mass observables: $m^2_{\nu_e} \geq 0, \sum_i m_i \geq 0$
  ex.: upper limit on sterile neutrino mixing $|U_{\alpha 4}|^2$

- when certain values of the predictions $\mu_i(\theta_\alpha)$ cannot be reached
  ex.: trigonometric dependencies:
  $\sin^2 2\theta_{23} \leq 1$, δ_{CP} dependence

- confidence regions from the standard $\Delta\chi^2$ contours will be only approximate.

- if large deviations from Gaussianity are expected, confidence regions have to be constructed by Monte Carlo methods.

CIs from explicit confidence belt construction

1. assume certain true values for θ and use the random number generator of your computer to generate an "artificial" realisation of the data x_i according to the p.d.f. predicted by the model
2. calculate the least square-statistic for that realisation: $\Delta X^2(\theta) = X^2(\theta) - X^2_{\text{min}}$ and store the value in a histogram
3. repeat steps 1 and 2 many times
4. repeat steps 1, 2, 3 for each value of θ
5. at each value for θ search for the cut-value $X^2_{\text{cut}}(\theta)$, such that $\Delta X^2(\theta)$ is larger than $X^2_{\text{cut}}(\theta)$ in 10% of the cases (for a 90% CI)
6. calculate the least square-statistic for the observed data $\Delta X^2_{\text{obs}}(\theta)$
7. the CI is given by the union of all values of θ for which $\Delta X^2_{\text{obs}}(\theta) < X^2_{\text{cut}}(\theta)$

Feldman, Cousins, PRD57, 3873 (1998), physics/9711021

- In the limit where Wilk's theorem holds ("Gaussian approximation" or "χ^2 approximation")
  - 1-dimensional χ^2 projections will be parabolas
  - p-dimensional regions will be p-dimensional ellipsoids
  - inclination of the ellipse in a 2-dim plane gives the correlation between those two parameters

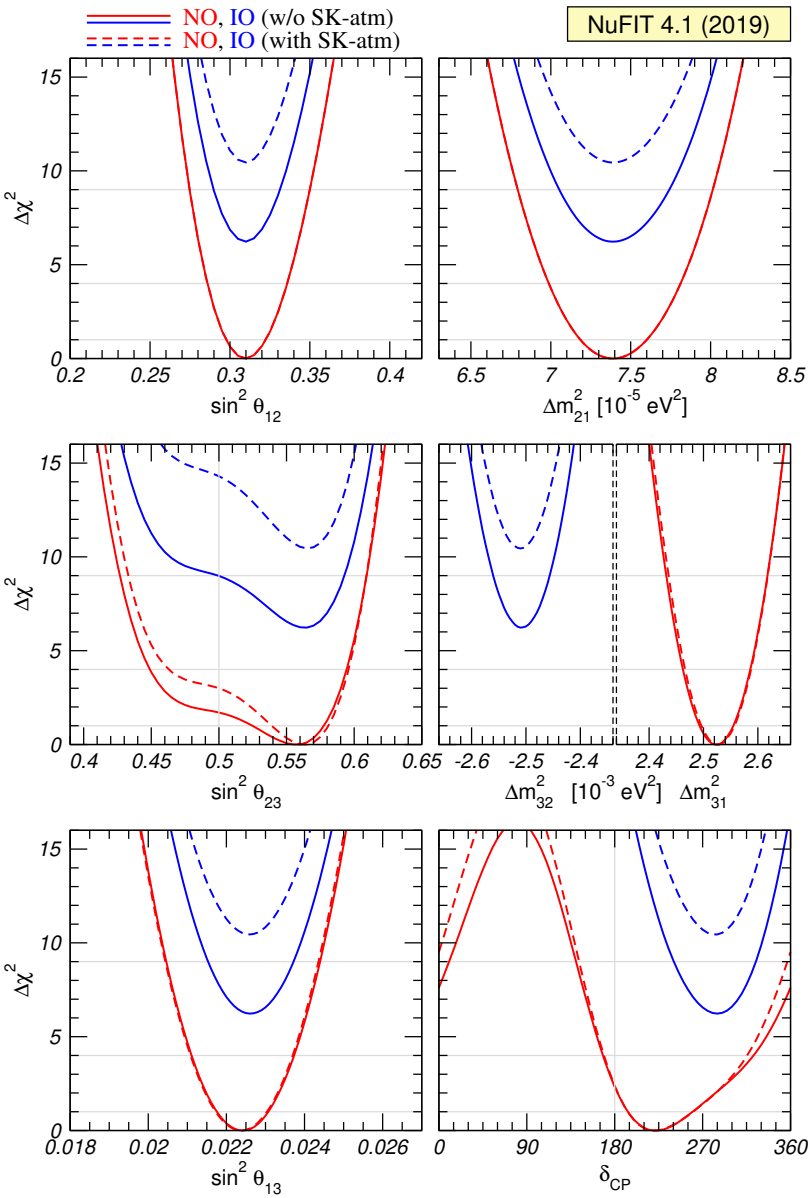
- In the θ_{12}, θ_{13}, Δm^2_{21} space we are close to Gaussian

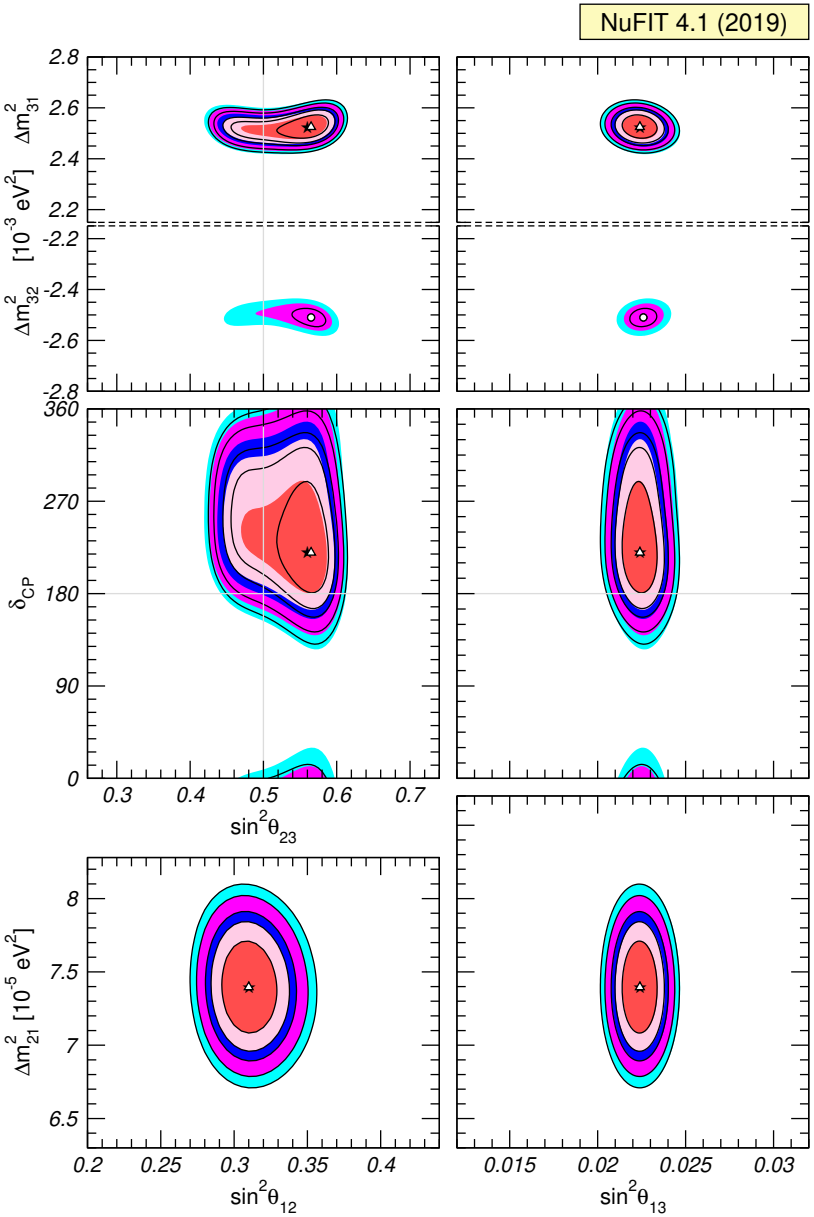
- non-Gaussianities are relevant:
  - mass ordering degeneracy Δm^2_{31}
  - octant degeneracy $\chi^2(\theta_{23})$
  - CP phase δ (periodic parameter space!)

- In these cases translation of $\Delta\chi^2$ values into CL (or probabilities) is only approximate.

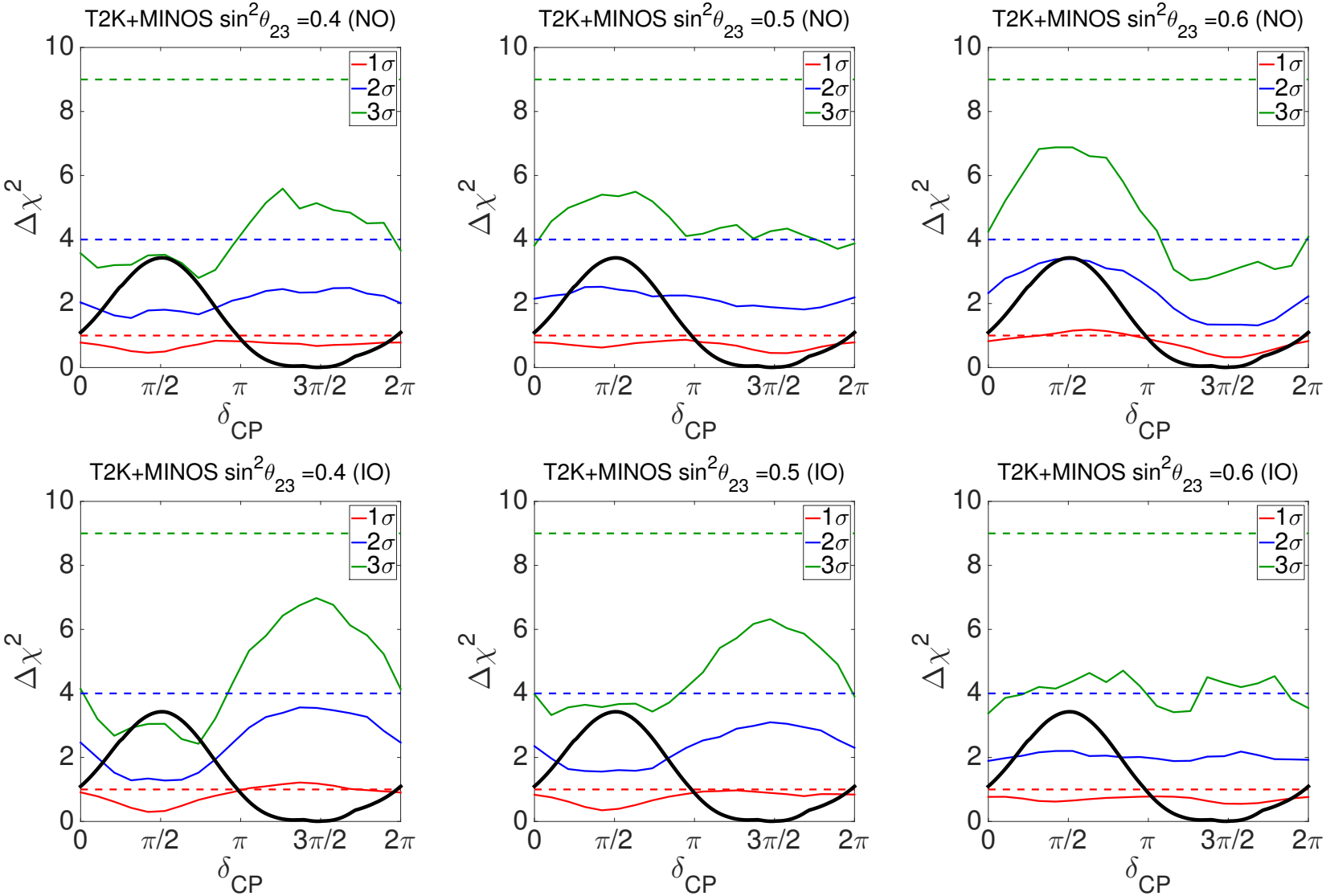
- In the limit where Wilk's theorem holds ("Gaussian approximation" or "χ^2 approximation")
  - 1-dimensional χ^2 projections will be parabolas
  - p-dimensional regions will be p-dimensional ellipsoids
  - inclination of the ellipse in a 2-dim plane gives the correlation between those two parameters
- In the θ_{12}, θ_{13}, Δm^2_{21} space we are close to Gaussian
- non-Gaussianities are relevant:
  - mass ordering degeneracy Δm^2_{31}
  - octant degeneracy $\chi^2(\theta_{23})$
  - CP phase δ (periodic parameter space!)
- In these cases translation of $\Delta\chi^2$ values into CL (or probabilities) is only approximate.

Global 3-flavour fit

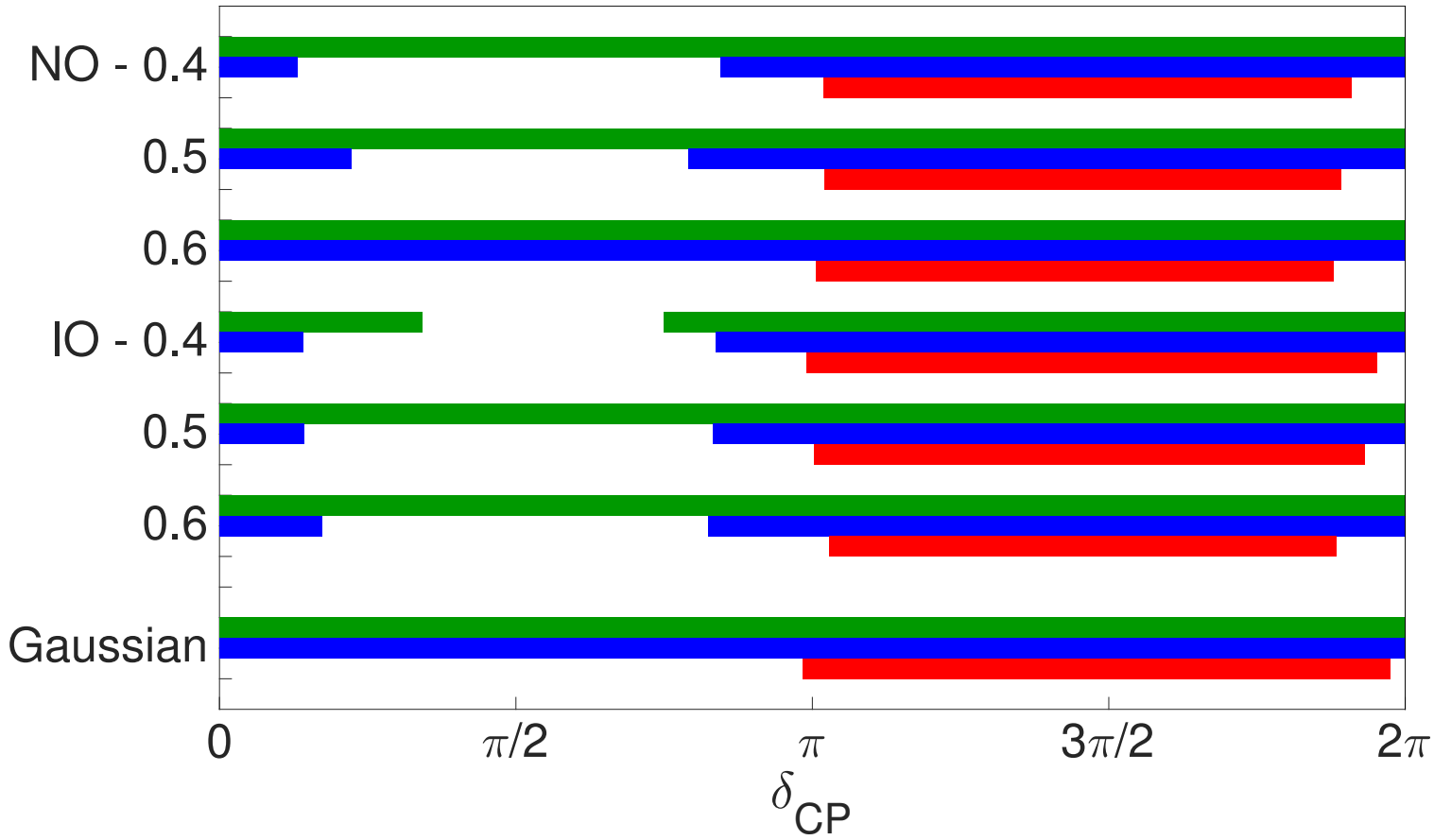




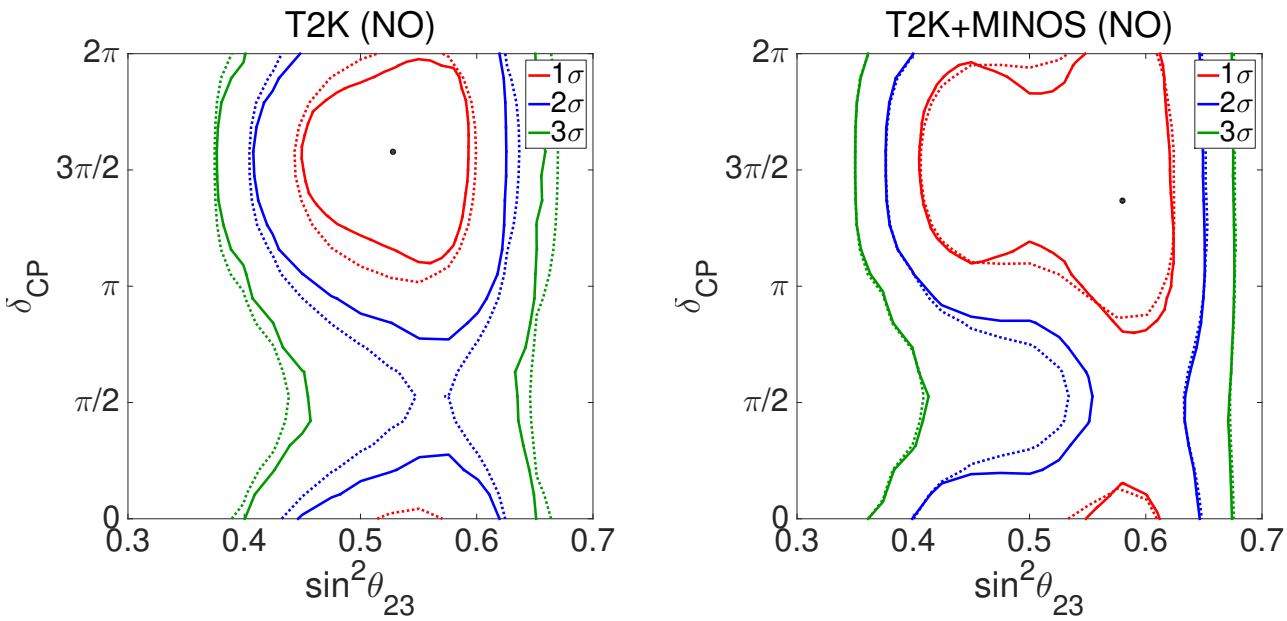
Elevant, Schwetz, 1506.07685



Elevant, Schwetz, 1506.07685



2-dimensional regions are closer to Gaussianity



Elevant, Schwetz, 1506.07685

see also Esteban et al., 1611.01514 [NuFit 3.0]

Invariance under variable change

Best fit values from χ^2_{min} and confidence intervals from $\Delta\chi^2 = \text{const}$ are invariant under (reasonable) variable transformations

Ex.: one variable $\theta \to y = g(\theta)$, e.g. $\theta \to y = \sin^2\theta$

- If $\chi^2(\theta)$ has a minimum at $\hat{\theta}$, $\chi^2(y)$ has a minimum at

$$\hat{y} = g(\hat{\theta})$$

- If $[\theta_1, \theta_2]$ is a confidence interval for θ obtained from requiring $\Delta\chi^2(\theta) = X_{\text{cut}}$, then the corresponding interval for y is

$$[y_1, y_2] = [g(\theta_1), g(\theta_2)]$$

unbinned likelihood versus χ^2

- binning implies loss of information
- whenever possible (practical) use unbinned likelihood
- if binning is necessary use as many bins as possible (practical)
  - rule of thumb: bin size comparable to the resolution
    (e.g., energy resolution in case of an energy spectrum)
    (smaller bins do not add more information, but do not hurt either)
  - if number of events per bin becomes small, an option is to use the "Poisson" version of χ^2 (see previous slide)

Combining several experiments

- consider M experiments.
- experiment ex consists of N_{ex} data points.
- each experiment has its own χ^2 function: $\chi^2_{ex}(\boldsymbol{\theta})$
- if there is no correlation between experiments, the combined χ^2 is simply

$$\chi^2_{glob}(\boldsymbol{\theta}) = \sum_{ex=1}^{M} \chi^2_{ex}(\boldsymbol{\theta}) \qquad \#\,\text{d.o.f.} = \sum_{ex=1}^{M} N_{ex}$$

  (or multiplying the likelihoods).
- possible correlations need to be taken into account by the covariance matrix or by introducing nuissance parameters (see later).
- any minimization over parameters has to be done for $\chi^2_{glob}(\boldsymbol{\theta})$, not the individual experiments

$$\min[f(x)] + \min[g(x)] \neq \min[f(x) + g(x)]$$

Combining several experiments

- consider M experiments.
- experiment ex consists of N_{ex} data points.
- each experiment has its own χ^2 function: $\chi^2_{ex}(\boldsymbol{\theta})$
- if there is no correlation between experiments, the combined χ^2 is simply

$$\chi^2_{glob}(\boldsymbol{\theta}) = \sum_{ex=1}^{M} \chi^2_{ex}(\boldsymbol{\theta}) \qquad \#\,\text{d.o.f.} = \sum_{ex=1}^{M} N_{ex}$$

  (or multiplying the likelihoods).

- possible correlations need to be taken into account by the covariance matrix or by introducing nuissance parameters (see later).
- any minimization over parameters has to be done for $\chi^2_{glob}(\boldsymbol{\theta})$, not the individual experiments

$$\min[f(x)] + \min[g(x)] \neq \min[f(x) + g(x)]$$

Combining several experiments

- consider M experiments.
- experiment ex consists of N_{ex} data points.
- each experiment has its own χ^2 function: $\chi^2_{ex}(\boldsymbol{\theta})$
- if there is no correlation between experiments, the combined χ^2 is simply

$$\chi^2_{glob}(\boldsymbol{\theta}) = \sum_{ex=1}^{M} \chi^2_{ex}(\boldsymbol{\theta}) \qquad \#\,\text{d.o.f.} = \sum_{ex=1}^{M} N_{ex}$$

  (or multiplying the likelihoods).
- possible correlations need to be taken into account by the covariance matrix or by introducing nuissance parameters (see later).
- any minimization over parameters has to be done for $\chi^2_{glob}(\boldsymbol{\theta})$, not the individual experiments

$$\min[f(x)] + \min[g(x)] \neq \min[f(x) + g(x)]$$

Combining several experiments

- consider M experiments.
- experiment ex consists of N_{ex} data points.
- each experiment has its own χ^2 function: $\chi^2_{ex}(\boldsymbol{\theta})$
- if there is no correlation between experiments, the combined χ^2 is simply

$$\chi^2_{glob}(\boldsymbol{\theta}) = \sum_{ex=1}^{M} \chi^2_{ex}(\boldsymbol{\theta}) \qquad \#\,\text{d.o.f.} = \sum_{ex=1}^{M} N_{ex}$$

  (or multiplying the likelihoods).
- possible correlations need to be taken into account by the covariance matrix or by introducing nuissance parameters (see later).
- any minimization over parameters has to be done for $\chi^2_{glob}(\boldsymbol{\theta})$, not the individual experiments

$$\min[f(x)] + \min[g(x)] \neq \min[f(x) + g(x)]$$

Bayes theorem and p.d.f.s for parameters

In contrast to the frequentist approach, Bayesians assign a probability distribution to the parameters of a model (or the model itself).

We can use Bayes theorem to get a p.d.f. for θ from the likelihood function of the data x: $f(x,\theta) = f(\theta|x)f(x) = f(x|\theta)f(\theta)$ or

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \propto \mathcal{L}(\theta)\pi(\theta)$$

- $f(x|\theta) = \mathcal{L}(\theta)$: likelihood function
- $f(\theta) = \pi(\theta)$ prior p.d.f of θ
- $f(\theta|x)$ posterior p.d.f. of θ

The data updates our "degree of belief" about θ from the prior $\pi(\theta)$ to the posterior p.d.f. $f(\theta|x)$

Bayesian parameter intervals

A Bayesian interval for θ containing a probability β is given by

$$\beta = \int_{\theta_1}^{\theta_2} f(\theta|x)d\theta$$

- $\theta_{1,2}$ not unique
- easy to define upper or lower limits
- "equal probability" intervals or "central" intervals
- N-dimensions: volume integration
  (often one defines the volume by constant-likelihood contours)
  integration is efficiently done by Markov chain Monte Carlo techniques

Bayesian intervals in general are not invariant under variable transformations. 1-dim example:

$$\beta = \int_{\theta_1}^{\theta_2} \mathcal{L}(\theta)\pi(\theta)d\theta$$

consider variable transformation $y = y(\theta)$

$$\int_{\theta_1}^{\theta_2} \mathcal{L}(\theta)\pi(\theta)d\theta = \int_{y(\theta_1)}^{y(\theta_2)} \mathcal{L}(\theta(y))\pi(\theta(y))\frac{d\theta}{dy}dy$$

BUT: the interval $[y_1, y_2]$ defined by

$$\beta = \int_{y_1}^{y_2} \mathcal{L}(\theta(y))\pi(\theta(y))dy$$

in general differs from $[y(\theta_1), y(\theta_2)]$
$\Rightarrow$ limits are not invariant under non-linear variable transformation!

How to choose the prior?

- constant prior?
- constant logarithmic prior? ("uninformative")
- non-normalizable prior ("impropper")?
- "objective priors" (allow for frequentist interpretation)
  require certain properties like maximum gain of information, or invariance under variable transformations
- result becomes independent of prior when the width of the likelihood is small compared the typical variation of the prior
- easy to include physical boundaries of a parameter

If the likelihood is much more peaked than the prior ("good data"), the result becomes independent of the prior and similar to frequentist limits.

How to choose the prior?

- constant prior?
- constant logarithmic prior? ("uninformative")
- non-normalizable prior ("impropper")?
- "objective priors" (allow for frequentist interpretation)
  require certain properties like maximum gain of information, or invariance under variable transformations
- result becomes independent of prior when the width of the likelihood is small compared the typical variation of the prior
- easy to include physical boundaries of a parameter

If the likelihood is much more peaked than the prior ("good data"), the result becomes independent of the prior and similar to frequentist limits.

Bayesian and Frequentist intervals

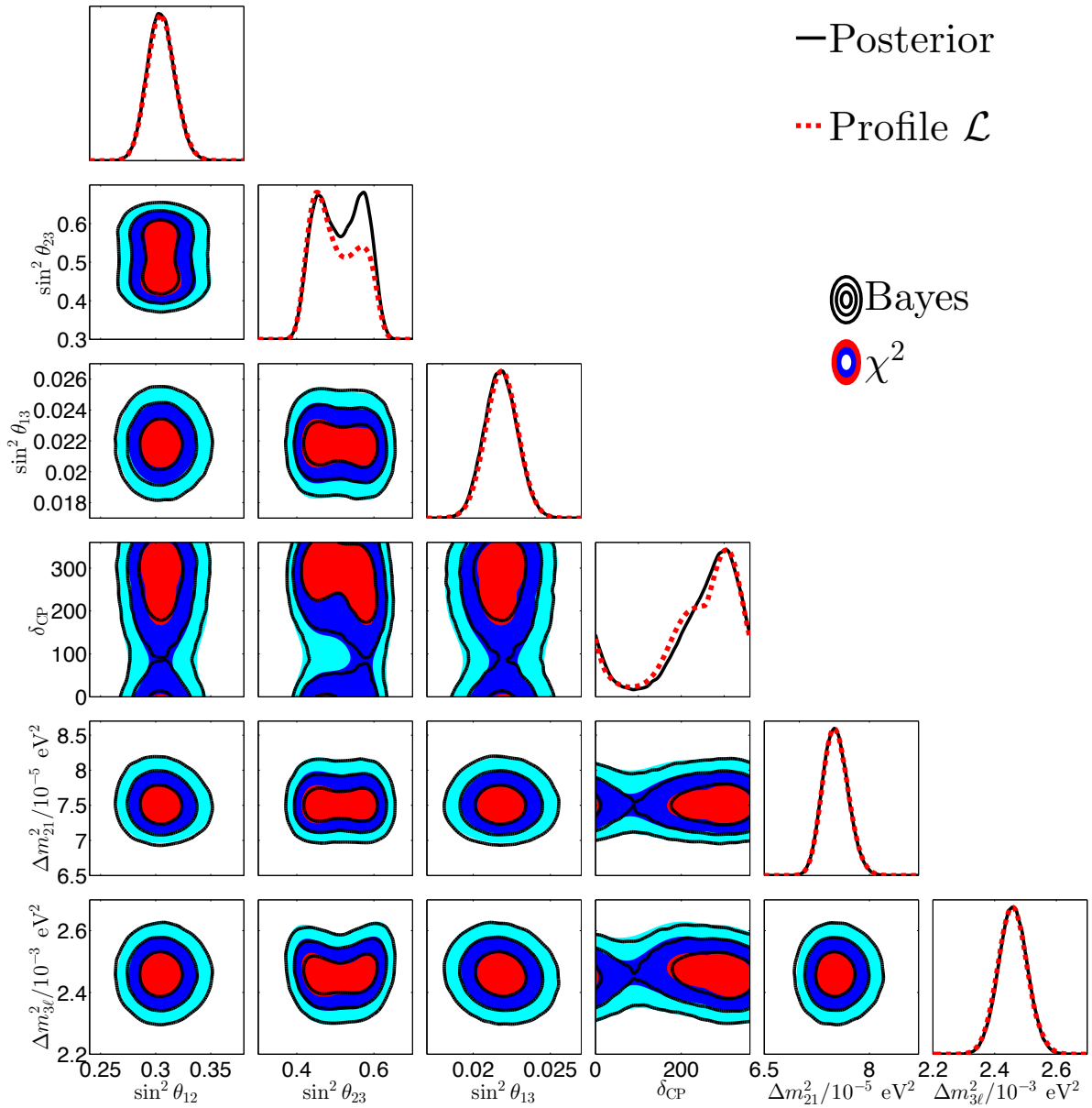
If "the problem is Gaussian", the χ^2 is a parabola

$$\chi^2 = \chi^2_{\text{min}} + \left(\frac{\theta - \hat{\theta}}{\sigma_\theta} \right)^2$$

and the likelihood is proportional to a Gaussian

$$\mathcal{L}(\theta) \propto e^{-\chi^2/2}$$

If furthermore, the prior is constant in θ, then bayesian (central) intervals and Frequentist intervals (from $\Delta\chi^2$) agree.



Bergstrom et al., 1507.04366

On multi-dimensional parameter spaces

Suppose you want to show regions at a CL β for q parameters x, and you are not interested in $P - q$ parameters y:

- use q d.o.f. and minimize wrt to y:
  "the q-dimensional region for x, irrespective of the values of y"
- use q d.o.f. and fix y to some values:
  "the q-dimensional region for x, assuming some true value of y"
  (ex.: upper bound on σ_{scat} for fixed m_χ)

On multi-dimensional parameter spaces

Suppose you want to show regions at a CL β for q parameters x, and you are not interested in $P - q$ parameters y:

- use q d.o.f. and minimize wrt to y:
  "the q-dimensional region for x, irrespective of the values of y"
- use q d.o.f. and fix y to some values:
  "the q-dimensional region for x, assuming some true value of y"
  (ex.: upper bound on σ_{scat} for fixed m_χ)

On multi-dimensional parameter spaces

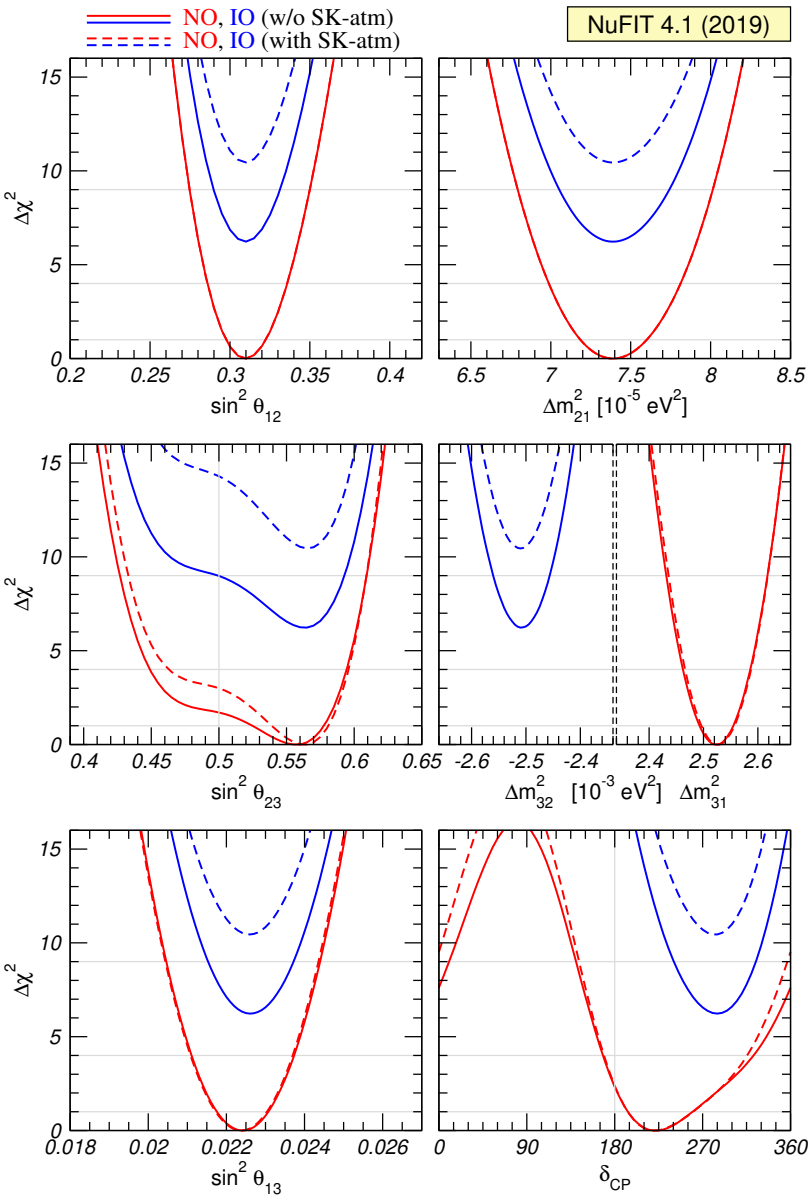
Suppose you want to show regions at a CL β for p parameters x, and you are not interested in $q = P - p$ parameters y [note: $\boldsymbol{\theta} = (x, y)$]:

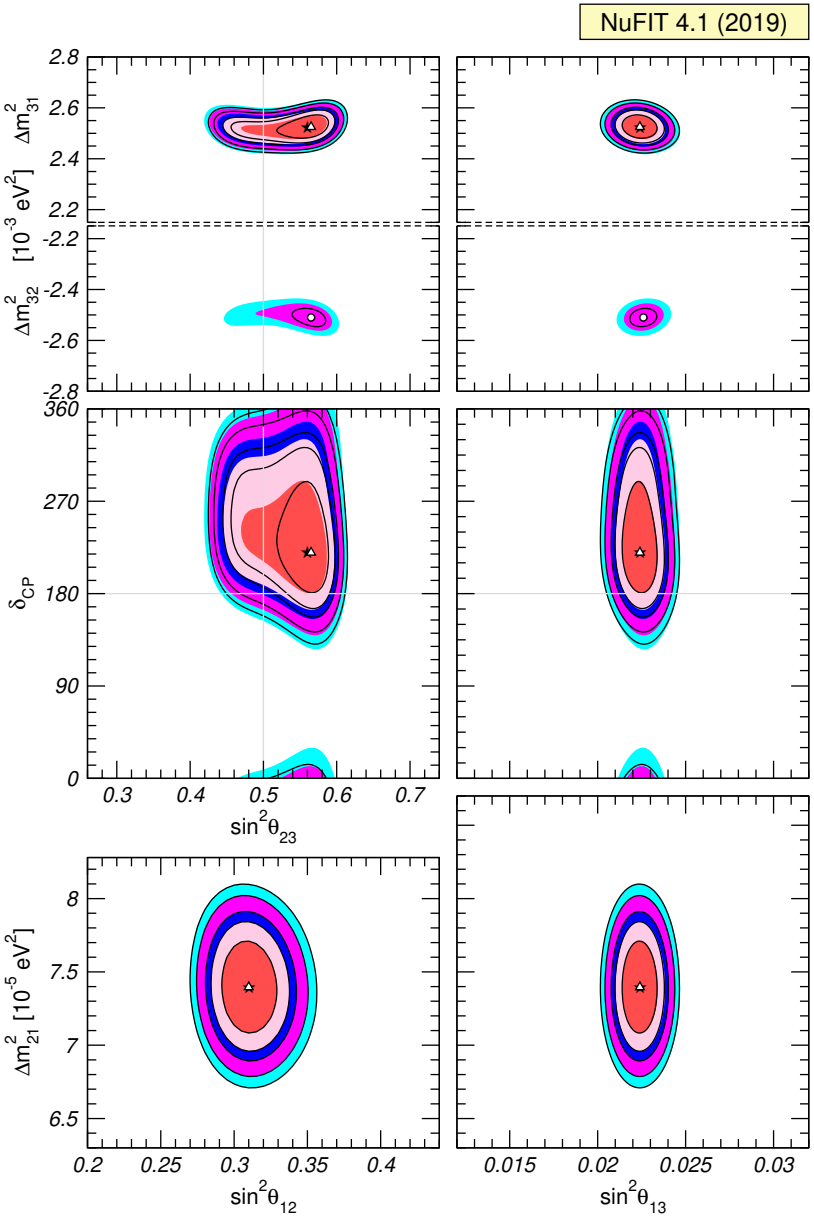
- use p d.o.f. and minimize wrt to y:
  "the p-dimensional region for x, irrespective of the values of y"

$\chi^2(\boldsymbol{\theta})$	$=$	$\chi^2_{\text{min}}(\hat{\boldsymbol{\theta}})$	$+$	$\Delta\chi^2(\boldsymbol{\theta})$
N		$N - P$		P
$\Delta\chi^2(x, y)$	$=$	$\Delta\chi^2_{\text{min},y}(x)$	$+$	$\delta\chi^2(x, y)$
P		$p = P - q$		q

$$\Delta\chi^2_{\text{min},y}(x) \equiv \text{min}[\Delta\chi^2(x, y); y] \qquad (p \text{ d.o.f.})$$

Example: 1-dim and 2-dim projections





Bayesian parameter marginalization

In a Bayesian framework it is straight forward to obtain the marginalized p.d.f. by integrating over nuisance parameters:

$$f(x,y) \propto \mathcal{L}(x,y)\pi(x,y)$$

$$f(x) = \int dy\, f(x,y)$$

If the prior factorizes $\pi(x,y) = \pi(x)\pi(y)$:

$$f(x) \propto \pi(x) \int dy\, \mathcal{L}(x,y)\pi(y)$$

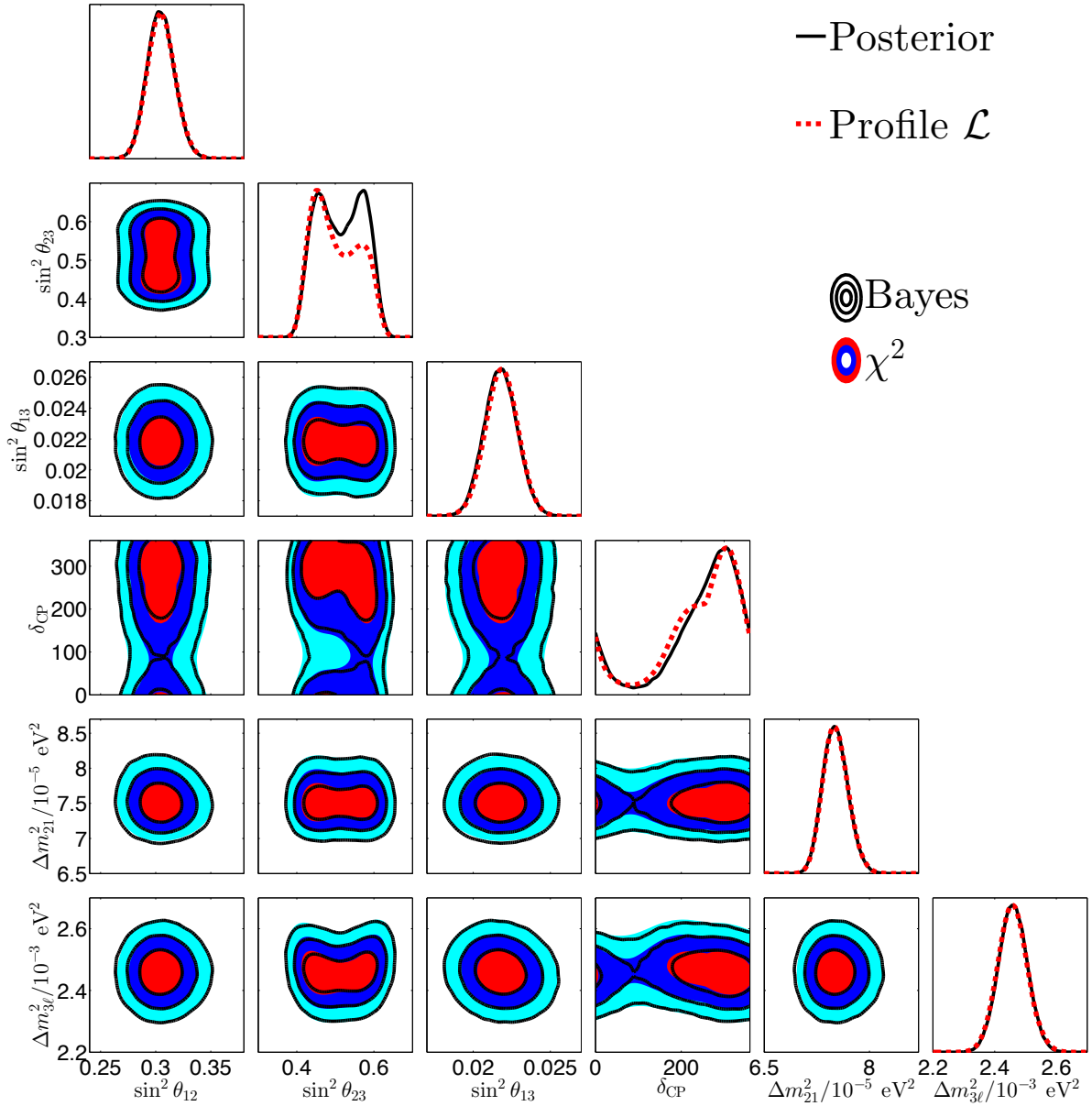
Bayesian parameter marginalization

In a Bayesian framework it is straight forward to obtain the marginalized p.d.f. by integrating over nuisance parameters:

$$f(x,y) \propto \mathcal{L}(x,y)\pi(x,y)$$
$$f(x) = \int dy\, f(x,y)$$

If the prior factorizes $\pi(x,y) = \pi(x)\pi(y)$:

$$f(x) \propto \pi(x) \int dy\, \mathcal{L}(x,y)\pi(y)$$



Bergstrom et al., 1507.04366

Summary marginalization

- frequentist: χ^2 is minimized (likelihood maximized) with respect to nuissance parameters $\rightarrow$ "profiling"

$$\chi^2(x) = \min_y \left[\chi^2(x, y)\right]$$

- Bayesian: posterior p.d.f. is integrated over nuissance parameters

$$f(x) = \int dy\, f(x, y)$$

results may differ dramatically, especially in multi-dimensional spaces

Summary marginalization

- frequentist: χ^2 is minimized (likelihood maximized) with respect to nuissance parameters $\rightarrow$ "profiling"

$$\chi^2(x) = \min_y \left[\chi^2(x,y)\right]$$

- Bayesian: posterior p.d.f. is integrated over nuissance parameters

$$f(x) = \int dy\, f(x,y)$$

results may differ dramatically, especially in multi-dimensional spaces

How to analyze data from neutrino oscillation experiments

Basic steps towards an analysis

- Suppose a given experiment divides the range of observation into N bins. The outcome is reported in number of observed events in each bin n_i.

  (Expect Poisson distribution for the number of events in each bin.)

- For given oscillation parameters

$$\boldsymbol{\theta} = (\theta_{12}, \theta_{13}, \theta_{23}, \delta_{\mathrm{CP}}, \Delta m^2_{21}, \Delta m^2_{31}) \qquad (P = 6)$$

  we can predict the expected number of events per bin $\mu_i(\boldsymbol{\theta})$.

- Build a χ^2, e.g. (more details later):

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[\frac{\mu_i(\boldsymbol{\theta}) - n_i}{\sigma_i} \right]^2$$

- Use $\chi^2(\boldsymbol{\theta})$ to perform a statistical analysis

Basic steps towards an analysis

- Suppose a given experiment divides the range of observation into N bins. The outcome is reported in number of observed events in each bin n_i.

  (Expect Poisson distribution for the number of events in each bin.)

- For given oscillation parameters

$$\boldsymbol{\theta} = (\theta_{12}, \theta_{13}, \theta_{23}, \delta_{\mathrm{CP}}, \Delta m^2_{21}, \Delta m^2_{31}) \qquad (P = 6)$$

  we can predict the expected number of events per bin $\mu_i(\boldsymbol{\theta})$.

- Build a χ^2, e.g. (more details later):

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[\frac{\mu_i(\boldsymbol{\theta}) - n_i}{\sigma_i} \right]^2$$

- Use $\chi^2(\boldsymbol{\theta})$ to perform a statistical analysis

Basic steps towards an analysis

- Suppose a given experiment divides the range of observation into N bins. The outcome is reported in number of observed events in each bin n_i.

  (Expect Poisson distribution for the number of events in each bin.)

- For given oscillation parameters

$$\boldsymbol{\theta} = (\theta_{12}, \theta_{13}, \theta_{23}, \delta_{\mathrm{CP}}, \Delta m^2_{21}, \Delta m^2_{31}) \qquad (P = 6)$$

  we can predict the expected number of events per bin $\mu_i(\boldsymbol{\theta})$.

- Build a χ^2, e.g. (more details later):

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[\frac{\mu_i(\boldsymbol{\theta}) - n_i}{\sigma_i} \right]^2$$

- Use $\chi^2(\boldsymbol{\theta})$ to perform a statistical analysis

Outline

Event rates in oscillation experiments

number of events in a $\nu_\alpha \to \nu_\beta$ oscillation experiment:

$$N(\boldsymbol{\theta}) = T\mathcal{N} \int dE_\nu \, \phi_{\nu_\alpha}(E_\nu) \, P_{\alpha\beta}(E_\nu; \boldsymbol{\theta}) \, \sigma_{\nu_\beta}(E_\nu)$$

T	exposure time
$\mathcal{N}$	number of target particles
ϕ_{ν_α}	neutrino flux of flavour α at detector
$P_{\alpha\beta}$	$\nu_\alpha \to \nu_\beta$ oscillation probability
σ_{ν_β}	detection cross section of neutrino ν_β

Event rates in oscillation experiments

number of events in a $\nu_\alpha \to \nu_\beta$ oscillation experiment:

$$N(\boldsymbol{\theta}) = T\mathcal{N} \int dE_\nu \, \phi_{\nu_\alpha}(E_\nu) \, P_{\alpha\beta}(E_\nu; \boldsymbol{\theta}) \, \sigma_{\nu_\beta}(E_\nu)$$

- in more realistic situations we need to take into account the characteristics of the particular experiment
- consider in more detail the actual observables
- typically it will involve more integrals
  Ex.: atmospheric neutrinos: integrate also over zenith angle, production height in atmosphere, ...

to compare with observation add expected background in each bin:

$$\mu_i(\boldsymbol{\theta}) = N_i(\boldsymbol{\theta}) + B_i$$

$\rightarrow$ can be used to build χ^2, for example:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - n_i]^2}{n_i}$$

includes only statistical errors $\rightarrow$ on systematics see later

Example: Reactor experiments

- source of $\bar{\nu}_e$ with few MeV $\rightarrow$ $\bar{\nu}_e$ disappearance
- detection reaction: inverse beta-decay

$$\bar{\nu}_e + p \rightarrow n + e^+$$

observe positron and neutron in coincidence

- visible energy:

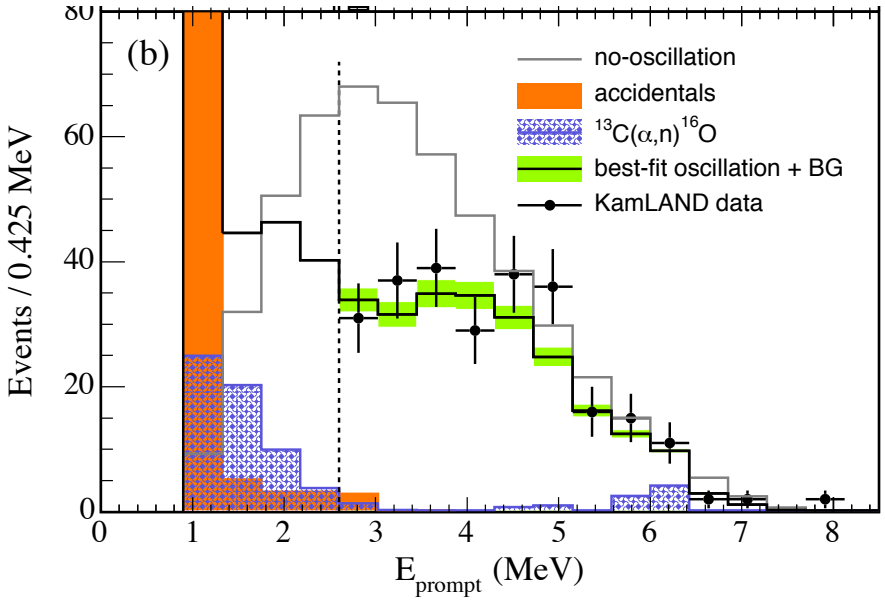
$$E_{\text{vis}} \approx E_{\text{kin}}^{e^+} + 2m_e = E_\nu - (m_n - m_p) + m_e + \mathcal{O}(E_\nu^2/m_n)$$
$$E_{\text{vis}} \approx E_\nu - 0.8\,\text{MeV}$$

$\rightarrow$ one-to-one relation between E_{vis} and E_ν

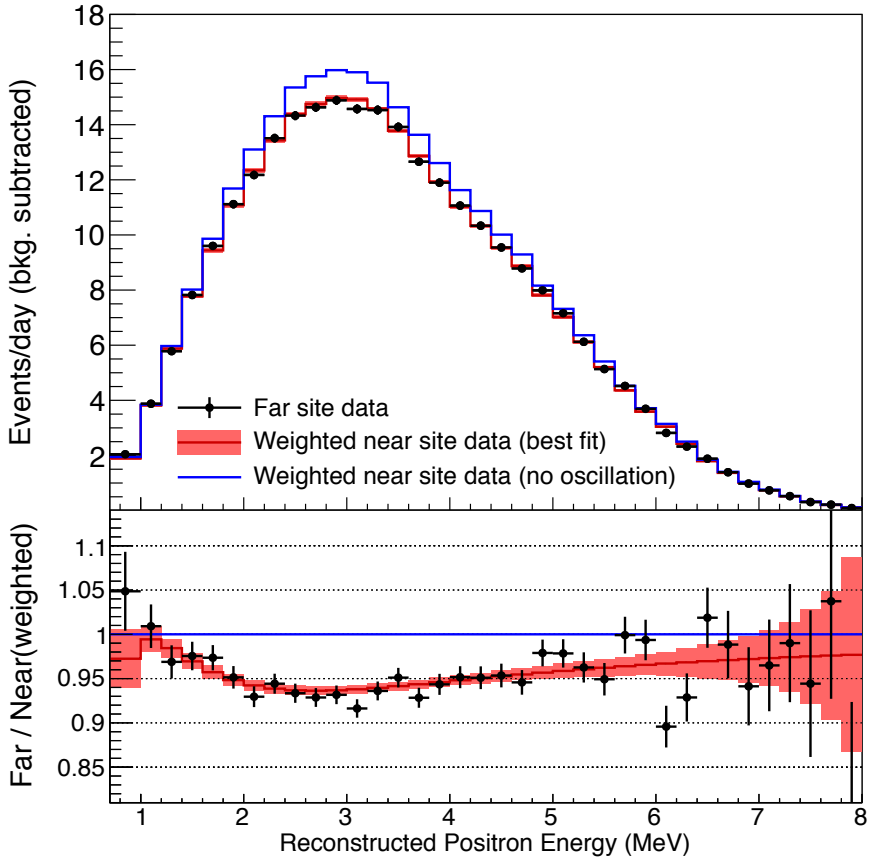
- accurate spectral information: number of inverse beta-decay events binned in visible energy

Example: Reactor experiments

KamLAND hep-ex/0406035



DayaBay 1505.03456



Number of events per bin

ideal experiment:

$$N_i(\boldsymbol{\theta}) = T\mathcal{N} \int_{E_{\mathrm{vis}}^{low,i}}^{E_{\mathrm{vis}}^{up,i}} dE_\nu \; \phi(E_\nu)\, P_{ee}(E_\nu;\boldsymbol{\theta})\, \sigma(E_\nu) \qquad E_\nu \approx E_{\mathrm{vis}} + 0.8\,\mathrm{MeV}$$

Number of events per bin

ideal experiment:

$$N_i(\boldsymbol{\theta}) = T\mathcal{N} \int_{E_{\mathrm{vis}}^{low,i}}^{E_{\mathrm{vis}}^{up,i}} dE_\nu \; \phi(E_\nu)\, P_{ee}(E_\nu;\boldsymbol{\theta})\, \sigma(E_\nu) \qquad E_\nu \approx E_{\mathrm{vis}} + 0.8\,\mathrm{MeV}$$

BUT: need to take into account energy resolution: a "true" $E_{\mathrm{vis}}^{\mathrm{true}}$ is reconstructed as E_{vis} with a certain probability distribution $R(E_{\mathrm{vis}}, E_{\mathrm{vis}}^{\mathrm{true}})$

$$N_i(\boldsymbol{\theta}) = T\mathcal{N} \int_{E_{\mathrm{vis}}^{low,i}}^{E_{\mathrm{vis}}^{up,i}} dE_{\mathrm{vis}} \int dE_\nu \, \phi(E_\nu)\, P_{ee}(E_\nu;\boldsymbol{\theta})\, \sigma(E_\nu) R(E_{\mathrm{vis}}, E_{\mathrm{vis}}^{\mathrm{true}})$$

$$E_\nu \approx E_{\mathrm{vis}}^{\mathrm{true}} + 0.8\,\mathrm{MeV}$$

can write this as

$$N_i(\boldsymbol{\theta}) = T\mathcal{N} \int dE_\nu \, \phi(E_\nu) \, P_{ee}(E_\nu; \boldsymbol{\theta}) \, \sigma(E_\nu) \, R_i(E_\nu)$$

$$R_i(E_\nu) \equiv \int_{E_{\text{vis}}^{low,i}}^{E_{\text{vis}}^{up,i}} dE_{\text{vis}} R(E_{\text{vis}}, E_{\text{vis}}^{\text{true}}) \qquad E_\nu \approx E_{\text{vis}}^{\text{true}} + 0.8\,\text{MeV}$$

often it is a good approximation to assume a Gaussian resolution function:

$$R(E_{\text{vis}}, E_{\text{vis}}^{\text{true}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(E_{\text{vis}} - E_{\text{vis}}^{\text{true}})^2}{2\sigma^2}\right] \qquad \sigma = \sigma(E_{\text{vis}}^{\text{true}})$$

$$R_i(E_\nu) = \frac{1}{2}\left[\text{erf}\left(\frac{E_{\text{vis}}^{up,i} - E_{\text{vis}}^{\text{true}}}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{E_{\text{vis}}^{low,i} - E_{\text{vis}}^{\text{true}}}{\sqrt{2}\sigma}\right)\right]$$

can write this as

$$N_i(\boldsymbol{\theta}) = T\mathcal{N} \int dE_\nu \, \phi(E_\nu) \, P_{ee}(E_\nu; \boldsymbol{\theta}) \, \sigma(E_\nu) \, R_i(E_\nu)$$

$$R_i(E_\nu) \equiv \int_{E_{\text{vis}}^{low,i}}^{E_{\text{vis}}^{up,i}} dE_{\text{vis}} R(E_{\text{vis}}, E_{\text{vis}}^{\text{true}}) \qquad E_\nu \approx E_{\text{vis}}^{\text{true}} + 0.8\,\text{MeV}$$

often it is a good approximation to assume a Gaussian resolution function:

$$R(E_{\text{vis}}, E_{\text{vis}}^{\text{true}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(E_{\text{vis}} - E_{\text{vis}}^{\text{true}})^2}{2\sigma^2} \right] \qquad \sigma = \sigma(E_{\text{vis}}^{\text{true}})$$

$$R_i(E_\nu) = \frac{1}{2}\left[\text{erf}\left(\frac{E_{\text{vis}}^{up,i} - E_{\text{vis}}^{\text{true}}}{\sqrt{2}\sigma} \right) - \text{erf}\left(\frac{E_{\text{vis}}^{low,i} - E_{\text{vis}}^{\text{true}}}{\sqrt{2}\sigma} \right) \right]$$

realistic "resolution function" $\rightarrow$ response matrix

Ex.: Daya Bay 1607.05378

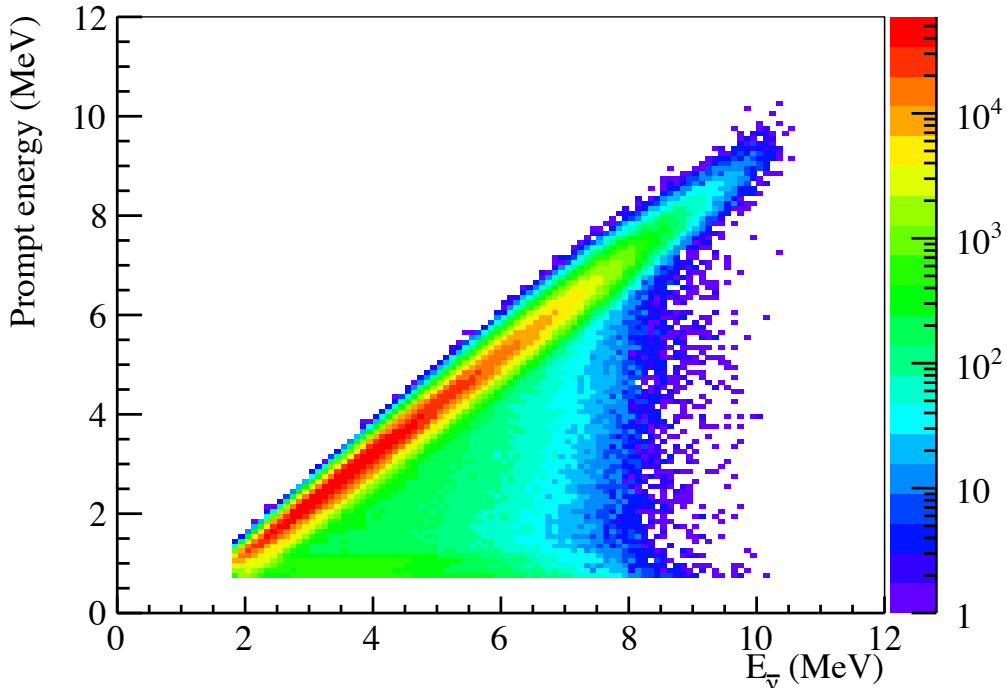


Fig. 21. The detector response matrix used to map antineutrino energy to the reconstructed energy. The IBD energy shift, IAV effect, nonlinearity, and energy resolution are included.

Ex.: KamLAND

multi-baseline source:

$$N_i(\boldsymbol{\theta}) = T\mathcal{N} \sum_r c_r \int dE_\nu \, \phi(E_\nu) \, P_{ee}(E_\nu, L_r; \boldsymbol{\theta}) \, \sigma(E_\nu) \, R_i(E_\nu)$$

$$c_r \propto \frac{P_r}{L_r^2}$$

Example: long-baseline experiment

- consider a $\nu_\mu \to \nu_e$ appearance experiment with $E_\nu \sim 1$ GeV (e.g., T2K, NOvA)
- detection reaction: $\nu_e + N \to e + X$
  significant energy is carried away by hadronic scattering products X

Example: long-baseline experiment

- consider a $\nu_\mu \to \nu_e$ appearance experiment with $E_\nu \sim 1$ GeV (e.g., T2K, NOvA)
- detection reaction: $\nu_e + N \to e + X$
  significant energy is carried away by hadronic scattering products X

assume only electron is observed and events are binned in electron energy

$$N_i(\boldsymbol{\theta}) = T\mathcal{N} \int dE_\nu \; \phi(E_\nu)\, P_{\mu e}(E_\nu; \boldsymbol{\theta}) \int_{E_e^{low,i}}^{E_e^{up,i}} dE_e \frac{d\sigma}{dE_e}(E_\nu)$$

$\to$ double integral even before including resolution function

Example: long-baseline experiment

- consider a $\nu_\mu \to \nu_e$ appearance experiment with $E_\nu \sim 1$ GeV (e.g., T2K, NOvA)
- detection reaction: $\nu_e + N \to e + X$
  significant energy is carried away by hadronic scattering products X

some detectors can use info on X to reconstruct $E_\nu \to$ bins in E_ν^{rec}

may require complicated cuts introducing energy dependent efficiences,...

Detector response function - migration matrix

$$N_i(\boldsymbol{\theta}) = T\mathcal{N} \int dE_\nu \, \phi(E_\nu) \, P_{\mu e}(E_\nu; \boldsymbol{\theta}) \, \sigma(E_\nu) \, \mathcal{R}_i(E_\nu)$$

$\mathcal{R}_i(E_\nu)$: detector response function

- describes the probability that an event with neutrino energy E_ν is reconstructed in the bin i
- the bins may label any observable (e.g., lepton energy, reconstr. neutrino energy, ...)
- $\mathcal{R}_i(E_\nu)$ can include many effects related to the detector (energy resolution, energy dep. efficiencies, differential cross sections, ...)
- if the integral over true neutrino energy is discretized $\mathcal{R}_i(E_\nu)$ becomes a matrix $\mathcal{R}_{ij} \rightarrow$ "migration matrix"

Detector response function - migration matrix

$$N_i(\boldsymbol{\theta}) = T\mathcal{N} \int dE_\nu \, \phi(E_\nu) \, P_{\mu e}(E_\nu; \boldsymbol{\theta}) \, \sigma(E_\nu) \, \mathcal{R}_i(E_\nu)$$

$\mathcal{R}_i(E_\nu)$: detector response function

can be conveniently done with the GLoBES software package
Huber, Lindner, Winter, hep-ph/0407333; Huber et al., hep-ph/0701187
http://www.mpi-hd.mpg.de/lin/globes/

Example: atmospheric neutrinos

consider an experiment observing muons induced by atmospheric neutrinos (e.g., INO, IceCube):

$$N_{ij}(\boldsymbol{\theta}) = T\mathcal{N} \int dE_\nu \int d\Omega\, \sigma(E_\nu)\, \mathcal{R}_{ij}(E_\nu, \Omega) \times \left[\phi_\mu(E_\nu, \Omega)\, P_{\mu\mu}(E_\nu, \Omega; \boldsymbol{\theta}) + \phi_e(E_\nu, \Omega)\, P_{e\mu}(E_\nu, \Omega; \boldsymbol{\theta})\right]$$

i	bin in muon energy
j	bin in muon zenith angle
$\phi_\alpha(E_\nu, \Omega)$	flux of ν_α with given E_ν and solid angle Ω

$\mathcal{R}_{ij}(E_\nu, \Omega)$: probability to reconstruct muon from a neutrino with energy E_ν coming from a solid angle Ω into the muon bin ij (includes double differential cross section)

(still simplified in several respects....)

Outline

- Can define:

$$\chi^2 = \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - n_i]^2}{\mu_i(\boldsymbol{\theta})} \quad \text{or} \quad \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - n_i]^2}{n_i}$$

- If the number of events is small in some bins ("Poisson χ^2"):

$$\chi^2 = 2 \sum_{i=1}^{N} \left[\mu_i(\theta) - n_i + n_i \log \frac{n_i}{\mu_i(\theta)} \right]$$

- If statistical errors include the ones from a subtracted background:

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{\mu_i(\theta) - n_i}{\sigma_i} \right]^2$$

- If there is correlation between bins:

$$\chi^2 = \sum_{i,j=1}^{N} [\mu_i(\theta) - n_i] V_{ij}^{-1} [\mu_j(\theta) - n_j]$$

- Can define:

$$\chi^2 = \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - n_i]^2}{\mu_i(\boldsymbol{\theta})} \quad \text{or} \quad \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - n_i]^2}{n_i}$$

- If the number of events is small in some bins ("Poisson χ^2"):

$$\chi^2 = 2 \sum_{i=1}^{N} \left[\mu_i(\boldsymbol{\theta}) - n_i + n_i \log \frac{n_i}{\mu_i(\boldsymbol{\theta})} \right]$$

- If statistical errors include the ones from a subtracted background:

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{\mu_i(\boldsymbol{\theta}) - n_i}{\sigma_i} \right]^2$$

- If there is correlation between bins:

$$\chi^2 = \sum_{i,j=1}^{N} [\mu_i(\boldsymbol{\theta}) - n_i] V_{ij}^{-1} [\mu_j(\boldsymbol{\theta}) - n_j]$$

- Can define:

$$\chi^2 = \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - n_i]^2}{\mu_i(\boldsymbol{\theta})} \quad \text{or} \quad \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - n_i]^2}{n_i}$$

- If the number of events is small in some bins ("Poisson χ^2"):

$$\chi^2 = 2 \sum_{i=1}^{N} \left[\mu_i(\boldsymbol{\theta}) - n_i + n_i \log \frac{n_i}{\mu_i(\boldsymbol{\theta})} \right]$$

- If statistical errors include the ones from a subtracted background:

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{\mu_i(\boldsymbol{\theta}) - n_i}{\sigma_i} \right]^2$$

- If there is correlation between bins:

$$\chi^2 = \sum_{i,j=1}^{N} [\mu_i(\boldsymbol{\theta}) - n_i] V_{ij}^{-1} [\mu_j(\boldsymbol{\theta}) - n_j]$$

- Can define:

$$\chi^2 = \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - n_i]^2}{\mu_i(\boldsymbol{\theta})} \quad \text{or} \quad \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - n_i]^2}{n_i}$$

- If the number of events is small in some bins ("Poisson χ^2"):

$$\chi^2 = 2 \sum_{i=1}^{N} \left[\mu_i(\boldsymbol{\theta}) - n_i + n_i \log \frac{n_i}{\mu_i(\boldsymbol{\theta})} \right]$$

- If statistical errors include the ones from a subtracted background:

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{\mu_i(\boldsymbol{\theta}) - n_i}{\sigma_i} \right]^2$$

- If there is correlation between bins:

$$\chi^2 = \sum_{i,j=1}^{N} [\mu_i(\boldsymbol{\theta}) - n_i] V_{ij}^{-1} [\mu_j(\boldsymbol{\theta}) - n_j]$$

Systematic uncertainties

Assume we have N experimental data points n_i with statistical error σ_i and theoretical predictions μ_i for each of the data points:

$$\chi^2 = \sum_{i=1}^{N} \frac{(\mu_i - n_i)^2}{\sigma_i^2}$$

$\mu_i(\theta)$ depends on the parameters of the model θ.

Consider the situation that μ_i depends also on additional parameters ξ, describing systematical uncertainties ("nuisance parameters"): $\mu_i(\theta, \xi)$

We may have some knowledge on ξ: mean values $\langle \xi_\alpha \rangle = \hat{\xi}_\alpha$ and uncertainty σ_α^ξ

Example

$$\mu_i(\theta) = \xi_1 \left(\xi_2 N_i(\theta) + \xi_3 B_i\right) \qquad \xi_\alpha = 1 \pm x_\alpha\%$$

$$\approx (1 + \delta_1 + \delta_2) N_i(\theta) + (1 + \delta_1 + \delta_3) B_i \qquad \delta_\alpha = \xi_\alpha - 1$$

- ξ_1 overall detector normalization
- ξ_2 overall signal normalization (e.g., flux uncertainty)
- ξ_3 background normalization

can be generalized to more complicated systematics, including energy dependent uncertainties (shape), energy scale,...

Example

$$\mu_i(\theta) = \xi_1 \left(\xi_2 N_i(\theta) + \xi_3 B_i\right) \qquad \xi_\alpha = 1 \pm x_\alpha\%$$

$$\approx (1 + \delta_1 + \delta_2) N_i(\theta) + (1 + \delta_1 + \delta_3) B_i \qquad \delta_\alpha = \xi_\alpha - 1$$

- ξ_1 overall detector normalization
- ξ_2 overall signal normalization (e.g., flux uncertainty)
- ξ_3 background normalization

can be generalized to more complicated systematics, including energy dependent uncertainties (shape), energy scale,...

Consider ξ at the same level as θ and add info to χ^2

$$\chi^2(\theta,\xi) = \sum_{i=1}^{N} \frac{[\mu_i(\theta,\xi) - n_i]^2}{\sigma_i^2} + \sum_{\alpha} \frac{(\xi_\alpha - \hat{\xi}_\alpha)^2}{(\sigma_\alpha^\xi)^2}$$

$$\chi^2(\theta) = \min_{\xi} \chi^2(\theta,\xi)$$

$\chi^2(\theta)$ is distributed as usual with $N = (N-P) + P$ dof

no conceptual issue also for $P \gtrsim N$

Consider ξ at the same level as θ and add info to χ^2

$$\chi^2(\theta, \xi) = \sum_{i=1}^{N} \frac{[\mu_i(\theta, \xi) - n_i]^2}{\sigma_i^2} + \sum_{\alpha} \frac{(\xi_\alpha - \hat{\xi}_\alpha)^2}{(\sigma_\alpha^\xi)^2}$$

$$\chi^2(\theta) = \min_{\xi} \chi^2(\theta, \xi)$$

$\chi^2(\theta)$ is distributed as usual with $N = (N - P) + P$ dof

no conceptual issue also for $P \gtrsim N$

Linearize the problem

$$\mu_i(\theta,\xi) \approx \mu_i(\theta,\hat{\xi}) + \sum_\alpha \frac{\partial \mu_i}{\partial \xi_\alpha}(\xi_\alpha - \hat{\xi}_\alpha)$$

define: $\mu_i(\theta,\hat{\xi}) \equiv \hat{\mu}_i(\theta)\,, \qquad \xi'_\alpha \equiv \frac{\xi_\alpha - \hat{\xi}_\alpha}{\sigma^\xi_\alpha}\,, \qquad R_{i\alpha} \equiv \sigma^\xi_\alpha \frac{\partial \mu_i}{\partial \xi_\alpha}$

$$\chi^2(\theta,\xi') = \sum_i \frac{[\hat{\mu}_i(\theta) + \sum_\alpha R_{i\alpha}\xi'_\alpha - n_i]^2}{\sigma_i^2} + \sum_\alpha {\xi'_\alpha}^2$$

$\chi^2(\theta,\xi')$ is quadratic in $\xi' \Rightarrow \frac{\partial \chi^2}{\partial \xi_\alpha} = 0$ is a linear system of equations

$\Rightarrow$ solve the system to obtain ξ_{min} and obtain $\chi^2(\theta) = \chi^2(\theta,\xi_{min})$

- this proceedure works fine if $\xi'_\alpha \lesssim 1$ and $(R\xi')_i \ll \mu_i$
- if $(R\xi')_i \sim \mu_i$, the prediction can become negative

Linearize the problem

$$\mu_i(\theta,\xi) \approx \mu_i(\theta,\hat{\xi}) + \sum_\alpha \frac{\partial \mu_i}{\partial \xi_\alpha}(\xi_\alpha - \hat{\xi}_\alpha)$$

define: $\mu_i(\theta,\hat{\xi}) \equiv \hat{\mu}_i(\theta)\,, \qquad \xi'_\alpha \equiv \frac{\xi_\alpha - \hat{\xi}_\alpha}{\sigma^\xi_\alpha}\,, \qquad R_{i\alpha} \equiv \sigma^\xi_\alpha \frac{\partial \mu_i}{\partial \xi_\alpha}$

$$\chi^2(\theta,\xi') = \sum_i \frac{[\hat{\mu}_i(\theta) + \sum_\alpha R_{i\alpha}\xi'_\alpha - n_i]^2}{\sigma_i^2} + \sum_\alpha {\xi'_\alpha}^2$$

$\chi^2(\theta,\xi')$ is quadratic in $\xi' \Rightarrow \frac{\partial \chi^2}{\partial \xi_\alpha} = 0$ is a linear system of equations
$\Rightarrow$ solve the system to obtain ξ_{min} and obtain $\chi^2(\theta) = \chi^2(\theta,\xi_{min})$

- this proceedure works fine if $\xi'_\alpha \lesssim 1$ and $(R\xi')_i \ll \mu_i$
- if $(R\xi')_i \sim \mu_i$, the prediction can become negative

Equivalence of pull and covariance approaches

- "pull" approach:

$$\chi^2_{\text{pull}}(\theta) = \min_{\xi} \chi^2(\theta, \xi)$$

- "covariance" approach:

$$V_{ij} = \sum_{\alpha} \frac{\partial \mu_i}{\partial \xi_\alpha} \frac{\partial \mu_j}{\partial \xi_\alpha} (\sigma^\xi_\alpha)^2 = \sum_{\alpha} R_{i\alpha} R_{j\alpha}$$

$$\chi^2_{\text{cov}}(\theta) = \sum_{ij} [\hat{\mu}_i(\theta) - n_i]^T S^{-1}_{ij} [\hat{\mu}_j(\theta) - n_j] \quad \text{with} \quad S_{ij} \equiv \sigma_i^2 \delta_{ij} + V_{ij}$$

Equivalence of pull and covariance approaches

- "pull" approach:

$$\chi^2_{\text{pull}}(\theta) = \min_{\xi} \chi^2(\theta, \xi)$$

- "covariance" approach:

$$V_{ij} = \sum_{\alpha} \frac{\partial \mu_i}{\partial \xi_\alpha} \frac{\partial \mu_j}{\partial \xi_\alpha} (\sigma^\xi_\alpha)^2 = \sum_{\alpha} R_{i\alpha} R_{j\alpha}$$

$$\chi^2_{\text{cov}}(\theta) = \sum_{ij} [\hat{\mu}_i(\theta) - n_i]^T S^{-1}_{ij} [\hat{\mu}_j(\theta) - n_j] \quad \text{with} \quad S_{ij} \equiv \sigma^2_i \delta_{ij} + V_{ij}$$

Exercise: proof that $\chi^2_{\text{pull}}(\theta) \equiv \chi^2_{\text{cov}}(\theta)$

Fogli, Lisi, Marrone, Montanino, Palazzo, PRD02 [hep-ph/0206162]

Simple example

Consider the case of a single systematic describing an over-all normalization uncertainty

$$\chi^2(\theta, \xi) = \sum_i \left[\frac{\mu_i(\theta)(1+\xi) - n_i}{\sigma_i} \right]^2 + \left(\frac{\xi}{\sigma_\xi} \right)^2$$

$$R_i = \mu_i(\theta)$$

covariance matrix for the covariance method: $S_{ij} = \delta_{ij}\sigma_i^2 + \mu_i \mu_j \sigma_\xi^2$

Simple example

Consider the case of a single systematic describing an over-all normalization uncertainty

$$\chi^2(\theta, \xi) = \sum_i \left[\frac{\mu_i(\theta)(1+\xi) - n_i}{\sigma_i} \right]^2 + \left(\frac{\xi}{\sigma_\xi} \right)^2$$

$$R_i = \mu_i(\theta)$$

covariance matrix for the covariance method: $S_{ij} = \delta_{ij}\sigma_i^2 + \mu_i \mu_j \sigma_\xi^2$

Exercise:

- minimize the χ^2 and calculate ξ_{min} and $\chi^2(\theta, \xi_{min})$
- consider the same systematic using the Poisson χ^2
  (check that your solution makes sense!)

Simple example

Consider the case of a single systematic describing an over-all normalization uncertainty

$$\chi^2(\theta,\xi) = \sum_i \left[\frac{\mu_i(\theta)(1+\xi) - n_i}{\sigma_i}\right]^2 + \left(\frac{\xi}{\sigma_\xi}\right)^2$$

$$R_i = \mu_i(\theta)$$

covariance matrix for the covariance method: $S_{ij} = \delta_{ij}\sigma_i^2 + \mu_i\mu_j\sigma_\xi^2$

for $\sigma_\xi \to \infty$ this corresponds to a shape-only analysis (free normalization)

exactly this method has been used by the Daya Bay collaboration for their 2012 analysis based on near-far comparison

Real-life example Daya Bay 1203.1669

The value of $\sin^2 2\theta_{13}$ was determined with a χ^2 constructed with pull terms accounting for the correlation of the systematic errors [28],

$$\chi^2 = \sum_{d=1}^{6} \frac{\left[M_d - T_d\left(1+\varepsilon+\sum_r \omega_r^d \alpha_r + \varepsilon_d\right) + \eta_d\right]^2}{M_d + B_d} + \sum_r \frac{\alpha_r^2}{\sigma_r^2} + \sum_{d=1}^{6}\left(\frac{\varepsilon_d^2}{\sigma_d^2} + \frac{\eta_d^2}{\sigma_B^2}\right), \quad (2)$$

where M_d are the measured IBD events of the d-th AD with backgrounds subtracted, B_d is the corresponding background, T_d is the prediction from neutrino flux, MC, and neutrino oscillations [29], ω_r^d is the fraction of IBD contribution of the r-th reactor to the d-th AD determined by baselines and reactor fluxes. The uncertainties are listed in Table III. The uncorrelated reactor uncertainty is σ_r (0.8%), σ_d (0.2%) is the uncorrelated detection uncertainty, and σ_B is the background uncertainty listed in Table II. The corresponding pull parameters are $(\alpha_r, \varepsilon_d, \eta_d)$. The detector- and reactor-related correlated

Real-life example Daya Bay 1203.1669

Exercise: study the χ^2 used in the Daya Bay paper

The value of $\sin^2 2\theta_{13}$ was determined with a χ^2 constructed with pull terms accounting for the correlation of the systematic errors [28],

$$\chi^2 = \sum_{d=1}^{6} \frac{\left[M_d - T_d\left(1+\varepsilon+\sum_r \omega_r^d \alpha_r + \varepsilon_d\right) + \eta_d\right]^2}{M_d + B_d} + \sum_r \frac{\alpha_r^2}{\sigma_r^2} + \sum_{d=1}^{6}\left(\frac{\varepsilon_d^2}{\sigma_d^2} + \frac{\eta_d^2}{\sigma_B^2}\right), \quad (2)$$

where M_d are the measured IBD events of the d-th AD with backgrounds subtracted, B_d is the corresponding background, T_d is the prediction from neutrino flux, MC, and neutrino oscillations [29], ω_r^d is the fraction of IBD contribution of the r-th reactor to the d-th AD determined by baselines and reactor fluxes. The uncertainties are listed in Table III. The uncorrelated reactor uncertainty is σ_r (0.8%), σ_d (0.2%) is the uncorrelated detection uncertainty, and σ_B is the background uncertainty listed in Table II. The corresponding pull parameters

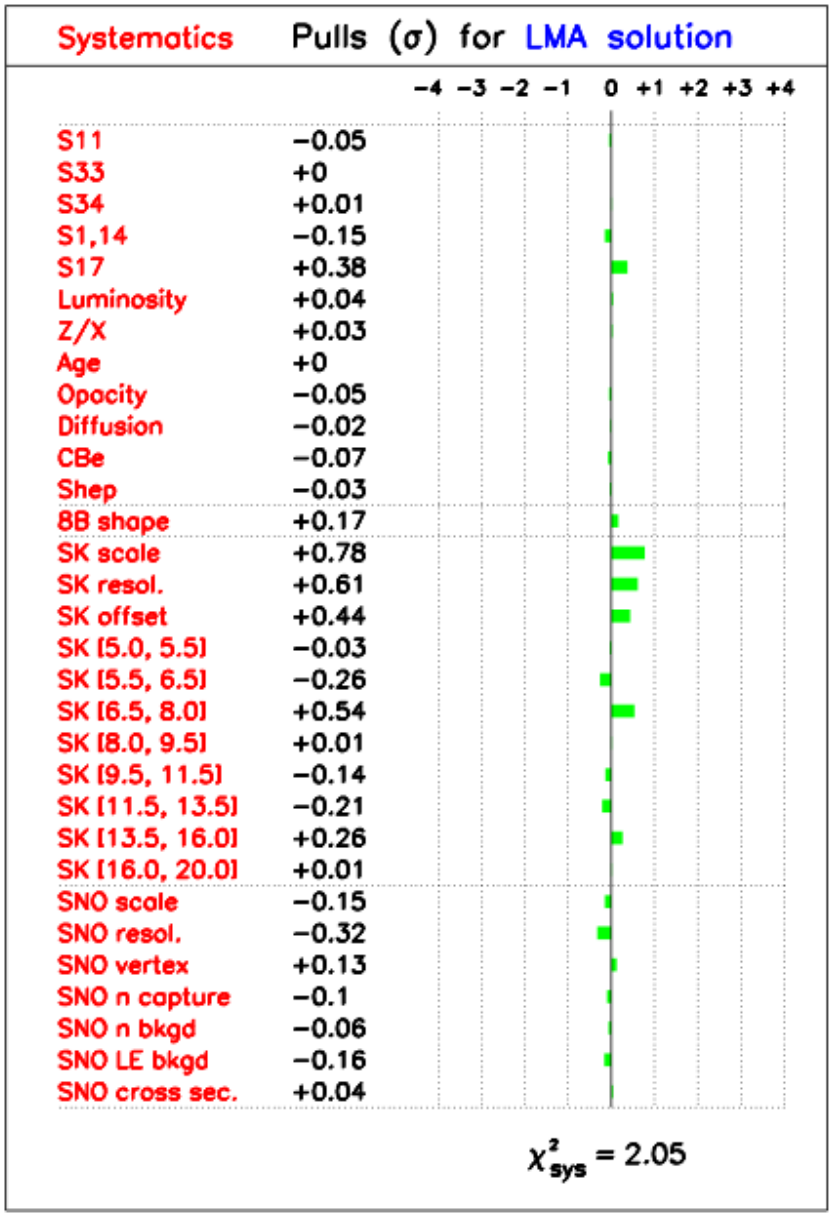
Pull versus covariance approaches

- Pull approach requires to solve a linear system of equations of dimension P (number of pulls)
- Covariance approach requires to invert the $N \times N$ covariance matrix (N number of bins)
- Depending on whether N is larger or smaller than P one or the other method may be preferred (often $P \ll N$)
- Pull method allows for more diagnostics of the fit, e.g.:
  - look at $\xi_{\alpha min}$ to identify a systematic with large "pull",
  - look at contours of θ versus ξ to identify correlations between systematics and parameters

Pull versus covariance approaches

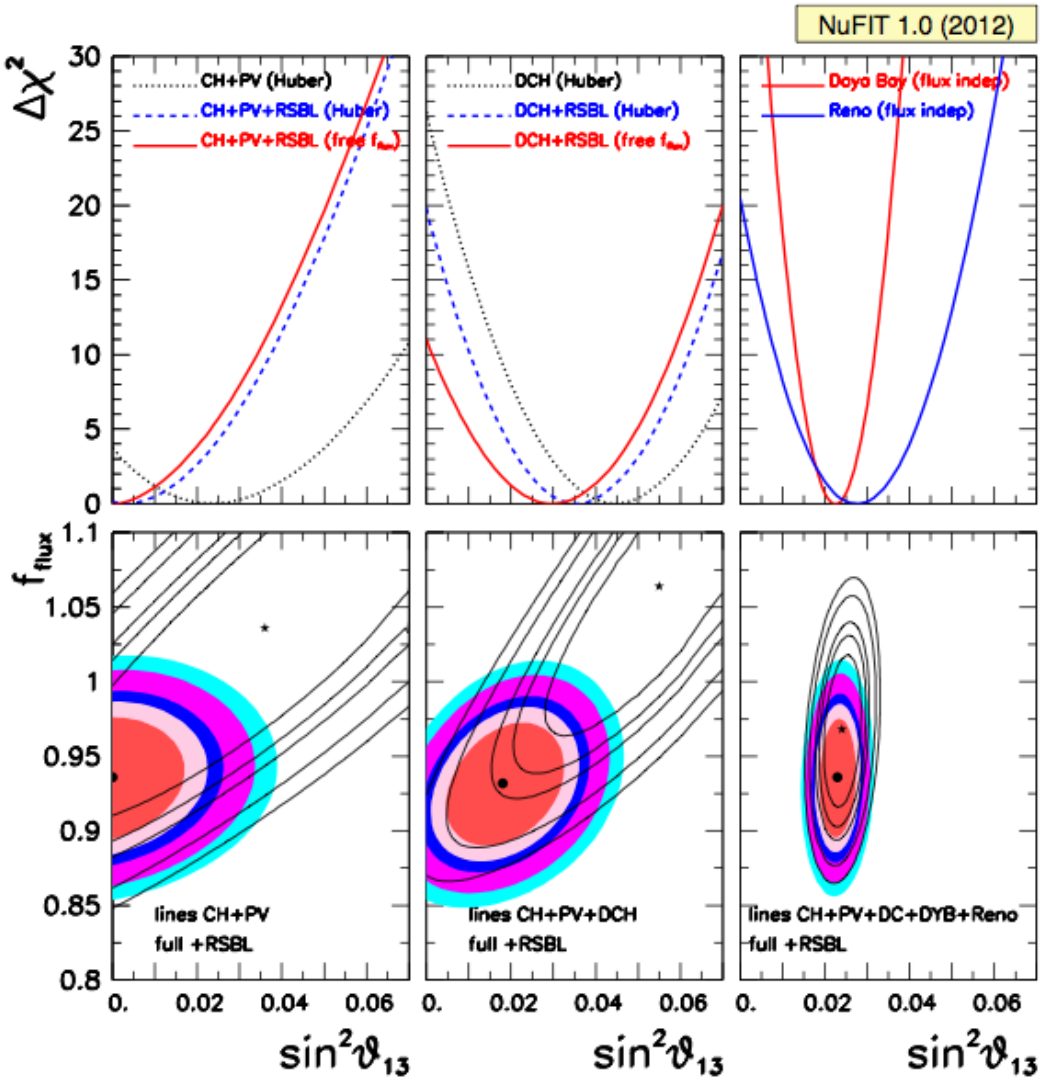
- Pull approach requires to solve a linear system of equations of dimension P (number of pulls)
- Covariance approach requires to invert the $N \times N$ covariance matrix (N number of bins)
- Depending on whether N is larger or smaller than P one or the other method may be preferred (often $P \ll N$)
- Pull method allows for more diagnostics of the fit, e.g.:
  - look at $\xi_{\alpha min}$ to identify a systematic with large "pull",
  - look at contours of θ versus ξ to identify correlations between systematics and parameters

Example for "pull diagram" from solar neutrino fit



Fogli et al hep-ph/0206162

Correlations between reactor flux normalization and θ_{13}



Poisson χ^2

The pull method can be generalized to the Poissonian form of the χ^2 which should be used in case of small event numbers per bin:

$$\chi^2(\theta, \xi_\alpha) = 2 \sum_{i=1}^{N} \left[\mu_i(\theta, \xi_\alpha) - n_i + n_i \log \frac{n_i}{\mu_i(\theta, \xi_\alpha)} \right] + \sum_\alpha \xi_\alpha^2$$

- allows to introduce correlated errors in the Poisson χ^2
- $\mu(\theta, \xi)$ can still be linearized in ξ, but the χ^2 will no longer be a quadratic function in $\xi \Rightarrow$ have to use numerical or semi-analytic methods to do the minimization

Comments - 1

- straight forward to generalize to correlated data and/or pulls:

$$\chi^2(\theta,\xi) = \sum_{i,j=1}^{N} [\mu_i(\theta,\xi) - n_i] V_{ij}^{-1} [\mu_j(\theta,\xi) - n_j] + \sum_{\alpha,\beta} (\xi_\alpha - \hat{\xi}_\alpha) W_{\alpha\beta}^{-1} (\xi_\beta - \hat{\xi}_\beta)$$

- can also be applied in the framework of likelihood analysis

$$\mathcal{L}(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi) \times \mathcal{L}_{\text{nuis}}(\xi)$$
$$\mathcal{L}(\theta) = \max_\xi \mathcal{L}(\theta,\xi)$$

$\mathcal{L}_{\text{nuis}}(\xi)$ contains all information we have on the nuisance parameters

If $\mathcal{L}(\theta,\xi)$ and/or $\mathcal{L}_{\text{nuis}}(\xi)$ are "complicated" the minimization (maximization) has to be done numerically.

Comments - 1

- straight forward to generalize to correlated data and/or pulls:

$$\chi^2(\theta,\xi) = \sum_{i,j=1}^{N} [\mu_i(\theta,\xi) - n_i] V_{ij}^{-1} [\mu_j(\theta,\xi) - n_j] + \sum_{\alpha,\beta} (\xi_\alpha - \hat{\xi}_\alpha) W_{\alpha\beta}^{-1} (\xi_\beta - \hat{\xi}_\beta)$$

- can also be applied in the framework of likelihood analysis

$$\mathcal{L}(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi) \times \mathcal{L}_{\text{nuis}}(\xi)$$
$$\mathcal{L}(\theta) = \max_\xi \mathcal{L}(\theta,\xi)$$

$\mathcal{L}_{\text{nuis}}(\xi)$ contains all information we have on the nuisance parameters

If $\mathcal{L}(\theta,\xi)$ and/or $\mathcal{L}_{\text{nuis}}(\xi)$ are "complicated" the minimization (maximization) has to be done numerically.

Comments - 2

- The methods discussed here for the treatment of systematic erros assume that systematic uncertainties are of statistical nature. Their effects on the analysis are encoded by assuming some random distribution for them (often Gaussian).
- Sometimes these assumptions are justified e.g. when the origin of the uncertainty is some measurment (e.g., normalization uncertainty).
- Sometimes these assumptions are not justified, in case of true "theoretical uncertainties" (e.g. nuclear matrix elements for neutrino-less double-beta decay).
- Frequentist interpretation in the strict sense is not clear
- pull method fits very natural in Bayesian framework:

$$\mathcal{L}(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi) \times \mathcal{L}_{\text{nuis}}(\xi) \quad \rightarrow$$

$$f(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi)\pi(\theta)\pi(\xi) \quad \rightarrow \quad f(\theta) = \int d\xi f(\theta,\xi)$$

Comments - 2

- The methods discussed here for the treatment of systematic erros assume that systematic uncertainties are of statistical nature. Their effects on the analysis are encoded by assuming some random distribution for them (often Gaussian).
- Sometimes these assumptions are justified e.g. when the origin of the uncertainty is some measurment (e.g., normalization uncertainty).
- Sometimes these assumptions are not justified, in case of true "theoretical uncertainties" (e.g. nuclear matrix elements for neutrino-less double-beta decay).
- Frequentist interpretation in the strict sense is not clear
- pull method fits very natural in Bayesian framework:

$$\mathcal{L}(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi) \times \mathcal{L}_{\text{nuis}}(\xi) \quad \rightarrow$$

$$f(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi)\pi(\theta)\pi(\xi) \quad \rightarrow \quad f(\theta) = \int d\xi f(\theta,\xi)$$

Comments - 2

- The methods discussed here for the treatment of systematic erros assume that systematic uncertainties are of statistical nature. Their effects on the analysis are encoded by assuming some random distribution for them (often Gaussian).
- Sometimes these assumptions are justified e.g. when the origin of the uncertainty is some measurment (e.g., normalization uncertainty).
- Sometimes these assumptions are not justified, in case of true "theoretical uncertainties" (e.g. nuclear matrix elements for neutrino-less double-beta decay).
- Frequentist interpretation in the strict sense is not clear
- pull method fits very natural in Bayesian framework:

$$\mathcal{L}(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi) \times \mathcal{L}_{\text{nuis}}(\xi) \quad \rightarrow$$

$$f(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi)\pi(\theta)\pi(\xi) \quad \rightarrow \quad f(\theta) = \int d\xi f(\theta,\xi)$$

Comments - 2

- The methods discussed here for the treatment of systematic erros assume that systematic uncertainties are of statistical nature. Their effects on the analysis are encoded by assuming some random distribution for them (often Gaussian).
- Sometimes these assumptions are justified e.g. when the origin of the uncertainty is some measurment (e.g., normalization uncertainty).
- Sometimes these assumptions are not justified, in case of true "theoretical uncertainties" (e.g. nuclear matrix elements for neutrino-less double-beta decay).
- Frequentist interpretation in the strict sense is not clear
- pull method fits very natural in Bayesian framework:

$$\mathcal{L}(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi) \times \mathcal{L}_{\text{nuis}}(\xi) \quad \rightarrow$$

$$f(\theta,\xi) = \mathcal{L}_{\text{data}}(\theta,\xi)\pi(\theta)\pi(\xi) \quad \rightarrow \quad f(\theta) = \int d\xi f(\theta,\xi)$$

Referenzes on pull method in neutrino context

- in the context of solar neutrinos
  G. L. Fogli, E. Lisi, A. Marrone, D. Montanino and A. Palazzo, Phys. Rev. D **66** (2002) 053010 [hep-ph/0206162]
- in the context of short-baseline oscillation experiments
  T. Schwetz, PhD thesis, Univ. Vienna 2002, see appendix A, available at request
- in the context of SuperKamiokande atmospheric neutrinos
  M. C. Gonzalez-Garcia and M. Maltoni, Phys. Rept. **460** (2008) 1 [arXiv:0704.1800], see appendix A
- in the context of future long-baseline oscillation experiment simulation
  P. Huber, M. Mezzetto and T. Schwetz, JHEP **0803** (2008) 021 [arXiv:0711.2950]

Outline

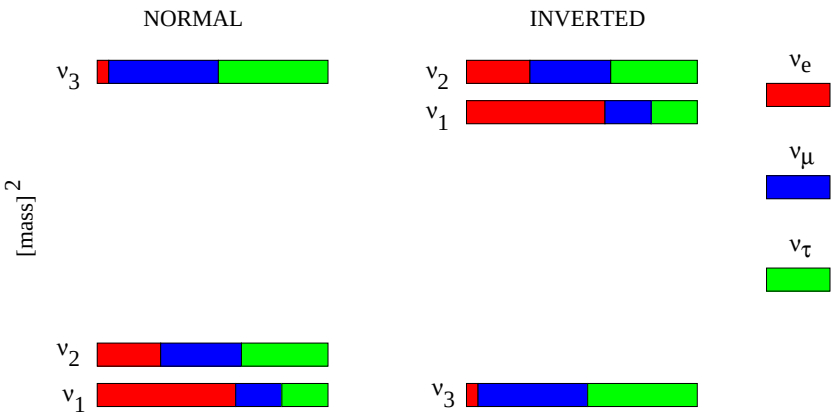
Hypothesis testing

- Want to decide whether data allows to reject or favour an hypothesis H_0 over an alternative hypothesis H_1

- simple hypotheses: depend on no free parameters
- composite hypotheses: depend on free parameters θ to be estimated from the data

Hypothesis testing

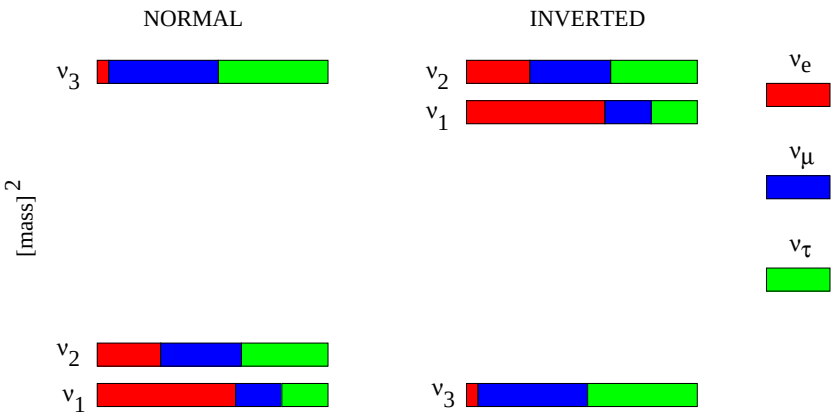
- Want to decide whether data allows to reject or favour an hypothesis H_0 over an alternative hypothesis H_1



- simple hypotheses: depend on no free parameters
- composite hypotheses: depend on free parameters θ to be estimated from the data

Hypothesis testing

- Want to decide whether data allows to reject or favour an hypothesis H_0 over an alternative hypothesis H_1

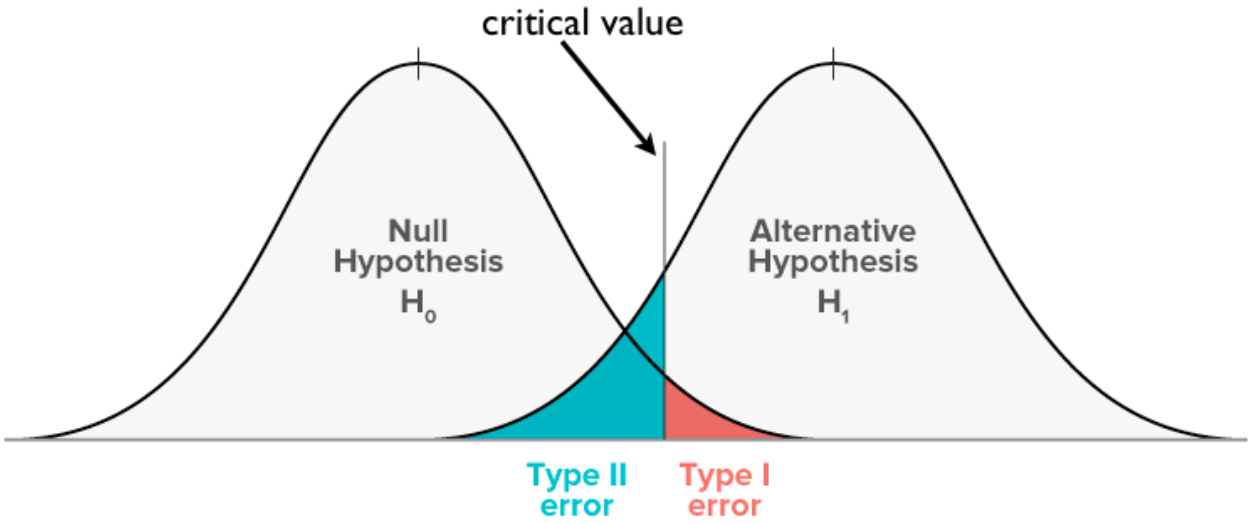


- simple hypotheses: depend on no free parameters
- composite hypotheses: depend on free parameters θ to be estimated from the data

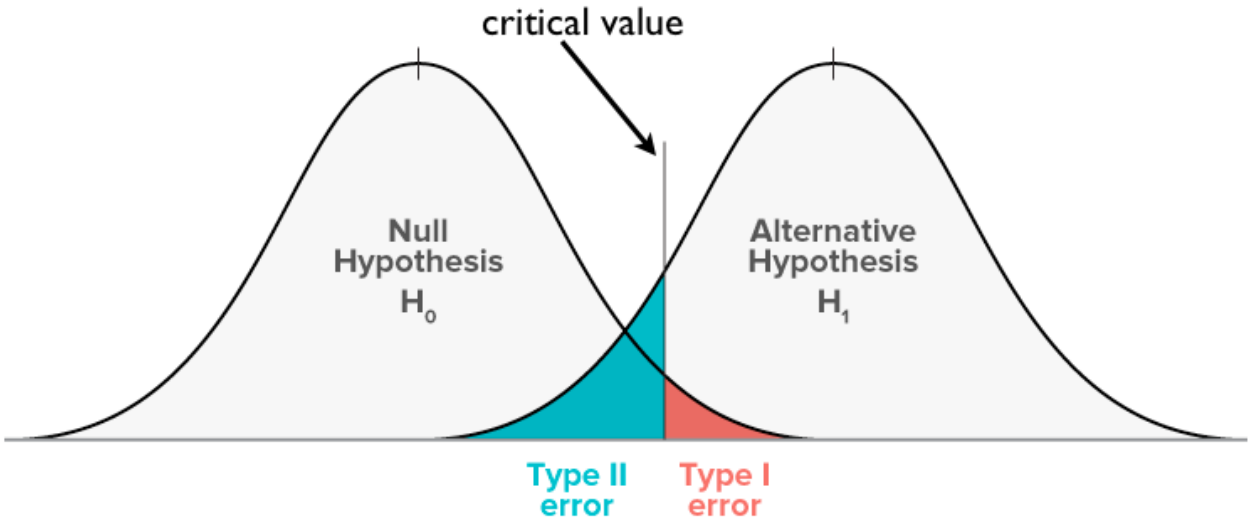
Simple hypotheses

- consider two simple hypotheses: H_0, H_1
- chose some statistic T (function of random variables, i.e., data)
- each hypothesis predicts a pdf for T: $f(T|H_i)$
- chose T such that small values favour H_0 and large values favour H_1

Simple hypotheses - errors of 1st and 2nd kind



Simple hypotheses - errors of 1st and 2nd kind

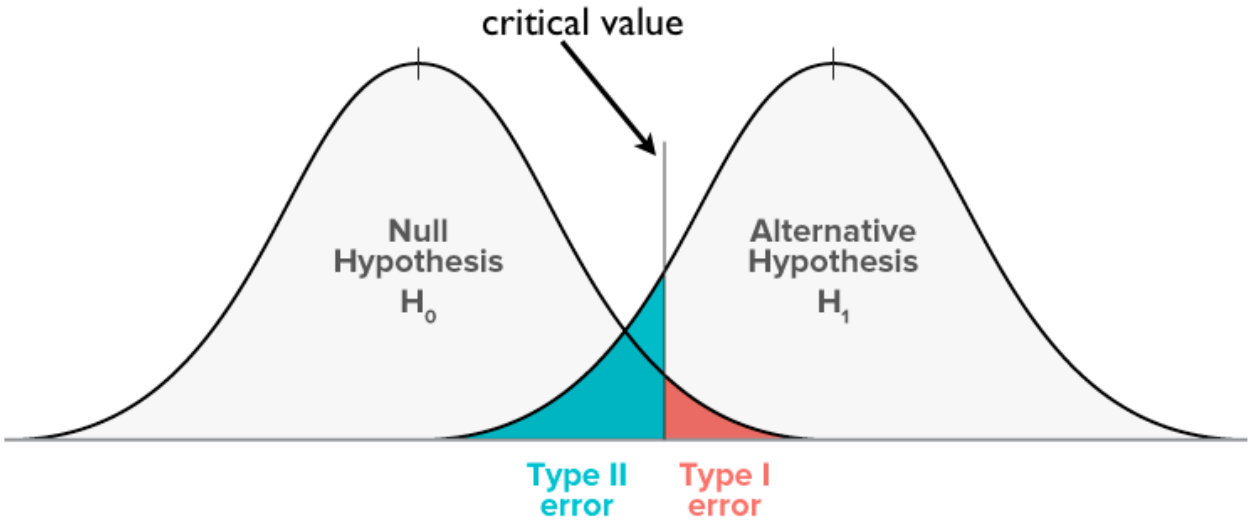


reject H_0 at the CL $(1-\alpha)$ if $T > T_c$ such that

$$\int_{T_c}^{\infty} dT\, f(T|H_0) = \alpha$$

α is the probability of rejecting H_0 although it is true
$\Rightarrow$ "error of the first kind"

Simple hypotheses - errors of 1st and 2nd kind



probab β of accepting H_0 although the alternative H_1 is true

$$\int_{-\infty}^{T_c} dT\, f(T|H_1) = \beta$$

$\Rightarrow$ "error of the second kind" or "power of the test" $(1-\beta)$

Simple hypotheses - comments

- a common choice (the "optimal one") is the likelihood ratio

$$T = \frac{f(x|H_1)}{f(x|H_0)} = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)}$$

- very often we deal with composite hypotheses...

Simple hypotheses - comments

- a common choice (the "optimal one") is the likelihood ratio

$$T = \frac{f(x|H_1)}{f(x|H_0)} = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)}$$

- very often we deal with composite hypotheses...

Composite hypotheses

- H_0 simple, H_1 composite
- H_0 composite, H_1 simple
- both H_0 and H_1 composite

if H_0 composite: need to reject it for all values of $\theta \in H_0$

$$\int_{T_c(\theta)}^{\infty} dT\, f(T|H_0(\theta)) = \alpha \quad \rightarrow \quad T_c = \max_{\theta \in H_0} T_c(\theta)$$

if H_1 composite: β will depend on $\theta \in H_1$

$$\int_{-\infty}^{T_c} dT\, f(T|H_1(\theta)) = \beta(\theta)$$

can quote e.g. "best" and "worst" power of the test

Composite hypotheses

- H_0 simple, H_1 composite
- H_0 composite, H_1 simple
- both H_0 and H_1 composite

if H_0 composite: need to reject it for all values of $\theta \in H_0$

$$\int_{T_c(\theta)}^{\infty} dT\, f(T|H_0(\theta)) = \alpha \quad \rightarrow \quad T_c = \max_{\theta \in H_0} T_c(\theta)$$

if H_1 composite: β will depend on $\theta \in H_1$

$$\int_{-\infty}^{T_c} dT\, f(T|H_1(\theta)) = \beta(\theta)$$

can quote e.g. "best" and "worst" power of the test

Composite hypotheses

- H_0 simple, H_1 composite
- H_0 composite, H_1 simple
- both H_0 and H_1 composite

if H_0 composite: need to reject it for all values of $\theta \in H_0$

$$\int_{T_c(\theta)}^{\infty} dT\, f(T|H_0(\theta)) = \alpha \quad \rightarrow \quad T_c = \max_{\theta \in H_0} T_c(\theta)$$

if H_1 composite: β will depend on $\theta \in H_1$

$$\int_{-\infty}^{T_c} dT\, f(T|H_1(\theta)) = \beta(\theta)$$

can quote e.g. "best" and "worst" power of the test

Composite hypotheses - comments

- we are completely free to chose any statistic T (power of the test will depend on this choice)
- again, often a LH ratio is a useful test statistic, e.g.

$$T = \frac{\max_{\theta \in H_1} \mathcal{L}(H_1)}{\max_{\theta \in H_0} \mathcal{L}(H_0)}$$

- in sufficiently Gaussian situations the pdf of T is still independent of θ

Application to the neutrino mass ordering

for extensive discussion see Blennow, Coloma, Huber, Schwetz, 1311.1822

test statistic motivated by LH ratio:

$$T = \min_{\theta \in \mathrm{IO}} \chi^2(\theta) - \min_{\theta \in \mathrm{NO}} \chi^2(\theta) \equiv \chi^2_{\mathrm{IO}} - \chi^2_{\mathrm{NO}},$$

under some conditions (similar to Wilk's theorem), T is normal distributed:

$$T = \mathcal{N}(\pm T_0, 2\sqrt{T_0}),$$

with

$$T_0^{\mathrm{NO}}(\theta_0) = \min_{\theta \in \mathrm{IO}} \sum_i \frac{[\mu_i^{\mathrm{NO}}(\theta_0) - \mu_i^{\mathrm{IO}}(\theta)]^2}{\sigma_i^2}$$

Application to the neutrino mass ordering

for extensive discussion see Blennow, Coloma, Huber, Schwetz, 1311.1822

test statistic motivated by LH ratio:

$$T = \min_{\theta \in \mathrm{IO}} \chi^2(\theta) - \min_{\theta \in \mathrm{NO}} \chi^2(\theta) \equiv \chi^2_{\mathrm{IO}} - \chi^2_{\mathrm{NO}},$$

under some conditions (similar to Wilk's theorem), T is normal distributed:

$$T = \mathcal{N}(\pm T_0, 2\sqrt{T_0}),$$

with

$$T_0^{\mathrm{NO}}(\theta_0) = \min_{\theta \in \mathrm{IO}} \sum_i \frac{[\mu_i^{\mathrm{NO}}(\theta_0) - \mu_i^{\mathrm{IO}}(\theta)]^2}{\sigma_i^2}$$

Application to the neutrino mass ordering

for extensive discussion see Blennow, Coloma, Huber, Schwetz, 1311.1822

test statistic motivated by LH ratio:

$$T = \min_{\theta \in \mathrm{IO}} \chi^2(\theta) - \min_{\theta \in \mathrm{NO}} \chi^2(\theta) \equiv \chi^2_{\mathrm{IO}} - \chi^2_{\mathrm{NO}},$$

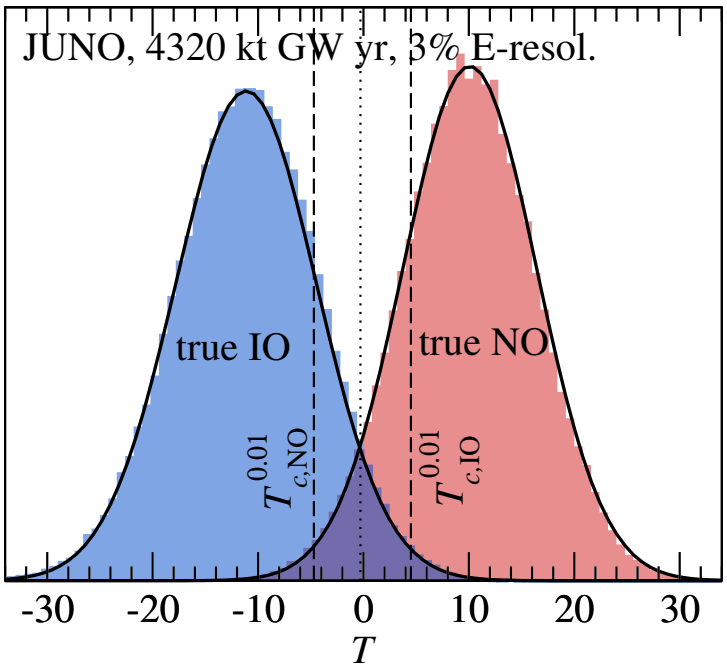
under some conditions (similar to Wilk's theorem), T is normal distributed:

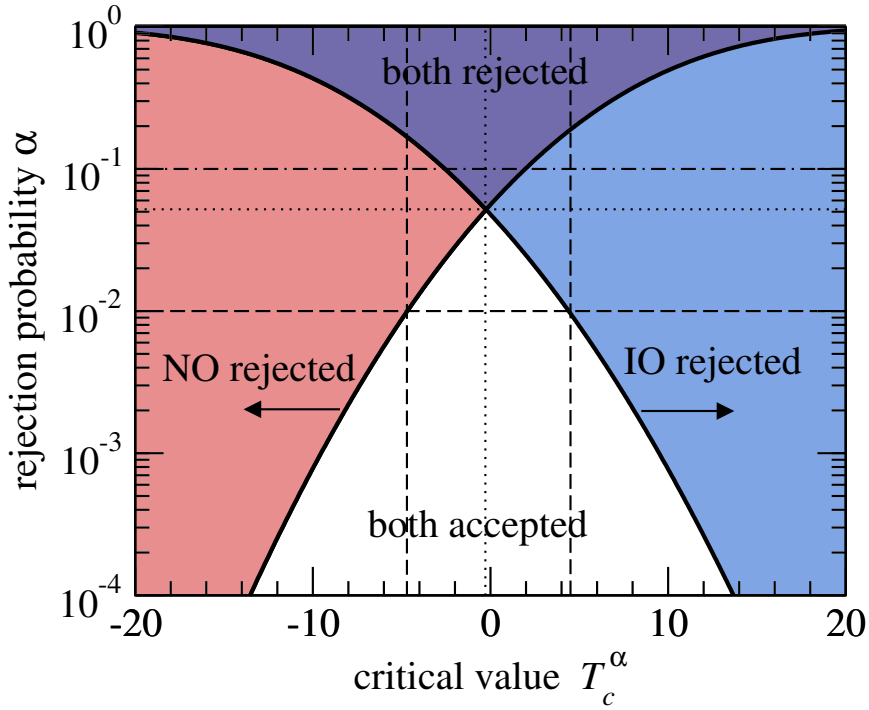
$$T = \mathcal{N}(\pm T_0, 2\sqrt{T_0}),$$

with

$$T_0^{\mathrm{NO}}(\theta_0) = \min_{\theta \in \mathrm{IO}} \sum_i \frac{[\mu_i^{\mathrm{NO}}(\theta_0) - \mu_i^{\mathrm{IO}}(\theta)]^2}{\sigma_i^2}$$

Application to the neutrino mass ordering





Application to the neutrino mass ordering

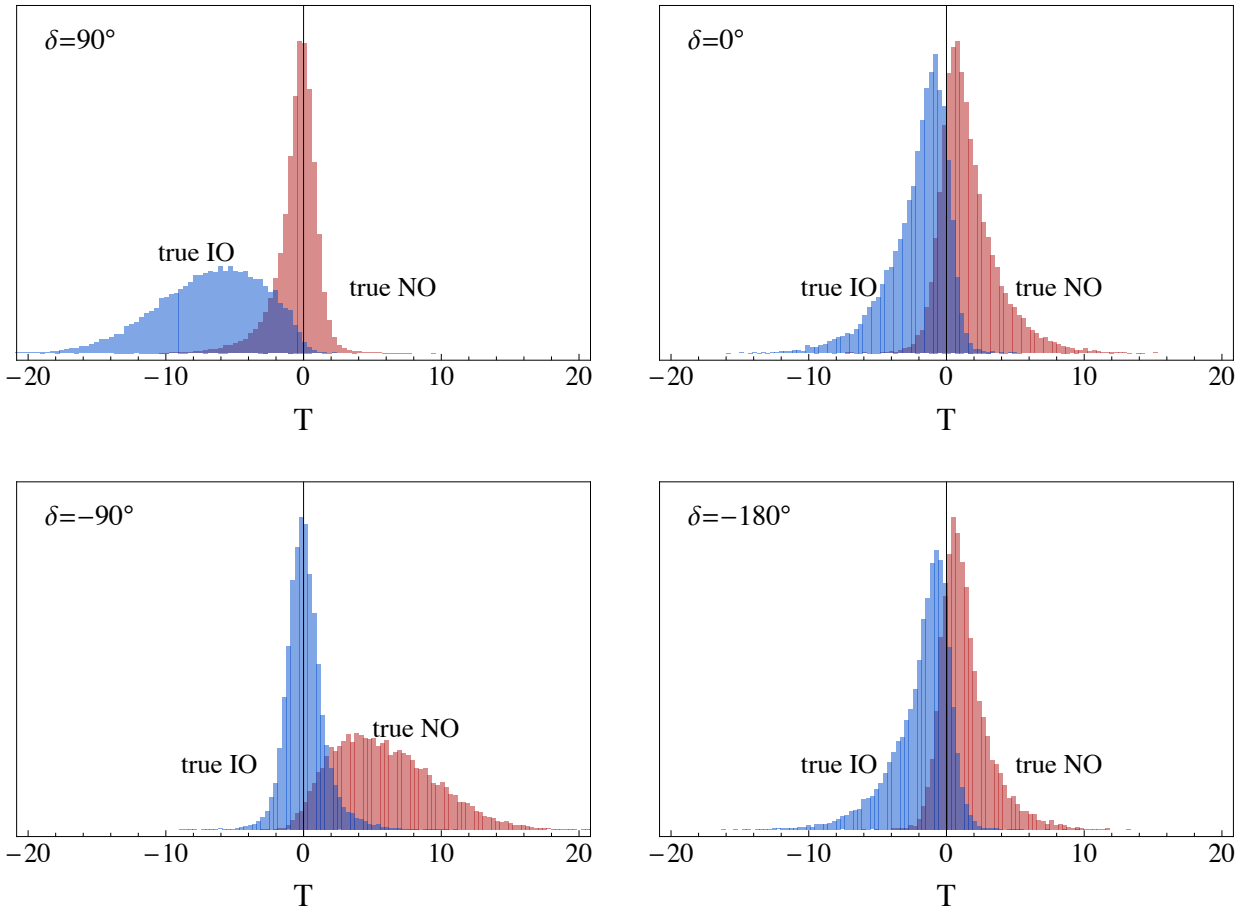


FIG. 7: The simulated distributions of the test statistic T in the NOνA experiment for different true values of δ, as indicated by the labels. The red (blue) distributions assume a true normal (inverted) ordering.

Median experiment

Median sensitivity corresponds to type II error rate of 50% $\Rightarrow$
with 50% chance the actual experiment will obtain a better/worse result

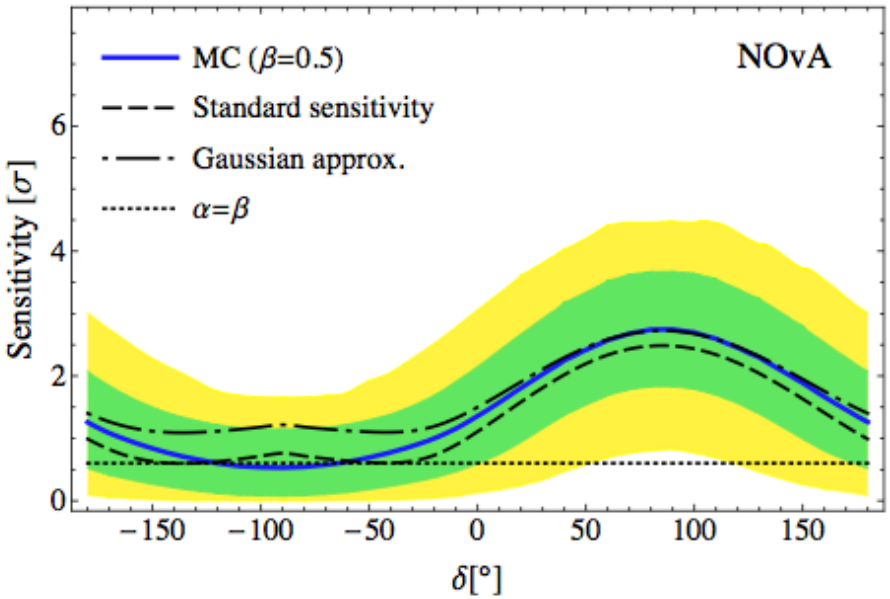
Instead of type I and II errors one can also quote the median sensitivity and its spread (again two numbers)

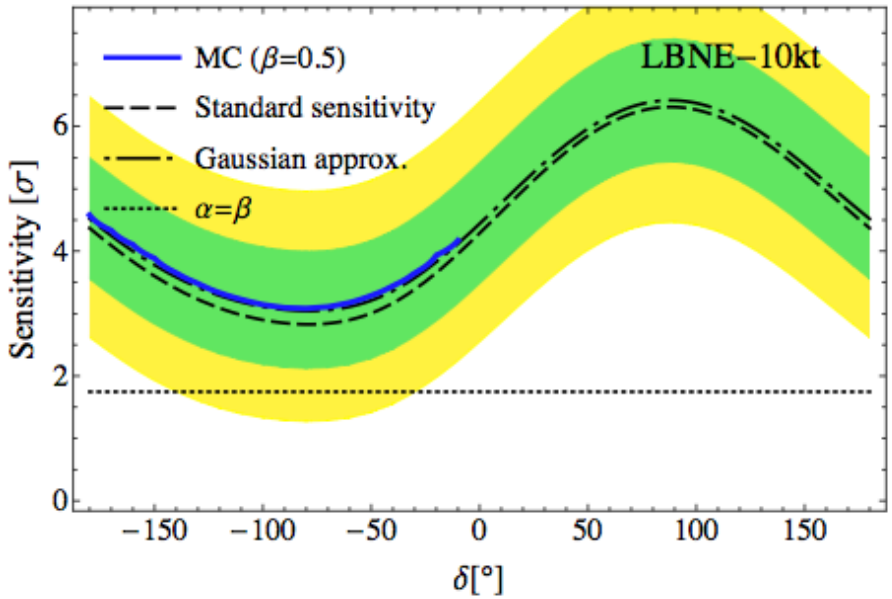
Median experiment

Median sensitivity corresponds to type II error rate of 50% $\Rightarrow$
with 50% chance the actual experiment will obtain a better/worse result

Instead of type I and II errors one can also quote the median sensitivity and its spread (again two numbers)

ex.: mass ordering sensitivity Blennow et al., 1311.1822





Median experiment

Consider the χ^2 using the predicted event rate as "data" (no statistical fluctuation):

$$\chi^2(\boldsymbol{\theta}; \boldsymbol{\theta}^{tr}) = \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - \mu_i(\boldsymbol{\theta}^{tr})]^2}{\mu_i(\boldsymbol{\theta}^{tr})}$$

$n_i = \mu_i(\boldsymbol{\theta}^{tr})$ can be considered as "most probable outcome" or the result of the "median experiment"

- interpret sensitivities based on the above χ^2 as median sensitivity, i.e., type II error rate of 50%.
  holds only approximately, in general needs to be checked by MC
  Schwetz, hep-ph/0612223, Blennow et al., 1311.1822

Median experiment

Consider the χ^2 using the predicted event rate as "data" (no statistical fluctuation):

$$\chi^2(\boldsymbol{\theta}; \boldsymbol{\theta}^{tr}) = \sum_{i=1}^{N} \frac{[\mu_i(\boldsymbol{\theta}) - \mu_i(\boldsymbol{\theta}^{tr})]^2}{\mu_i(\boldsymbol{\theta}^{tr})}$$

$n_i = \mu_i(\boldsymbol{\theta}^{tr})$ can be considered as "most probable outcome" or the result of the "median experiment"

- this is by far the most common method in the literature to calculate sensitivities of neutrino oscillation epxeriments

GLoBES software is designed primarily for this purporse

Huber, Lindner, Winter, hep-ph/0407333; Huber et al., hep-ph/0701187

http://www.mpi-hd.mpg.de/lin/globes/

Nested hypotheses

H_0 and H_1 are related by a continuous parameter, ex.:

- KATRIN: $H_0 : m_\nu = 0$ and $H_1 : m_\nu > 0$
- MO: $H_0 : \Delta m^2_{31} > 0$ (NO) and $H_1 : \Delta m^2_{31} < 0$ (IO)

hypothesis testing becomes related to parameter estimation:

- consider confidence interval for θ and check whether the interval at $(1-\alpha)$ CL covers the value of θ_0 corresponding to the null hypothesis
  $\rightarrow$ probab. of error of first kind $= \alpha$
- error of second kind is given by the probability that the $(1-\alpha)$ CL interval covers θ_0 although the true value $\theta \neq \theta_0$
  (will of course depend on the value of θ)

Nested hypotheses

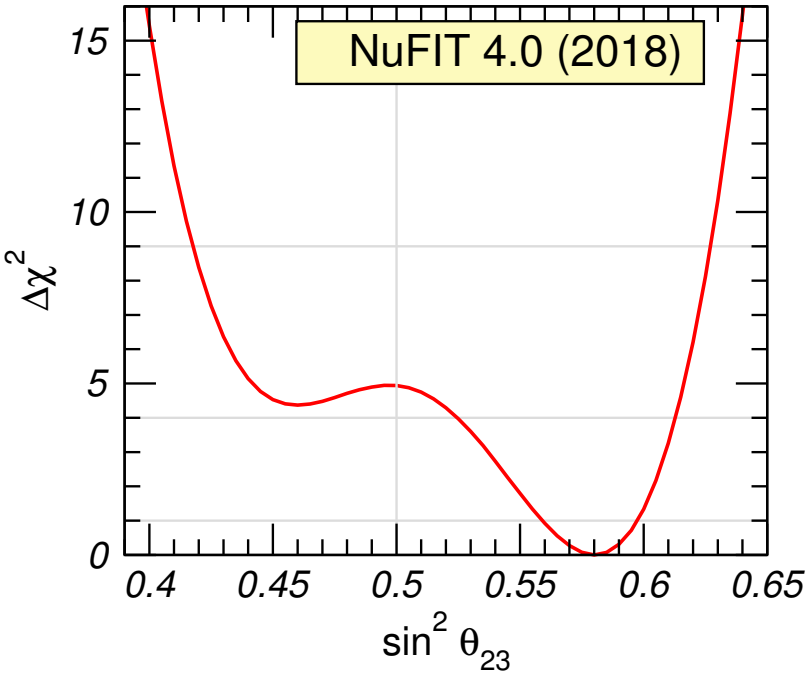
H_0 and H_1 are related by a continuous parameter, ex.:

- KATRIN: $H_0 : m_\nu = 0$ and $H_1 : m_\nu > 0$
- MO: $H_0 : \Delta m^2_{31} > 0$ (NO) and $H_1 : \Delta m^2_{31} < 0$ (IO)

hypothesis testing becomes related to parameter estimation:

- consider confidence interval for θ and check whether the interval at $(1-\alpha)$ CL covers the value of θ_0 corresponding to the null hypothesis $\rightarrow$ probab. of error of first kind $= \alpha$
- error of second kind is given by the probability that the $(1-\alpha)$ CL interval covers θ_0 although the true value $\theta \neq \theta_0$
  (will of course depend on the value of θ)

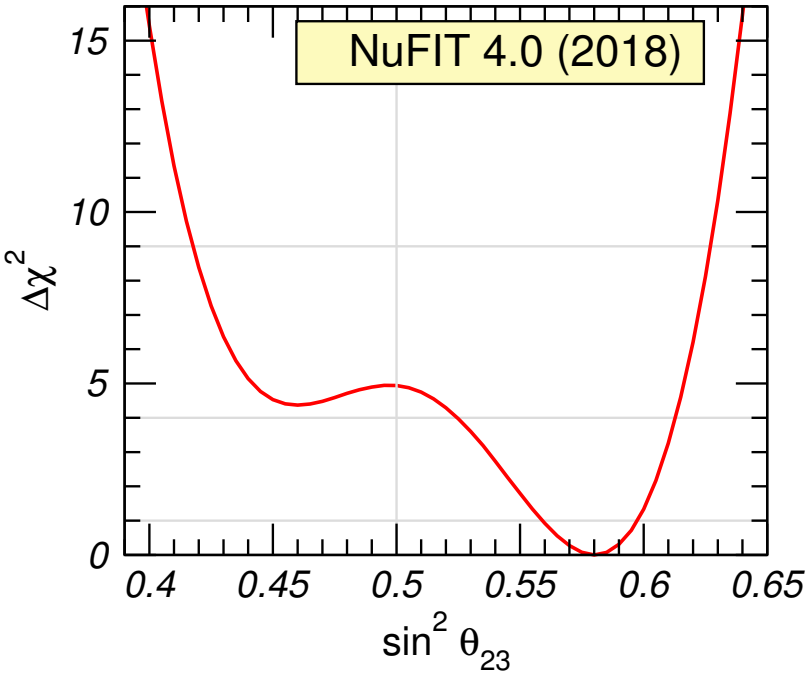
Nested hypotheses, ex.: θ_{23} maximal mixing and octant



- maximal mixing: $H_0 : \theta_{23} = 45°$
  $\Delta\chi^2(H_0) = 4.9 \rightarrow$
  $(\sqrt{4.9} = 2.2)\sigma$ for 1 dof
- first octant: $H_0 : \theta_{23} < 45°$
  $\Delta\chi^2(H_0) = 4.3 \rightarrow$
  $(\sqrt{4.3} = 2.1)\sigma$ for 1 dof

significance went down in NuFit 4.1: 1.4σ

Nested hypotheses, ex.: θ_{23} maximal mixing and octant



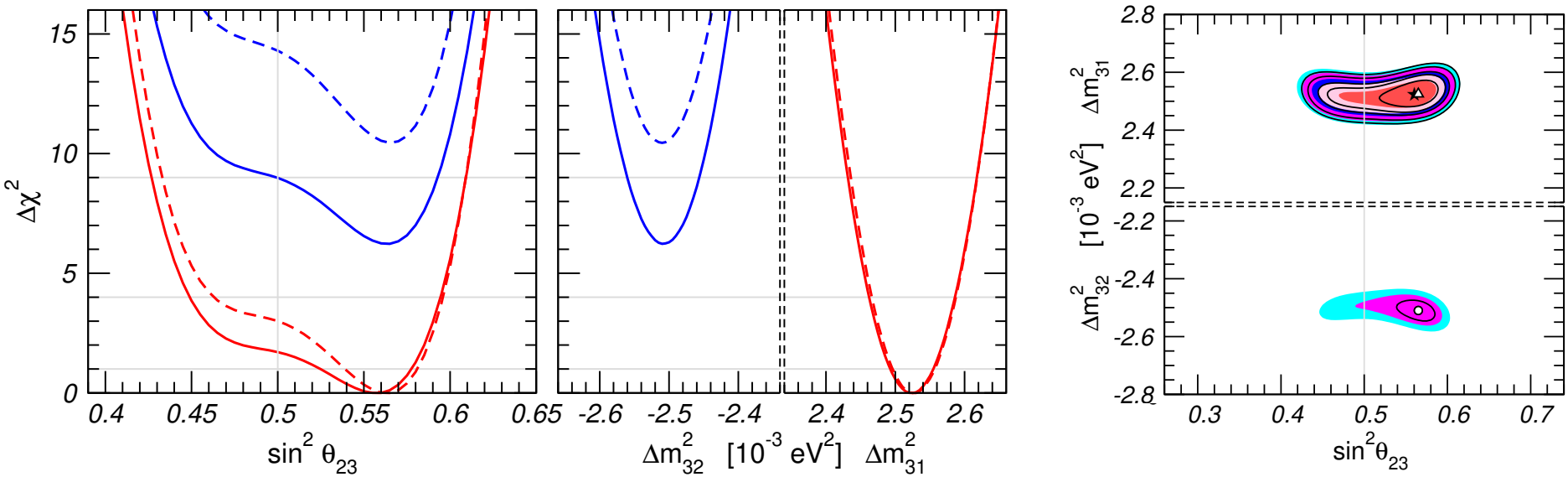
- maximal mixing: $H_0 : \theta_{23} = 45°$
  $\Delta\chi^2(H_0) = 4.9 \rightarrow$
  $(\sqrt{4.9} = 2.2)\sigma$ for 1 dof
- first octant: $H_0 : \theta_{23} < 45°$
  $\Delta\chi^2(H_0) = 4.3 \rightarrow$
  $(\sqrt{4.3} = 2.1)\sigma$ for 1 dof

significance went down in NuFit 4.1: 1.4σ

Mass ordering from real data (global fit)

usually MO significance is interpreted in terms of parameter estimation (nested models)

$$\Delta\chi^2 = \chi^2_{\text{min,IO}} - \chi^2_{\text{min,glob}}$$



NuFit 4.1: $\Delta\chi^2 = 10.4$ (how many dof?)

Bayesian model selection

- In the Bayesian framework we can make statements on the relative belief that H_0 or H_1 is true (usually called "models").
- Calculate "Bayesian odds" of $M_1 : M_2$

suppose we want to compare two hypotheses ("models") M_1, M_2.
each model depends on n_i parameters θ_i
there is a given set of observations ("data") D

Bayesian model selection

- In the Bayesian framework we can make statements on the relative belief that H_0 or H_1 is true (usually called "models").
- Calculate "Bayesian odds" of $M_1 : M_2$

suppose we want to compare two hypotheses ("models") M_1, M_2.
each model depends on n_i parameters $\boldsymbol{\theta}_i$
there is a given set of observations ("data") D

Use Bayes' theorem to calculate probability for model M_i given data:

$$P(M_i|D) = \frac{P(D|M_i)}{P(D)}\pi(M_i) \propto Z_i\pi(M_i)$$

$$P(D|M_i) = Z_i = \int d\boldsymbol{\theta}_i\, f(D|\boldsymbol{\theta}_i, M_i)\pi(\boldsymbol{\theta}_i) \qquad \text{"evidence"}$$

Use Bayes' theorem to calculate probability for model M_i given data:

$$P(M_i|D) = \frac{P(D|M_i)}{P(D)}\pi(M_i) \propto Z_i\pi(M_i)$$

$$P(D|M_i) = Z_i = \int d\boldsymbol{\theta}_i\, f(D|\boldsymbol{\theta}_i, M_i)\pi(\boldsymbol{\theta}_i) \qquad \text{"evidence"}$$

the evidence is the normalization factor in the posterior for the parameters:

$$f(\boldsymbol{\theta}_i|D) = \frac{f(D|\boldsymbol{\theta}_i, M_i)}{Z_i}\pi(\boldsymbol{\theta}_i) \quad \text{...posterior p.d.f. for } \boldsymbol{\theta}_i \text{ given } M_i$$

remember: $f(D|\boldsymbol{\theta}_i, M_i) = \mathcal{L}(\boldsymbol{\theta}_i, M_i)$

Use Bayes' theorem to calculate probability for model M_i given data:

$$P(M_i|D) = \frac{P(D|M_i)}{P(D)}\pi(M_i) \propto Z_i\pi(M_i)$$

$$P(D|M_i) = Z_i = \int d\boldsymbol{\theta}_i\, f(D|\boldsymbol{\theta}_i, M_i)\pi(\boldsymbol{\theta}_i) \qquad \text{"evidence"}$$

relative odds for M_1 versus M_2 after data:

$$M_1 : M_2 = \frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{\pi(M_1)}{\pi(M_2)}$$

The "Bayes factor" determines how much the data changes our degree of belief in model 1 versus model 2:

$$B = \frac{Z_1}{Z_2}$$

Jeffrey scale

$\lvert\log(\text{odds})\rvert$	odds	$\Pr(M_1\lvert\mathbf{D})$	Strength of evidence
< 1.0	$\lesssim 3:1$	$\lesssim 0.75$	Inconclusive
1.0	$\simeq 3:1$	$\simeq 0.75$	Weak evidence
2.5	$\simeq 12:1$	$\simeq 0.92$	Moderate evidence
5.0	$\simeq 150:1$	$\simeq 0.993$	Strong evidence

Table 1. *The Jeffreys scale, used for interpretation of Bayes factors, odds, and model probabilities. The posterior model probabilities for the preferred model are calculated assuming only two competing hypotheses and equal prior probabilities. Note that* log *denotes the natural logarithm.*

$$\text{odds} = B \text{ for } \pi(M_1) = \pi(M_2) = 0.5$$

Bayesian evidence

The evidence describes the overlap of the prior and the likelihood

$$Z_i = \int d\boldsymbol{\theta}_i \, \mathcal{L}(\boldsymbol{\theta}_i, M_i)\pi(\boldsymbol{\theta}_i)$$

- models with large overlap of prior and likelihood are favoured
- models with many parameters are penalized (volume factor)

example: nested models

- $M_1 : \theta$ free parameter with prior $\pi(\theta)$
- $M_0 : \theta = \theta_0 \rightarrow \pi(\theta) = \delta(\theta - \theta_0)$

$$Z_1 = \int d\theta \, \mathcal{L}(\theta)\pi(\theta)\,, \qquad Z_0 = \mathcal{L}(\theta_0)$$

Bayesian evidence

The evidence describes the overlap of the prior and the likelihood

$$Z_i = \int d\boldsymbol{\theta}_i \, \mathcal{L}(\boldsymbol{\theta}_i, M_i)\pi(\boldsymbol{\theta}_i)$$

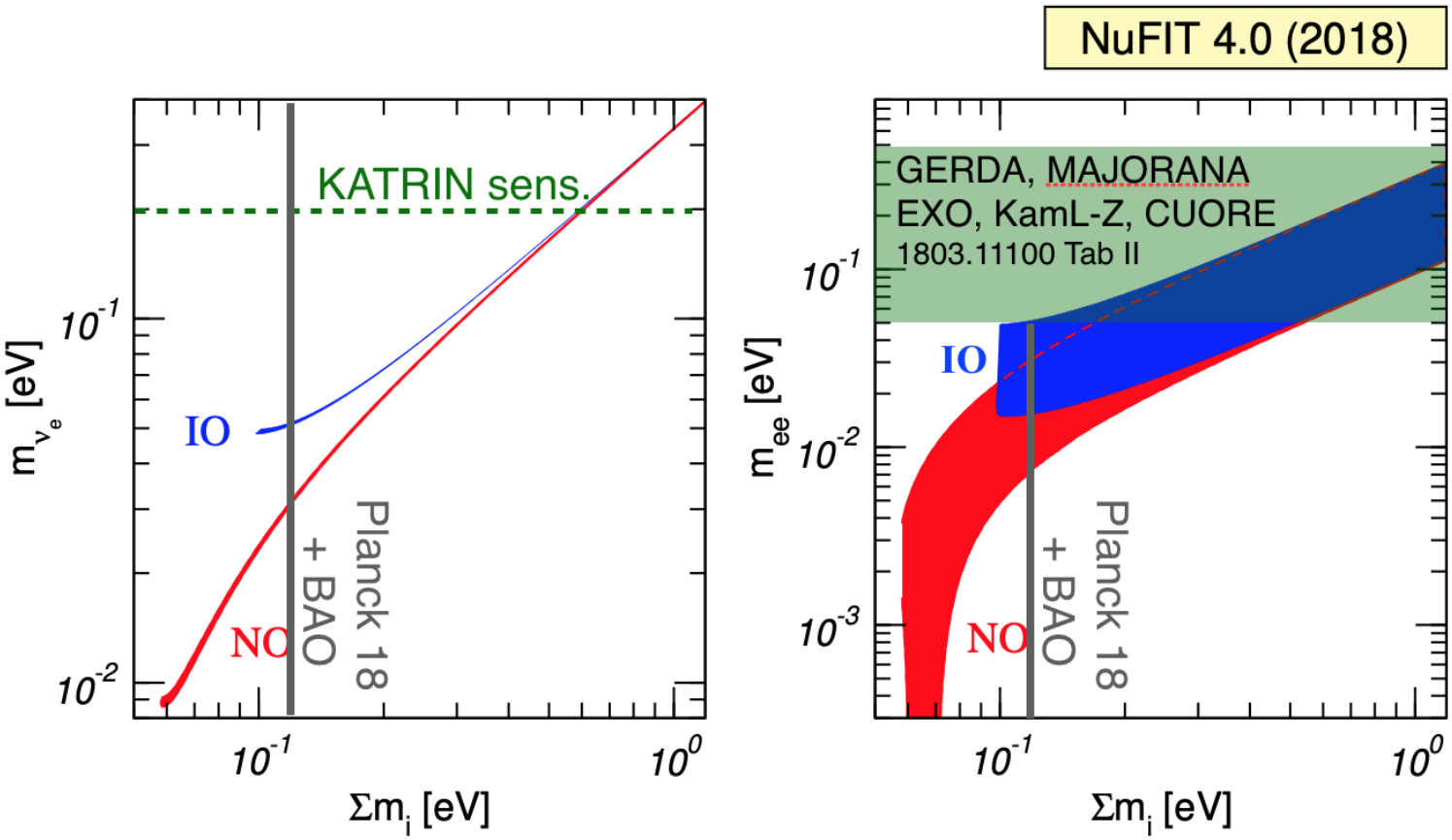
- models with large overlap of prior and likelihood are favoured
- models with many parameters are penalized (volume factor)

example: nested models

- $M_1 : \theta$ free parameter with prior $\pi(\theta)$
- $M_0 : \theta = \theta_0 \rightarrow \pi(\theta) = \delta(\theta - \theta_0)$

$$Z_1 = \int d\theta \, \mathcal{L}(\theta)\pi(\theta)\,, \qquad Z_0 = \mathcal{L}(\theta_0)$$

Neutrino mass ordering from cosmology



Neutrino mass ordering from cosmology

$$\sum \equiv \sum_{i=1}^{3} m_i$$

$$= \begin{cases} m_0 + \sqrt{\Delta m_{21}^2 + m_0^2} + \sqrt{\Delta m_{31}^2 + m_0^2} & \text{(NO)} \\ m_0 + \sqrt{|\Delta m_{32}^2| + m_0^2} + \sqrt{|\Delta m_{32}^2| - \Delta m_{21}^2 + m_0^2} & \text{(IO)} \end{cases}$$

minimal values for $m_0 = 0$:

$$\sum \geq \begin{cases} \sqrt{\Delta m_{21}^2} + \sqrt{\Delta m_{31}^2} & = 58.5 \pm 0.48\,\text{meV} \quad \text{(NO)} \\ \sqrt{|\Delta m_{32}^2|} + \sqrt{|\Delta m_{32}^2| - \Delta m_{21}^2} & = 98.6 \pm 0.85\,\text{meV} \quad \text{(IO)} \end{cases}$$

Neutrino mass ordering from cosmology

$$\sum \equiv \sum_{i=1}^{3} m_i$$

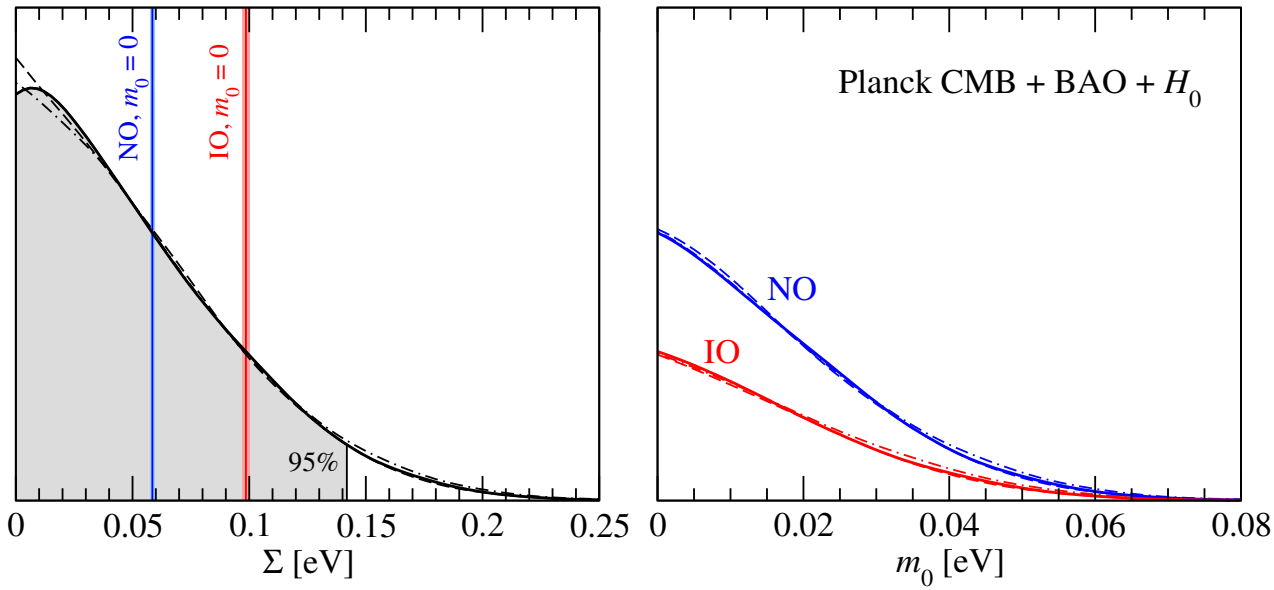
$$= \begin{cases} m_0 + \sqrt{\Delta m_{21}^2 + m_0^2} + \sqrt{\Delta m_{31}^2 + m_0^2} & \text{(NO)} \\ m_0 + \sqrt{|\Delta m_{32}^2| + m_0^2} + \sqrt{|\Delta m_{32}^2| - \Delta m_{21}^2 + m_0^2} & \text{(IO)} \end{cases}$$

minimal values for $m_0 = 0$:

$$\sum \geq \begin{cases} \sqrt{\Delta m_{21}^2} + \sqrt{\Delta m_{31}^2} & = 58.5 \pm 0.48\,\text{meV} \quad \text{(NO)} \\ \sqrt{|\Delta m_{32}^2|} + \sqrt{|\Delta m_{32}^2| - \Delta m_{21}^2} & = 98.6 \pm 0.85\,\text{meV} \quad \text{(IO)} \end{cases}$$

Neutrino mass ordering from cosmology

Hannestad, Schwetz, 1606.04691

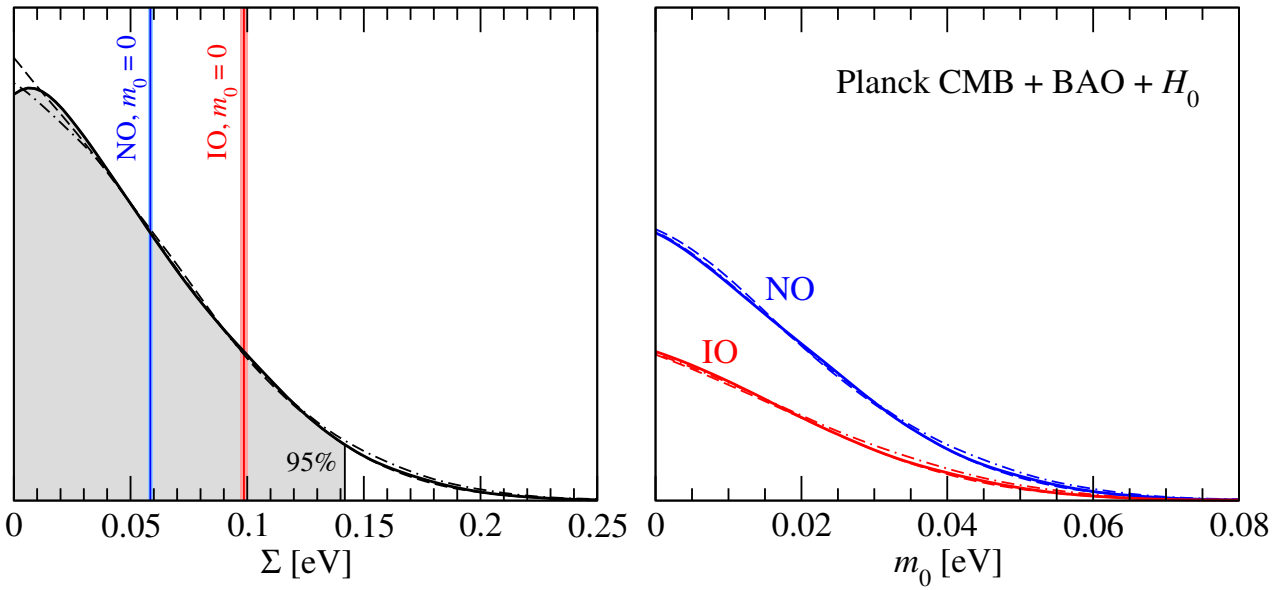


relative odds (assuming flat prior for m_0):

$$IO : NO = \frac{Z_{IO}}{Z_{NO}} \frac{\pi(IO)}{\pi(NO)} = \frac{\int_0^\infty dm_0 \, \mathcal{L}(D|m_0, IO)}{\int_0^\infty dm_0 \, \mathcal{L}(D|m_0, NO)} \frac{\pi(IO)}{\pi(NO)} \approx \frac{1}{2} \frac{\pi(IO)}{\pi(NO)}$$

Neutrino mass ordering from cosmology

Hannestad, Schwetz, 1606.04691

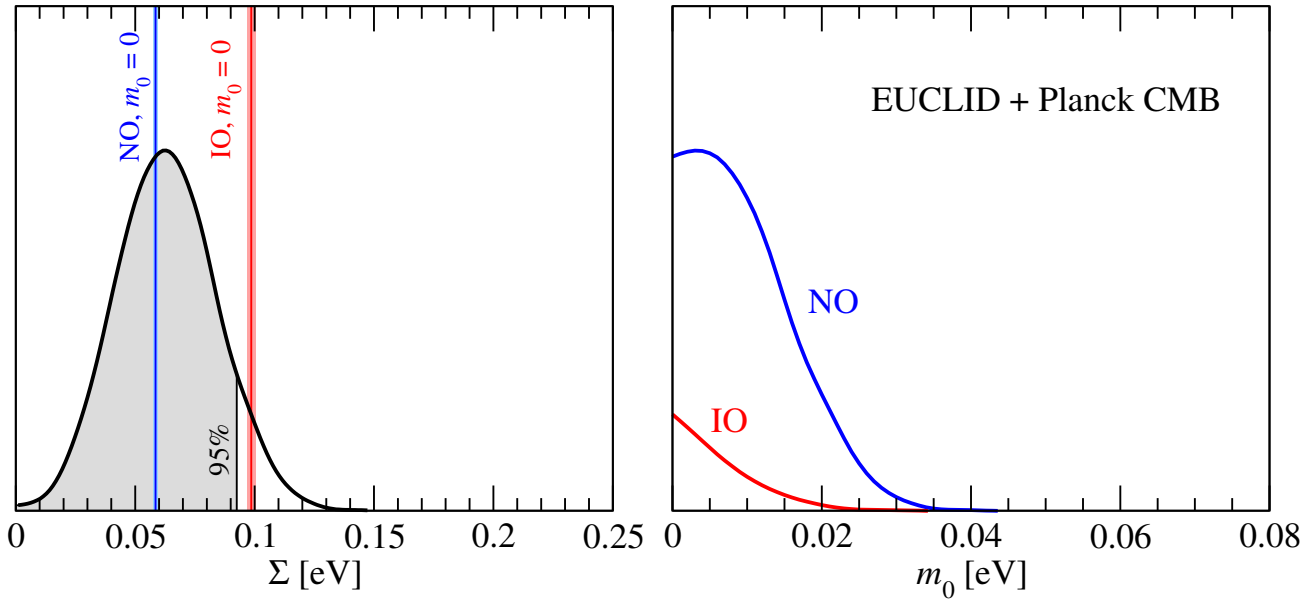


relative odds (assuming flat prior for m_0):

$$IO : NO = \frac{Z_{IO}}{Z_{NO}} \frac{\pi(IO)}{\pi(NO)} = \frac{\int_0^\infty dm_0 \, \mathcal{L}(D|m_0, IO)}{\int_0^\infty dm_0 \, \mathcal{L}(D|m_0, NO)} \frac{\pi(IO)}{\pi(NO)} \approx \frac{1}{2} \frac{\pi(IO)}{\pi(NO)}$$

Neutrino mass ordering from cosmology

Hannestad, Schwetz, 1606.04691



relative odds (assuming flat prior for m_0):

$$IO : NO = \frac{Z_{IO}}{Z_{NO}} \frac{\pi(IO)}{\pi(NO)} = \frac{\int_0^\infty dm_0 \, \mathcal{L}(D|m_0, IO)}{\int_0^\infty dm_0 \, \mathcal{L}(D|m_0, NO)} \frac{\pi(IO)}{\pi(NO)} \approx \frac{1}{12} \frac{\pi(IO)}{\pi(NO)}$$

Final word on priors

Astrophysics > Cosmology and Nongalactic Astrophysics

Cosmology and the neutrino mass ordering

Steen Hannestad, Thomas Schwetz

(Submitted on 15 Jun 2016 (v1), last revised 18 Nov 2016 (this version, v2))

We propose a simple method to quantify a possible exclusion of the inverted neutrino mass ordering from cosmological bounds on the sum of the neutrino masses. The method is based on Bayesian inference and allows for a calculation of the posterior odds of normal versus inverted ordering. We apply the method for a specific set of current data from Planck CMB data and large-scale structure surveys, providing an upper bound on the sum of neutrino masses of 0.14 eV at 95% CL. With this analysis we obtain posterior odds for normal versus inverted ordering of about 2:1. If cosmological data is combined with data from oscillation experiments the odds reduce to about 3:2. For an exclusion of the inverted ordering from cosmology at more than 95% CL, an

Astrophysics > Cosmology and Nongalactic Astrophysics

Strong Evidence for the Normal Neutrino Hierarchy

Fergus Simpson, Raul Jimenez, Carlos Pena-Garay, Licia Verde

(Submitted on 9 Mar 2017)

The configuration of the three neutrino masses can take two forms, known as the normal and inverted hierarchies. We compute the Bayesian evidence associated with these two hierarchies. Previous studies found a mild preference for the normal hierarchy, and this was driven by the asymmetric manner in which cosmological data has confined the available parameter space. Here we identify the presence of a second asymmetry, which is imposed by data from neutrino oscillations. By combining constraints on the squared-mass splittings with the limit on the sum of neutrino masses of $\Sigma m_\nu < 0.13$ eV, we infer odds of 42:1 in favour of the normal hierarchy, which is classified as "strong" in the Jeffreys' scale. We explore how these odds may evolve in light of higher precision cosmological data, and discuss the implications of this finding with regards to the nature of neutrinos.

de.arXiv.org > astro-ph > arXiv:1606.04691

Astrophysics > Cosmology and Nongalactic Astrophysics

Cosmology and the neutrino mass ordering

Steen Hannestad, Thomas Schwetz

(Submitted on 15 Jun 2016 (v1), last revised 18 Nov 2016 (this version, v2))

We propose a simple method to quantify a possible exclusion of the inverted neutrino mass ordering from cosmological bounds on the sum of the neutrino masses. The method is based on Bayesian inference and allows for a calculation of the posterior odds of normal versus inverted ordering. We apply the method for a specific set of current data from Planck CMB data and large-scale structure surveys, providing an upper bound on the sum of neutrino masses of 0.14 eV at 95% CL. With this analysis we obtain posterior odds for normal versus inverted ordering of about 2:1. If cosmological data is combined with data from oscillation experiments the odds reduce to about 3:2. For an exclusion of the inverted ordering from cosmology at more than 95% CL, an

de.arXiv.org > astro-ph > arXiv:1703.03425

Astrophysics > Cosmology and Nongalactic Astrophysics

Strong Evidence for the Normal Neutrino Hierarchy

Fergus Simpson, Raul Jimenez, Carlos Pena-Garay, Licia Verde

(Submitted on 9 Mar 2017)

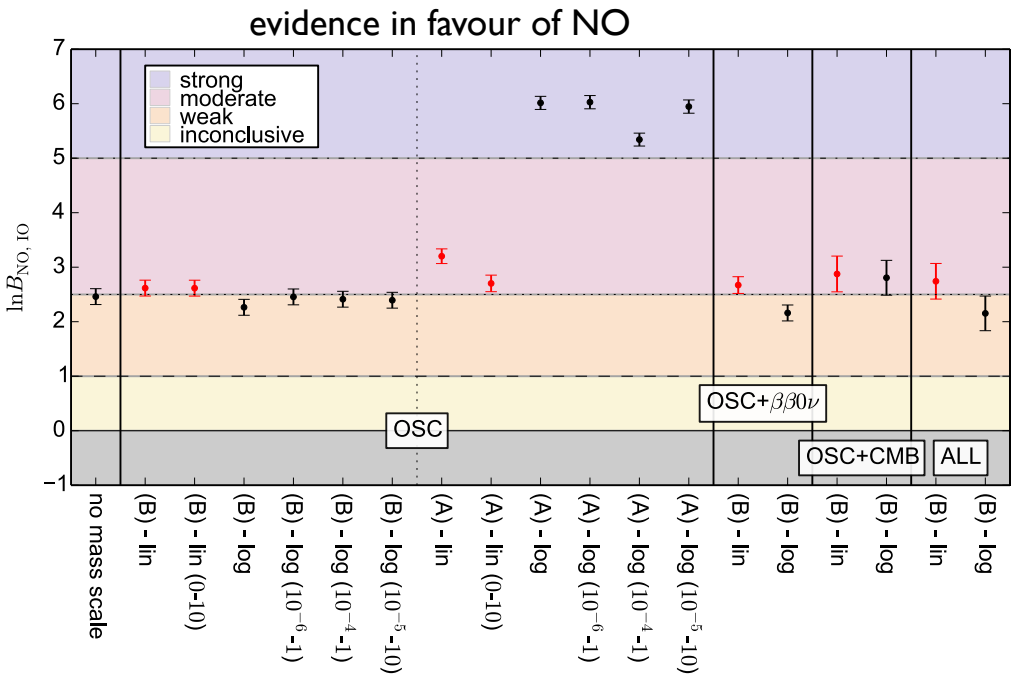
The configuration of the three neutrino masses can take two forms, known as the normal and inverted hierarchies. We compute the Bayesian evidence associated with these two hierarchies. Previous studies found a mild preference for the normal hierarchy, and this was driven by the asymmetric manner in which cosmological data has confined the available parameter space. Here we identify the presence of a second asymmetry, which is imposed by data from neutrino oscillations. By combining constraints on the squared-mass splittings with the limit on the sum of neutrino masses of $\Sigma m_\nu < 0.13$ eV, we infer odds of 42:1 in favour of the normal hierarchy, which is classified as "strong" in the Jeffreys' scale. We explore how these odds may evolve in light of higher precision cosmological data, and discuss the implications of this finding with regards to the nature of neutrinos.

Watch out for assumptions about priors!!

see comment in arXiv:1703.04585

Model A			Model B		
Parameter	Prior	Range	Parameter	Prior	Range
m_1/eV	linear log	$0-1$ $10^{-5}-1$	$m_{\text{lightest}}/\text{eV}$	linear log	$0-1$ $10^{-5}-1$
m_2/eV	linear log	$0-1$ $10^{-5}-1$	$\Delta m^2_{21}/\text{eV}^2$	linear	$5\times10^{-5}-10^{-4}$
m_3/eV	linear log	$0-1$ $10^{-5}-1$	$\lvert\Delta m^2_{31}\rvert/\text{eV}^2$	linear	$1.5\times10^{-3}-3.5\times10^{-3}$

Archidiacono, de Salas, Gariazzo, Mena, Ternes, Tortola, 1801.04946



- assuming a log prior in the 3 masses prefers strongly NO (just from oscillation data!)