Statistics

or “How to find answers to your questions”

Pietro Vischia[1]

[1] CP3 — IRMP, Université catholique de Louvain

UCLouvain
Institut de recherche
en mathématique et physique

CP3—IRMP, Intensive Course on Statistics for HEP, 21/01–18/02 2022

Program for today

Lesson 2
Estimating a physical quantity

- Sufficiency principle
- Likelihood Principle
- Estimators and maximum likelihood
- Profile likelihood ratio

UCLouvain
Institut de recherche
en mathématique et physique

- Schedule: five days of lectures (Every Friday for the next five weeks)
  - 2h morning lecture, virtual coffee break midway (09:30–11:45)
  - 2h (probably less) afternoon exercise session, virtual coffee break midway (13:30–15:45)
- Many interesting references, nice reading list for your career
  - Papers mostly cited in the topical slides
  - Some cool books cited here and there and in the appendix
- Unless stated otherwise, figures belong to P. Vischia for inclusion in my upcoming textbook on Statistics for HEP
  (textbook to be published by Springer in 2021)
  - Or I forgot to put the reference, let me know if you spot any figure obviously lacking reference, so that I can fix it
  - I cannot put the recordings publicly online as "massive online course", so I will distribute them only to registered participants, and have to ask you to not record yourself. I hope you understand.
- Your feedback is crucial for improving these lectures (a feedback form will be provided at the end of the lectures)!
  - You can also send me an email during the lectures: if it is something I can fix for the next day, I'll gladly do so!

The Bureaucracy

- This course provides 3 credits for the UCLouvain doctoral school (CDD Sciences)
  - If you need it recognized by another doctoral school, you have to ask to your school
  - Besides the certificate, I am available at supplying additional information (e.g. detailed schedule) or activity (exam?)
- People connecting online: certificates will be provided by checking connection logs
  - The only way I have to check if you connected to most lectures is to check the Zoom logs
  - Make sure you connect with a recognizable email address (or let me know which unrecognizable address belongs to you)
- This course contributes to the activities of the Excellence of Science (EOS) Be.h network, https://be-h.be/

- I will pop up every now and then some questions
- I will open a link, and you'll be able to answer by going to www.menti.com and inserting a code
- Totally anonymous (no access even for me to any ID information, not even the country): don't be afraid to give a wrong answer!
  - The purpose is making you think, not having 100% correct answers!
- First question of the day is purely a logistics matter
  Question time: ROOT
  - The direct links are accessible to me only: you'll see in your screens the code in a second :)
- The slides of each lecture will be available one minute after the end of the lecture
  - To encourage you to really try answering without looking at the answers

UCLouvain
Institut de recherche
en mathématique et physique

- **Lesson 1 - Fundaments**
  - Bayesian and frequentist probability, theory of measure, correlation and causality, distributions
- **Lesson 2 - Point and Interval estimation**
  - Maximum likelihood methods, confidence intervals, most probable values, credible intervals
- **Lesson 3 - Advanced interval estimation, test of hypotheses**
  - Interval estimation near the physical boundary of a parameter
  - Frequentist and Bayesian tests, CLs, significance, look-elsewhere effect, reproducibility crysis
- **Lesson 4 - Commonly-used methods in particle physics**
  - Unfolding, ABCD, ABC, MCMC, estimating efficiencies
- **Lesson 5 - Machine Learning**
  - Overview and mathematical foundations, generalities most used algorithms, automatic Differentiation and Deep Learning

Lesson 2
Point and Interval estimation

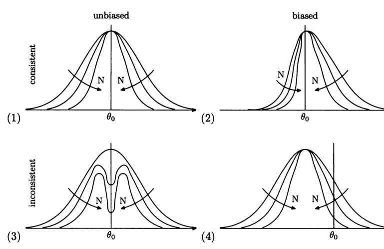
Estimating a physical quantity

Estimators

- Set $\vec{x} = (x_1, ..., x_N)$ of N statistically independent observations x_i, sampled from a p.d.f. $f(x)$.
- Mean and width of $f(x)$ (or some parameter of it: $f(x;\vec{\theta})$, with $\vec{\theta} = (\theta_1, ..., \theta_M)$ unknown)
  - In case of a linear p.d.f., the vector of parameters would be $\vec{\theta} = (\mathit{intercept}, \mathit{slope})$
- We call estimator a function of the observed data $\vec{x}$ which returns numerical values $\hat{\vec{\theta}}$ for the vector $\vec{\theta}$.
- $\hat{\vec{\theta}}$ is (asymptotically) consistent if it converges to $\vec{\theta}_{true}$ for large N:

$$\lim_{N\to\infty} \hat{\vec{\theta}} = \vec{\theta}_{true}$$

- $\hat{\vec{\theta}}$ is unbiased if its bias is zero, $\vec{b} = 0$
  - Bias of $\hat{\vec{\theta}}$: $\vec{b} := E[\hat{\vec{\theta}}] - \vec{\theta}_{true}$
  - If bias is known, can redefine $\hat{\vec{\theta}}' = \hat{\vec{\theta}} - \vec{b}$, resulting in $\vec{b}' = 0$.
- $\hat{\vec{\theta}}$ is efficient if its variance $V[\hat{\vec{\theta}}]$ is the smallest possible

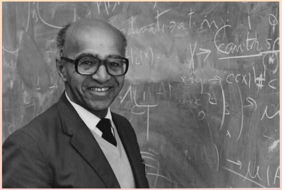


Plot from James, 2nd ed.

- An estimator is robust when it is insensitive to small deviations from the underlying distribution (p.d.f.) assumed (ideally, one would want distribution-free estimates, without assumptions on the underlying p.d.f.)

Sufficient statistic

- A test statistic is a function of the data (a quantity derived from the data sample)
- When $X \sim f(X|\theta)$, a statistic $T = T(\boldsymbol{X})$ is sufficient for θ if the density function $f(\boldsymbol{X}|T)$ is independent of θ
  - If T is a sufficient statistic for θ, then also any strictly monotonic $g(T)$ is sufficient for θ
- Minimal sufficient statistic: a sufficient statistic that is a function of all other sufficient statistics for θ
- The statistic T carries as much information about θ as the original data $\boldsymbol{X}$
  - No other function can give any further information about θ
  - Same inference from data $\boldsymbol{X}$ with model M and from sufficient statistic $T(\boldsymbol{X})$ with model M'
- **Rao–Blackwell theorem:** if $g(X)$ is an estimator for θ and T is a sufficient statistic, then the conditional expectation of $g(X)$ given $T(X)$ is never a worse estimator of θ
  - Practical procedure: build a ballpark estimator $g(X)$, then condition it on a $T(X)$ to obtain a better estimator
- **The Sufficiency Principle:** Two observations X and Y that factorize through the same value of $T(\cdot)$, i.e. s.t. $T(x) = T(y)$, must lead to the same inference about θ



Images from AmStat magazine and from Illinois.edu

Example of sufficient statistic

- Given some data $1, 2, 3, 4, 5$, you may want to estimate the population mean
  - Consider the sample mean $\hat{x} = \frac{1+2+3+4+5}{5} = 3$ as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic? Question time: Sufficient statistic

Example of sufficient statistic

- Given some data $1, 2, 3, 4, 5$, you may want to estimate the population mean
  - Consider the sample mean $\hat{x} = \frac{1+2+3+4+5}{5} = 3$ as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic? Question time: Sufficient statistic
  - If you only knew the sample mean of 3, you would estimate the population mean to be 3 anyway, regardless of having the data or not
  - Knowing the data (the set $1, 2, 3, 4, 5$) or knowing only the sample mean does not improve our estimate for the population mean

Example of sufficient statistic

- Given some data $1, 2, 3, 4, 5$, you may want to estimate the population mean
  - Consider the sample mean $\hat{x} = \frac{1+2+3+4+5}{5} = 3$ as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic? Question time: Sufficient statistic
  - If you only knew the sample mean of 3, you would estimate the population mean to be 3 anyway, regardless of having the data or not
  - Knowing the data (the set $1, 2, 3, 4, 5$) or knowing only the sample mean does not improve our estimate for the population mean
- Estimate the binomial probability of obtaining r heads in N coin tosses
  - Record heads and tails, with their order: $HTTHHHTHHTTTHTHTH$
  - Can we somehow improve by identifying a sufficient statistic? Question time: Sufficient Statistic

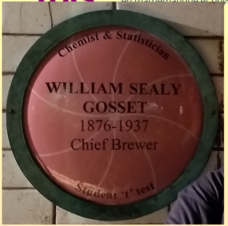
Example of sufficient statistic

- Given some data $1, 2, 3, 4, 5$, you may want to estimate the population mean
  - Consider the sample mean $\hat{x} = \frac{1+2+3+4+5}{5} = 3$ as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic? Question time: Sufficient statistic
  - If you only knew the sample mean of 3, you would estimate the population mean to be 3 anyway, regardless of having the data or not
  - Knowing the data (the set $1, 2, 3, 4, 5$) or knowing only the sample mean does not improve our estimate for the population mean
- Estimate the binomial probability of obtaining r heads in N coin tosses
  - Record heads and tails, with their order: $HTTHHHTHHTTTHTHTH$
  - Can we somehow improve by identifying a sufficient statistic? Question time: Sufficient Statistic
  - What happens if we record only the number of heads? (remember that the binomial p.d.f. is: $P(r) = \binom{N}{r} p^r (1-p)^{N-r}, r = 0, 1, \ldots, N$)

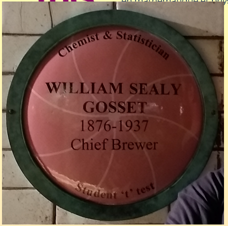
UCLouvain
Institut de recherche
en mathématique et physique

- Given some data $1, 2, 3, 4, 5$, you may want to estimate the population mean
  - Consider the sample mean $\hat{x} = \frac{1+2+3+4+5}{5} = 3$ as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic? Question time: Sufficient statistic
  - If you only knew the sample mean of 3, you would estimate the population mean to be 3 anyway, regardless of having the data or not
  - Knowing the data (the set $1, 2, 3, 4, 5$) or knowing only the sample mean does not improve our estimate for the population mean
- Estimate the binomial probability of obtaining r heads in N coin tosses
  - Record heads and tails, with their order: $HTTHHHTHHTTTHTHTH$
  - Can we somehow improve by identifying a sufficient statistic? Question time: Sufficient Statistic
  - What happens if we record only the number of heads? (remember that the binomial p.d.f. is: $P(r) = \binom{N}{r} p^r (1-p)^{N-r}, r = 0, 1, \dots, N$)
  - Recording only the number of heads (no tails, no order) gives exactly the same information
  - Data can be reduced; we only need to store a sufficient statistic (the distribution $f(\boldsymbol{X}|T)$ is independent of θ)
  - **Storage needs are reduced!!!**

Ancillary statistic and pivotal quantities

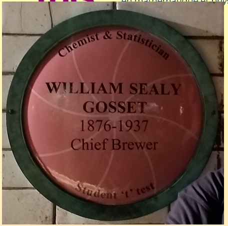


- Pivotal quantity: its distribution does not depend on the parameters
  - For a $Gaus(\mu, \sigma^2)$ p.d.f., $\frac{\bar{X}-\mu}{S/\sqrt{N}} \sim t_{student}$ is a pivot
  - See exercise this afternoon
- Ancillary statistic for a parameter θ: a statistic $f(X)$ which does not depend on θ
  - Concept linked to that of *(minimal) sufficient statistic*; (maximal) data reduction while retaining all Fisher information about θ
- Can an ancillary statistic can give information about θ even if it does not depend on it? QT! Ancillary



- Pivotal quantity: its distribution does not depend on the parameters
  - For a $Gaus(\mu, \sigma^2)$ p.d.f., $\frac{\bar{X}-\mu}{S/\sqrt{N}} \sim t_{student}$ is a pivot
  - See exercise this afternoon
- Ancillary statistic for a parameter θ: a statistic $f(X)$ which does not depend on θ
  - Concept linked to that of *(minimal) sufficient statistic*; (maximal) data reduction while retaining all Fisher information about θ
- Can an ancillary statistic can give information about θ even if it does not depend on it? QT! Ancillary
- Yes!
  - Sample X_1 and X_2 from $P_\theta(X=\theta) = P_\theta(X=\theta+1) = P_\theta(X=\theta+2) = \frac{1}{3}$
  - Ancillary statistic: $R := X_2 - X_1$ (no information about θ)
  - Minimal sufficient statistic: $M := \frac{X_1+X_2}{2}$
  - Sample point $(M=m, R=r)$: either $\theta = m$, or $\theta = m-1$, or $\theta = m-2$
  - If $R=2$, then necessarily $X_1 = m-1$ and $X_2 = m-2$; Therefore necessarily $\theta = m-1$

Ancillary statistic and pivotal quantities



- Pivotal quantity: its distribution does not depend on the parameters
  - For a $Gaus(\mu, \sigma^2)$ p.d.f., $\frac{\bar{X}-\mu}{S/\sqrt{N}} \sim t_{student}$ is a pivot
  - See exercise this afternoon
- Ancillary statistic for a parameter θ: a statistic $f(X)$ which does not depend on θ
  - Concept linked to that of *(minimal) sufficient statistic*; (maximal) data reduction while retaining all Fisher information about θ
- Can an ancillary statistic can give information about θ even if it does not depend on it? QT! Ancillary
- Yes!
  - Sample X_1 and X_2 from $P_\theta(X = \theta) = P_\theta(X = \theta + 1) = P_\theta(X = \theta + 2) = \frac{1}{3}$
  - Ancillary statistic: $R := X_2 - X_1$ (no information about θ)
  - Minimal sufficient statistic: $M := \frac{X_1+X_2}{2}$
  - Sample point $(M = m, R = r)$: either $\theta = m$, or $\theta = m - 1$, or $\theta = m - 2$
  - If $R = 2$, then necessarily $X_1 = m - 1$ and $X_2 = m - 2$; Therefore necessarily $\theta = m - 1$
- Knowledge of R alone carries no information on θ, but increases the precision on an estimate of θ (Cox, Efron, Hinckley)!
- Powerful tool to improve data reduction capabilities (save money...)
- Also employed for asymptotic likelihood expressions
  - Also impact on approximate expressions for significance

- The information of a set of observations should increase with the number of observations
  - Double the data should result in double the information if the data are independent
- Information should be conditional on what we want to learn from the experiment
  - Data which are irrelevant to our hypothesis should carry zero information relative to our hypothesis
- Information should be related to precision
  - The greatest the information carried by the data, the better the precision of our result

UCLouvain
Institut de recherche
en mathématique et physique

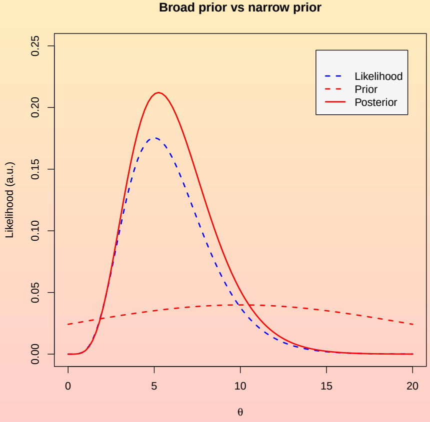
- Common enunciation: given a set of observed data $\vec{x}$, the likelihood function $L(\vec{x};\theta)$ contains all the information that is relevant to the estimation of the parameter θ contained in the data sample
  - The likelihood function is seen as a function of θ, for a fixed set (a particular realization) of observed data $\vec{x}$
  - The likelihood is used to define the information contained in a sample

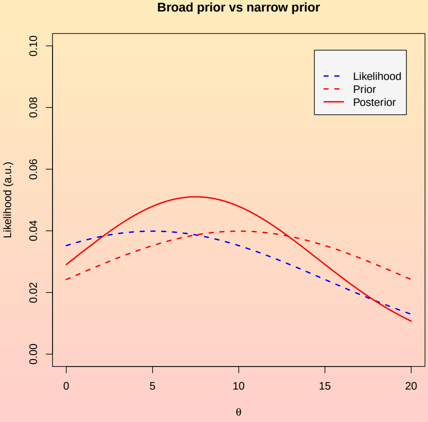
- Bayesian statistics automatically satisfies the likelihood principle
  - $P(\theta|\vec{x}) \propto L(\vec{x};\theta) \times \pi(\theta)$: the only quantity depending on the data is the likelihood
  - *Information* as a broad way of saying *all the possible inferences about* θ
  - “Probably tomorrow will rain”
- Frequentist statistics: *information* more strictly as *Fisher information* (connection with curvature of $L(\vec{x};\theta)$)
  - Usually does not comply (have to consider the hypothetical set of data that might have been obtained)
  - Need to recast question in terms of hypotetical data
  - Example: tail areas from sampling distributions obtained with toys
  - Even in forecasts: computer simulations of the day of tomorrow, or counting the past frequency of correct forecasts by the grandpa feeling arthritis in the shoulder
  - “The sentence -tomorrow it will rain- is probably true”
- The Likelihood Principle is quite vague: no practical prescription for drawing inference from the likelihood
  - Bayesian Maximum a-posteriori (MAP) estimator automatically maximizes likelihood
  - Maximum Likelihood estimator (MLE) maximizes likelihood automatically, but some foundational issues

- Two likelihoods differing by only a normalization factor are equivalent
  - Implies that information resides in the shape of the likelihood
- George Bernard: replace a dataset D with a dataset $D+Z$, where Z is the result of tossing a coin
  - Assume that the coin toss is independent on the parameter θ you seek to determine
  - Sampling probability: $p(DZ|\theta) = p(D|\theta)p(Z)$
  - The coin toss tells us nothing about the parameter θ beyond what we already learn by considering D only
  - Any inference we do with D must therefore be the same as any inference we do with $D+Z$
  - In particular, normalizations cancel out in ratio: $\frac{\mathcal{L}_1}{\mathcal{L}_2} = \frac{p(DZ|\theta_1 I)}{p(DZ|\theta_2 I)} = \frac{p(D|\theta_1 I)}{p(D|\theta_2 I)}$
- Do you believe probability comes from the imperfect knowledge of the observer?
  - Then the likelihood principle does not seem too profound besides the mathematical simplifications it allows
- Do you believe that probability is a physical phaenomenon arising from *randomness*?
  - Then the likelihood principle has for you a profound meaning of valid principle of inference

Likelihood and Fisher Information

- A very narrow likelihood will provide much information about θ_{true}
  - The posterior probability will be more localized than the prior in the regimen in which the likelihood function dominates the product $L(\vec{x};\vec{\theta}) \times \pi$
  - Ideally we'd want to connect this with the Fisher Information, which therefore be large
- A very broad likelihood will not carry much information, and ideally the computed Fisher Information will be small
- What's a reasonable definition of Fisher Information based on the likelihood function? Question time: Likelihood and Information





UCLouvain
Institut de recherche
en mathématique et physique

- Score: $\frac{\partial}{\partial\theta} lnL(X;\theta)$
- Under broad regularity conditions, if $X \sim f(x|\theta_{true})$ the expectation of the score calculated for $\theta = \theta_{true}$ is zero

$$E\Big[\frac{\partial}{\partial\theta} lnL(X;\theta)|\theta = \theta_{true}\Big] = \frac{\partial}{\partial\theta}\int f(x|\theta_{true})dx = \frac{\partial}{\partial\theta}1 = 0$$

- **Fisher Information:** the variance of the score

$$I(\theta) = E\Big[\Big(\frac{\partial}{\partial\theta} lnL(X;\theta)\Big)^2|\theta_{true}\Big] = \int \Big(\frac{\partial}{\partial\theta} lnf(x|\theta)\Big)^2 f(x|\theta)dx \geq 0$$

- Under some regularity conditions, and when the likelihood is twice differentiable, then you can “exchange” the exponent and the number of derivations

$$I(\theta) = -E\Big[\Big(\frac{\partial^2}{\partial\theta^2} lnL(X;\theta)\Big)^2|\theta_{true}\Big]$$

UCLouvain
Institut de recherche
en mathématique et physique

- The narrowness of the likelihood can be estimated by looking at its curvature
- The curvature is the second derivative with respect to the parameter of interest
- A very narrow (peaked) likelihood is characterized by a very large and positive curvature $-\frac{\partial^2 lnL}{\partial\theta^2}$
- The second derivative of the likelihood is linked to the Fisher Information

$$I(\theta) = -E\left[\frac{\partial^2 lnL}{\partial\theta^2}\right] = E\left[\left(\frac{\partial lnL}{\partial\theta}\right)^2\right]$$

Fisher Information and Jeffreys priors

- When changing variable, the change of parameterization must not result in a change of the information
  - The information is a property of the data only, through the likelihood—that summarizes them completely (likelihood principle)
- Search for a parametrization $\theta'(\theta)$ in which the Fisher Information is constant
- Compute the prior as a function of the new variable

$$\begin{aligned}
\pi(\theta) = \pi(\theta')\left|\frac{d\theta'}{d\theta}\right| &\propto \sqrt{E\left[\left(\frac{\partial lnN}{\partial\theta'}\right)^2\right]}\left|\frac{\partial\theta'}{\partial\theta}\right| \\
&= \sqrt{E\left[\left(\frac{\partial lnL}{\partial\theta'}\frac{\partial\theta'}{\partial\theta}\right)^2\right]} \\
&= \sqrt{E\left[\left(\frac{\partial lnL}{\partial\theta}\right)^2\right]} \\
&= \sqrt{I(\theta)}
\end{aligned}$$

- For any θ, $\pi(\theta) = \sqrt{I(\theta)}$; with this choice, the information is constant under changes of variable
- Such priors are called Jeffreys priors, and assume different forms depending on the type of parametrization
  - Location parameters: uniform prior
  - Scale parameters: prior $\propto \frac{1}{\theta}$
  - Poisson processes: prior $\propto \frac{1}{\sqrt{\theta}}$

The Maximum Likelihood Method 1/

- Let $\vec{x} = (x_1, ..., x_N)$ be a set of N statistically independent observations x_i, sampled from a p.d.f. $f(x; \vec{\theta})$ depending on a vector of parameters
- Under independence of the observations, the likelihood function factorizes into the individual p.d.f. s

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^{N} f(x_i, \vec{\theta})$$

- The maximum-likelihood estimator is the $\vec{\theta}_{ML}$ which maximizes the joint likelihood

$$\vec{\theta}_{ML} := argmax_{\theta} \left(L(\vec{x}, \vec{\theta}) \right)$$

  - The maximum must be global
- Numerically, it's usually easier to minimize

$$-\ lnL(\vec{x}; \vec{\theta}) = -\sum_{i=1}^{N} lnf(x_i, \vec{\theta})$$

  - Easier working with sums than with products
  - Easier minimizing than maximizing
- If the minimum is far from the range of permitted values for $\vec{\theta}$, then the minimization can be performed by finding solutions to

$$-\frac{lnL(\vec{x}; \vec{\theta})}{\partial \theta_j} = 0$$

  - It is assumed that the p.d.f. s are correctly normalized, i.e. that $\int f(\vec{x}; \vec{\theta}) dx = 1$ ($\rightarrow$ integral does not depend on $\vec{\theta}$)

UCLouvain
Institut de recherche
en mathématique et physique

- Solutions to the likelihood minimization are found via numerical methods such as MINOS
  - Fred James' Minuit: https://root.cern.ch/root/htmldoc/guides/minuit2/Minuit2.html
- $\vec{\theta}_{ML}$ is an estimator $\rightarrow$ let's study its properties!
  1. **Consistent:** $\lim_{N\rightarrow\infty} \vec{\theta}_{ML} = \vec{\theta}_{true}$;
  2. **Unbiased:** only asymptotically. $\vec{b} \propto \frac{1}{N}$, so $\vec{b} = 0$ only for $N \rightarrow \infty$;
  3. **Efficient:** $V[\vec{\theta}_{ML}] = \frac{1}{I(\theta)}$
  4. **Invariant:** for change of variables $\psi = g(\theta)$; $\hat{\psi}_{ML} = g(\vec{\theta}_{ML})$
- $\vec{\theta}_{ML}$ is only asymptotically unbiased, and therefore it does not always represent the best trade-off between bias and variance
- Remember that in frequentist statistics $L(\vec{x};\vec{\theta})$ is not a p.d.f. . In Bayesian statistics, the posterior probability is a p.d.f.:

$$P(\vec{\theta}|\vec{x}) = \frac{L(\vec{x}|\vec{\theta})\pi(\vec{\theta})}{\int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})d\vec{\theta}}$$

  - Note that if the prior is uniform, $\pi(\vec{\theta}) = k$, then the MLE is also the maximum of the posterior probability, $\vec{\theta}_{ML} = maxP(\vec{\theta}|\vec{x})$.

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.
- The joint p.d.f. can be factorized

$$f(t_1, ...t_N; \tau) = \prod_i f(t_i; \tau)$$

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.
- The joint p.d.f. can be factorized

$$f(t_1, ...t_N;\tau) = \prod_i f(t_i;\tau)$$

- For a particular set of N measurements t_i, the p.d.f. can be written as a function of τ only, $L(\tau) := f(t_i;\tau)$

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.
- The joint p.d.f. can be factorized

$$f(t_1, ...t_N; \tau) = \prod_i f(t_i; \tau)$$

- For a particular set of N measurements t_i, the p.d.f. can be written as a function of τ only, $L(\tau) := f(t_i; \tau)$
- Now all you need to do is to maximize the likelihood

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau}e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.
- The joint p.d.f. can be factorized

$$f(t_1,...t_N;\tau) = \prod_i f(t_i;\tau)$$

- For a particular set of N measurements t_i, the p.d.f. can be written as a function of τ only, $L(\tau) := f(t_i;\tau)$
- Now all you need to do is to maximize the likelihood
- The logarithm of the likelihood, $lnL(\tau) = \sum \left(ln\frac{1}{\tau} - \frac{t_i}{\tau} \right)$, can be maximized analytically

$$\frac{\partial lnL(\tau)}{\partial \tau} = \sum_i \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) \equiv 0$$

Nuclear Decay with Maximum Likelihood Method

UCLouvain
Institut de recherche
en mathématique et physique

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient? QT: N D 1

Nuclear Decay with Maximum Likelihood Method

UCLouvain
Institut de recherche
en mathématique et physique

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient? QT: N D 1
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N} \sum_i t_i\right] = \frac{1}{N^2} \sum_i V[t_i] = \frac{\tau^2}{N}$$

Nuclear Decay with Maximum Likelihood Method

UCLouvain
Institut de recherche
en mathématique et physique

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient? QT: N D 1
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N}\sum_i t_i\right] = \frac{1}{N^2}\sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. Fill the table!

	Consistent	Unbiased	Efficient
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1+\ldots+t_N}{N}$			
$\hat{\tau} = \frac{t_1+\ldots+t_N}{N-1}$			
$\hat{\tau} = t_i$			

Table: Properties of different estimators of the half life for a nuclear decay.

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient? QT: N D 1
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N}\sum_i t_i\right] = \frac{1}{N^2}\sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. Fill the table!

	Consistent	Unbiased	Efficient
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1+\ldots+t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1+\ldots+t_N}{N-1}$			
$\hat{\tau} = t_i$			

Table: Properties of different estimators of the half life for a nuclear decay.

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient? QT: N D 1
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N} \sum_i t_i\right] = \frac{1}{N^2} \sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. Fill the table! Question time: Nuclear Decay 2

	Consistent	Unbiased	Efficient
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1 + \ldots + t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1 + \ldots + t_N}{N-1}$	✓	✗	✗
$\hat{\tau} = t_i$			

Table: Properties of different estimators of the half life for a nuclear decay.

Nuclear Decay with Maximum Likelihood Method

fnrs LA LIBERTÉ DE CHERCHER
UCLouvain
Institut de recherche
en mathématique et physique

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient? QT: N D 1
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N}\sum_i t_i\right] = \frac{1}{N^2}\sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. Fill the table! Question time: Nuclear Decay 2

	Consistent	Unbiased	Efficient
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1+\ldots+t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1+\ldots+t_N}{N-1}$	✓	✗	✗
$\hat{\tau} = t_i$	✗	✓	✗

Table: Properties of different estimators of the half life for a nuclear decay.

Why $\hat{\tau} = t_i$ is unbiased

- Bias: $b = E[\hat{\tau}] - \tau$
  - Note: if you don't know the true value, you must simulate the bias of the method
  - Generate toys with known parameters, and check what is the estimate of the parameter for the toy data
  - If there is a bias, correct for it to obtain an unbiased estimator
- t_i is an individual observation, which is still sampled from the original factorized p.d.f. $f(t_i;\tau) = \frac{1}{\tau}e^{-\frac{t_i}{\tau}}$
- The expected value of t_i is therefore still $E[\hat{\tau}] = E[t_i] = \tau$
- $\hat{\tau} = t_i$ is therefore unbiased!

	Consistent	Unbiased	Efficient
$\hat{\tau} = t_i$	✗	✓	✗

Table: Properties of different estimators of the half life for a nuclear decay.

- We usually want to optimize both bias $\vec{b}$ and variance $V[\hat{\vec{\theta}}]$
- While we can optimize each one separately, optimizing them simultaneously leads to none being optimally optimized, in genreal
  - Optimal solutions in two dimensions are often suboptimal with respect to the optimization of just one of the two properties
- The variance is linked to the width of the likelihood function, which naturally leads to linking it to the curvature of $L(\vec{x};\vec{\theta})$ near the maximum
- However, the curvature of $L(\vec{x};\vec{\theta})$ near the maximum is linked to the Fisher information, as we have seen
- Information is therefore a limiting factor for the variance (no data set contains infinite information, variance cannot collapse to zero)
- Variance of an estimator satisfies the Rao-Cramér-Frechet (RCF) bound

$$V[\hat{\theta}] \geq \frac{1}{\hat{\theta}}$$

- Rao-Cramer-Frechet (RCF) bound
  $V[\hat{\theta}] \geq \frac{(1+\partial b/\partial\theta)^2}{-E\left[\partial^2 lnL/\partial\theta^2\right]}$
  - In multiple dimensions, link with the information is maintaned via the full Fisher Information Matrix: $I_{ij} = E\left[\partial^2 lnL/\partial\theta_i\partial\theta_j\right]$
- Approximations
  - Neglect the bias ($b = 0$)
  - Inequality is an approximate equality (true for large data samples)
- $V[\hat{\theta}] \simeq \frac{1}{-E\left[\partial^2 lnL/\partial\theta^2\right]}$
- Estimate of the variance of the estimate of the parameter!
- $\hat{V}[\hat{\theta}] \simeq \frac{1}{-E\left[\partial^2 lnL/\partial\theta^2\right]|_{\theta=th\hat{e}ta}}$
- For a generic unbiased estimator, can define *efficiency* of the estimator as

$$e(\hat{\theta}) := \frac{I(\theta)^{-1}}{V[\hat{\theta}]}$$

- The efficiency of a generic unbiased estimator, because of the RCF bound, is always $e(\hat{\theta}) \leq 1$

- For multidimensional parameters, we can build the information matrix with elements:

$$
\begin{aligned}
I_{jk}(\vec{\theta}) &= -E\Big[\sum_{i}^{N} \frac{\partial^2 lnf(x_i;\vec{\theta})}{\partial\theta_k\partial\theta_k}\Big] \\
&= N\int \frac{1}{f}\frac{\partial f}{\partial\theta_j}\frac{\partial f}{\partial\theta_k}dx
\end{aligned}
$$

- (the last equality is due to the integration interval not being dependent on $\vec{\theta}$)

UCLouvain
Institut de recherche
en mathématique et physique

- We have calculated the variance of the MLE in the simple case of the nuclear decay
- Analytic calculation of the variance is not always possible
- Write the variance approximately as:

$$V[\hat{\theta}] \geq \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{-E\left[\frac{\partial^2 lnL}{\partial \theta^2}\right]}$$

- This expression is valid for any estimator, but if applied to the MLE then we can note $\vec{\theta}_{ML}$ is efficient and asymptotically unbiased
- Therefore, when $N \to \infty$ then $b = 0$ and the variance approximate to the RCF bound, and $\geq$ becomes $\simeq$:

$$V[\vec{\theta}_{ML}] \simeq \frac{1}{-E\left[\frac{\partial^2 lnL}{\partial \theta^2}\right]\Big|_{\theta=\vec{\theta}_{ML}}}$$

- For a Gaussian p.d.f., $f(x;\vec{\theta}) = N(\mu,\sigma)$, the likelihood can be written as:

$$L(\vec{x};\vec{\theta}) = ln\Big[-\frac{(\vec{x}-\vec{\theta})^2}{2\sigma^2}\Big]$$

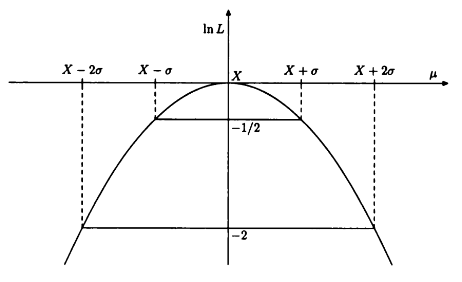
- Moving away from the maximum of $L(\vec{x};\vec{\theta})$ by one unit of σ, the likelihood assumes the value $\frac{1}{2}$, and the area enclosed in $[\vec{\theta}-\sigma,\vec{\theta}+\sigma]$ will be—because of the properties of the Normal distribution—equal to 68.3%.

How to extract an interval from the likelihood function 2/

- We can therefore write

$$P\Big((\vec{x}-\vec{\theta})^2 \leq \sigma)\Big) = 68.3\%$$
$$P(-\sigma \leq \vec{x}-\vec{\theta} \leq \sigma) = 68.3\%$$
$$P(\vec{x}-\sigma \leq \vec{\theta} \leq \vec{x}+\sigma) = 68.3\%$$

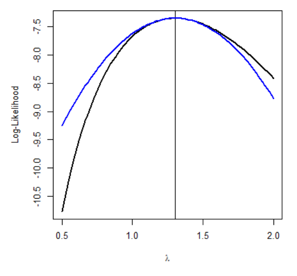
- Taking into account that it is important to keep in mind that probability is a property of sets, in frequentist statistics
  - Confidence interval: interval with a fixed probability content
- This process for computing a confidence interval is exact for a Gaussian p.d.f.
  - Pathological cases reviewed later on (confidence belts and Neyman construction)
- Practical prescription:
  - Point estimate by computing the Maximum Likelihood Estimate
  - Confidence interval by taking the range delimited by the crossings of the likelihood function with $\frac{1}{2}$ (for 68.3% probability content, or 2 for 95% probability content—2σ, etc)

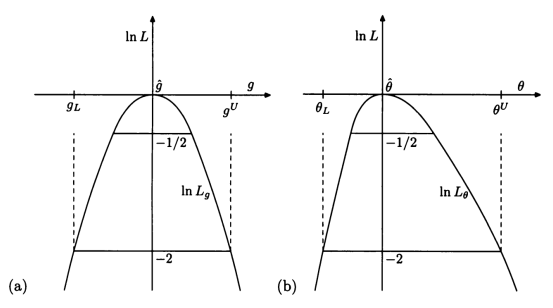


Plot from James, 2nd ed.

How to extract an interval from the likelihood function 3/

- MLE is invariant for monotonic transformations of θ
  - This applies not only to the maximum of the likelihood, but to all relative values
  - The likelihood ratio is therefore an invariant quantity (we'll use it for hypothesis testing)
  - Can transform the likelihood such that $log(L(\vec{x};\vec{\theta}))$ is parabolic, but not necessary (MINOS/Minuit)
- When the p.d.f. is not normal, either assume it is, and use symmetric intervals from Gaussian tails...
  - This yields symmetric approximate intervals
  - The approximation is often good even for small amounts of data
- ...or use asymmetric intervals by just looking at the crossing of the $log(L(\vec{x};\vec{\theta}))$ values
  - Naturally-arising asymmetrical intervals
  - No gaussian approximation
- In any case (even asymmetric intervals) still based on asymptotic expansion
  - Method is exact only to $\mathcal{O}(\frac{1}{N})$





Plot from James, 2nd ed.

UCLouvain
Institut de recherche
en mathématique et physique

- Theorem: for <u>any p.d.f.</u> $f(x|\vec{\theta})$, in the large numbers limit $N \rightarrow \infty$, the likelihood can always be approximated with a gaussian:

$$L(\vec{x};\vec{\theta}) \propto_{N\rightarrow\infty} e^{-\frac{1}{2}(\vec{\theta}-\vec{\theta}_{ML})^T H(\vec{\theta}-\vec{\theta}_{ML})}$$

- where H is the information matrix $I(\vec{\theta})$.
- Under these conditions, $V[\vec{\theta}_{ML}] \rightarrow \frac{1}{I(\vec{\theta}_{ML})}$, and the intervals can be computed as:

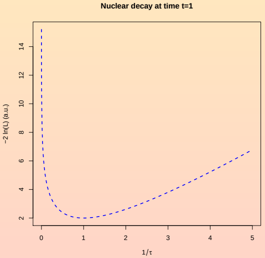
$$\Delta lnL := lnL(\theta') - lnL_{max} = -\frac{1}{2}$$

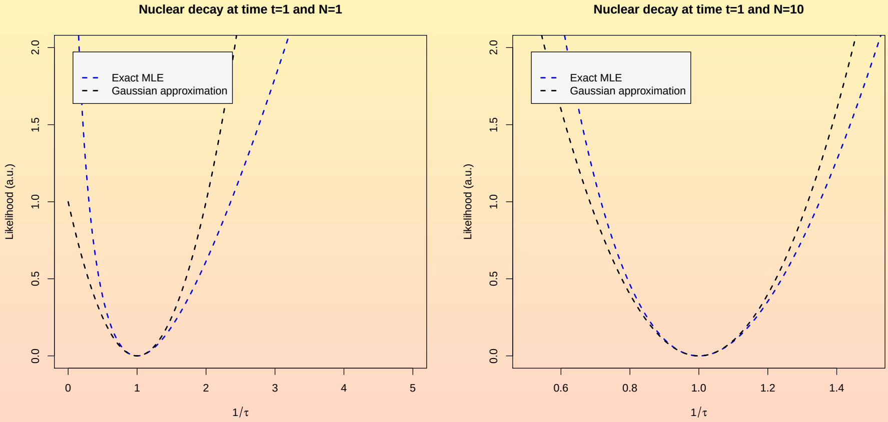
- The resulting interval has in general a larger probability content than the one for a gaussian p.d.f., but the approximation grows better when N increases
  - The interval <u>overcovers the true value $\vec{\theta}_{true}$</u>

UCLouvain
Institut de recherche en mathématique et physique

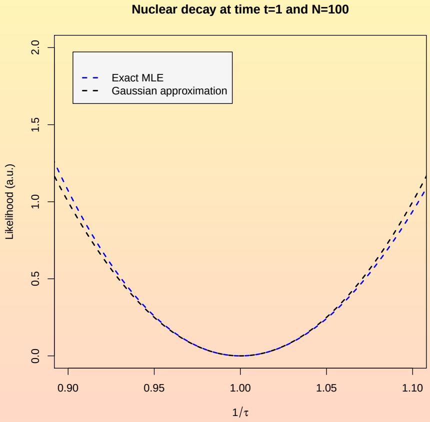
- $\vec{\theta}_{true}$ is therefore stimated as $\hat{\theta} = \vec{\theta}_{ML} \pm \sigma$. This is another situation in which frequentist and Bayesian statistics differ in the interpretation of the numerical result
- Frequentist: $\vec{\theta}_{true}$ is fixed
  - "if I repeat the experiment many times, computing each time a confidence interval around $\vec{\theta}_{ML}$, on average 68.3% of those intervals will contain $\vec{\theta}_{true}$"
  - Coverage: "the interval covers the true value with 68.3% probability"
  - Direct consequence of the probability being a property of <u>data sets</u>
- Bayesian: $\vec{\theta}_{true}$ is not fixed
  - "the true value $\vec{\theta}_{true}$ will be in the range $[\vec{\theta}_{ML} - \sigma, \vec{\theta}_{ML} + \sigma]$ with a probabilty of 68.3%"
  - This corresponds to giving a value for the posterior probability of the parameter $\vec{\theta}_{true}$

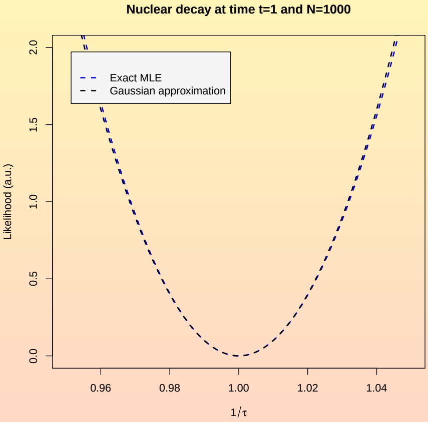
- How good is the approximation $L(\vec{x};\vec{\theta}) \propto exp\Big[-\frac{1}{2}(\vec{\theta}-\vec{\theta}_{MLE})^T H(\vec{\theta}-\vec{\theta}_{ML})\Big]$?
  - Here H is the information matrix $I(\vec{\theta})$
  - True only to $\mathcal{O}(\frac{1}{N})$
  - In these conditions, $V[\vec{\theta}_{ML}] \rightarrow \frac{1}{I(\vec{\theta}_{ML})}$
  - Intervals can be derived by crossings: $\Delta lnL = lnL(\theta') - lnL_{max} = k$
- This afternoon: we'll convince ourselves of how good is this approximation in case of the nuclear decay!




Nuclear decay at time t=1 and N=1
Exact MLE
Gaussian approximation
Likelihood (a.u.)
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
1/τ
Nuclear decay at time t=1 and N=10
Exact MLE
Gaussian approximation
Likelihood (a.u.)
0.0
0.5
1.0
1.5
2.0
0.6
0.8
1.0
1.2
1.4
1/τ

Non-normal likelihoods and Gaussian approximation — 3





UCLouvain
Institut de recherche
en mathématique et physique

- The convergence of the likelihood $L(\vec{x};\vec{\theta})$ to a gaussian is a direct consequence of the central limit theorem
- Take a set of measurements $\vec{x} = (x_i, ..., x_N)$ affected by experimental errors that results in uncertainties $\sigma_1, ..., \sigma_N$ (not necessarily equal among each other)
- In the limit of a large number of events, $M \to \infty$, the random variable built summing M measurements is gaussian-distributed:

$$Q := \sum_{j=1}^{M} x_j \sim N\Big(\sum_{j=1}^{M} x_j, \sum_{j=1}^{M} \sigma_j^2\Big), \qquad \forall\, f(x,\vec{\theta})$$

- The demonstration runs by expanding in series the characteristic function $y_i = \frac{x_j - \mu_j}{\sqrt{\sigma_j}}$
- The theorem is valid for any p.d.f. $f(x,\vec{\theta})$ that is reasonably peaked around its expected value.
  - If the p.d.f. has large tails, the bigger contributions from values sampled from the tails will have a large weight in the sum, and the distribution of Q will have non-gaussian tails
  - The consequence is an alteration of the probability of having sums Q outside of the gaussian

UCLouvain
Institut de recherche
en mathématique et physique

- The condition $M \to \infty$ is reasonably valid if the sum is of many small contributions.
- How large does M need to be for the approximation to be reasonably good? Question time: Central Limit

- The condition $M \to \infty$ is reasonably valid if the sum is of many small contributions.
- How large does M need to be for the approximation to be reasonably good? Question time: Central Limit
- This afternoon we'll check!

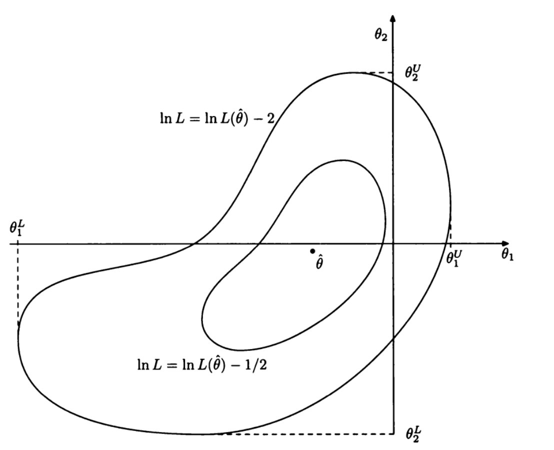
And in many dimensions...

- Construct $log\mathcal{L}$ contours and determine confidence intervals by MINOS
- Elliptical contours correspond to gaussian Likelihoods
  - The closer to MLE, the more elliptical the contours, even in non-linear problems
  - All models are linear in a sufficiently small region
- Nonlinear regions not problematic (no parabolic transformation of $log\mathcal{L}$ needed)
  - MINOS accounts for non-linearities by following the likelihood contour

- Confidence intervals for each parameter

$$\max_{\theta_j, j\neq i} log\mathcal{L}(\theta) = log\mathcal{L}(\hat{\theta}) - \lambda$$

- $\lambda = \frac{Z^2_{1-\beta}}{2}$
  - $\lambda = 1/2$ for $\beta = 0.683$ ("1σ")
  - $\lambda = 2$ for $\beta = 0.955$ ("2σ")



Plot from James, 2nd ed.

Profile likelihood ratio step by step for cross sections — Expected events

- We used to compute the total cross section of a given process by applying the naïve formula

$$\sigma = \frac{N_{data} - N_{bkg}}{\epsilon L} .$$

  - N_{sig} estimated from $N_{data} - N_{bkg}$ for the measured integrated luminosity L
  - The acceptance ϵ accounts for th. branching fractions fiducial region for the measurement (fiducial region: generator-level selection which defines the phase space of the measurement)
- Nowadays we model everything into the likelihood function
- $p(x|\mu, \theta)$ pdf for the observable x to assume a certain value in a single event
  - $\mu := \frac{\sigma}{\sigma_{pred}}$ (single- or multi-dimensional) *parameter of interest* (POI). A multiplier of the predicted cross section: *signal strength*
  - θ (generally multi-dimensional) *nuisance parameter* representing all the uncertainties affecting the measurement.
- Extend to a data set of many events $X = \{x_1, ..., x_n\}$ by taking the product of the single-event p.d.f.s.

$$\prod_{e=1}^{n} p(x_e|\mu, \theta)$$

Profile likelihood ratio step by step for cross sections — Expected events

- We used to compute the total cross section of a given process by applying the naïve formula

$$\sigma = \frac{N_{data} - N_{bkg}}{\epsilon L} .$$

  - N_{sig} estimated from $N_{data} - N_{bkg}$ for the measured integrated luminosity L
  - The acceptance ϵ accounts for th. branching fractions fiducial region for the measurement (fiducial region: generator-level selection which defines the phase space of the measurement)
- Nowadays we model everything into the likelihood function
- $p(x|\mu, \theta)$ pdf for the observable x to assume a certain value in a single event
  - $\mu := \frac{\sigma}{\sigma_{pred}}$ (single- or multi-dimensional) *parameter of interest* (POI). A multiplier of the predicted cross section: *signal strength*
  - θ (generally multi-dimensional) *nuisance parameter* representing all the uncertainties affecting the measurement.
- Extend to a data set of many events $X = \{x_1, ..., x_n\}$ by taking the product of the single-event p.d.f.s.

$$\prod_{e=1}^{n} p(x_e|\mu, \theta)$$

- The number of events in the data set is however a random variable itself!
  - Poisson distribution with mean equal to the number of events ν we expect from theory
- *Marked Poisson model*

$$f(X|\nu(\mu, \theta), \mu, \theta) = Pois(n|\nu(\mu, \theta)) \prod_{e=1}^{n} p(x_e|\mu, \theta) .$$

Pleasant quality read: Vischia, 2019 doi:10.1016/j.revip.2020.100046 ☺

- Both μ and θ act on the individual pdfs for the observable and on the expectation for the global amount of events
- Incorporate systematic uncertainties as nuisance parameter θ: Conway, 2011 in CERN-2011-006115
  - Constrain the terms in the fit: constraint interpreted as prior coming from the auxiliary measurement
  - θ estimated with uncertainty $\delta\theta$
  - Often Gaussian pdf, unless θ has a physical bound at zero: then log-normal (rejects negative values)
- Likelihood $\mathcal{L}(\mu, \theta; X)$: take the marked Poisson model $f(X|\nu(\mu,\theta), \mu, \theta)$ and condition on the observed value of X
- MLE: $\hat{\mu} := argmax_{\mu}\mathcal{L}(\mu, \theta; X)$ still depends on the nuisance parameters θ

$$\mathcal{L}(\boldsymbol{n}, \boldsymbol{\alpha}^{\mathbf{0}}|\mu, \boldsymbol{\alpha}) = \prod_{i\in bins} \mathcal{P}(n_i|\mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j\in syst} \mathcal{G}(\alpha_j^0|\alpha_j, \delta\alpha_j)$$

$$\downarrow$$

$$\mathcal{L}(\boldsymbol{n}, 0|\mu, \boldsymbol{\alpha}) = \prod_{i\in bins} \mathcal{P}(n_i|\mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j\in syst} \mathcal{G}(0|\alpha_j, 1)$$

Pleasant quality read: Vischia, 2019 doi:10.1016/j.revip.2020.100046 ☺

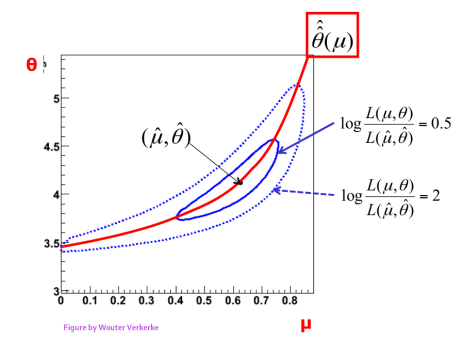
Eliminate dependence on the nuisance parameters

UCLouvain
Institut de recherche
en mathématique et physique

- Likelihood ratio!

$$\lambda(\mu) := \frac{\mathcal{L}(\mu, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu}, \hat{\theta})} .$$

- Denominator $\mathcal{L}(\hat{\mu}, \hat{\theta})$ is computed for the values of μ and θ which jointly maximize the likelihood function.
  - *Profiling*: eliminating the dependence on the nuisance parameters by taking their conditional maximum likelihood estimate
  - Bayesians normally marginalize (integrate) rather than profiling (see Demortier, 2002)
- The maximum of the likelihood ratio yields the point estimate for μ
- The second derivative of the maximum likelihood ratio yields intervals on the parameter μ
  - Tomorrow: the tricky cases (e.g. point estimate near the physical range allowed for the parameter)



Pleasant quality read: Vischia, 2019 doi:10.1016/j.revip.2020.100046 ☺

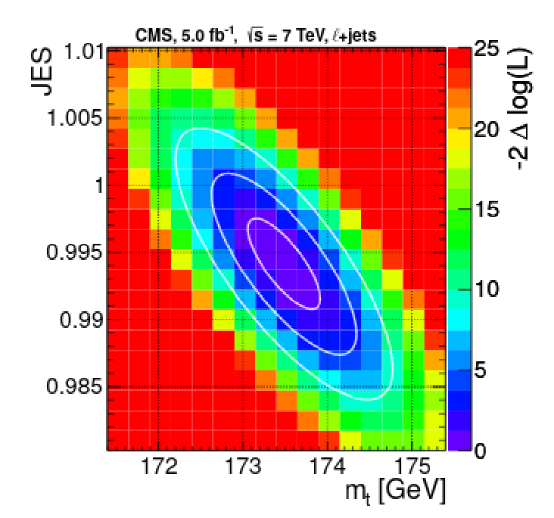
What do I need to profile?

UCLouvain
Institut de recherche
en mathématique et physique

- The likelihood ratio $\lambda(\mu) = \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})}$
- Conceptually, you can run the experiment many times (e.g. toys) and record the value of the test statistic
- The test statistic can therefore be seen as a distribution
- Asymptotically, $\lambda(\mu) \sim exp\Big[-\frac{1}{2}\chi^2\Big]\Big(1+\mathcal{O}(\frac{1}{\sqrt{N}})\Big)$ (Wilks Theorem, under some regularity conditions—continuity of the likelihood and up to 2nd derivatives, existence of a maximum, etc)
  - The χ^2 distribution depends only on a single parameter, the number of degrees of freedom
  - It follows that the test statistic is <u>independent of the values of the nuisance parameters</u>
  - Useful: you don't need to make toys in order to find out how is $\lambda(\mu)$ distributed!

What is a nuisance parameter?

- Sometimes the classification into POI and nuisance parameter washes out
- Maybe you data and your method can provide information on a systematic uncertainty



Plot from doi:10.1007/JHEP12(2012)105

- More often, the analysis is not sensitive enough to treat an uncertainty as POI and measure it
- The fit can still constrain the nuisance parameter that is profiled
- Indirectly provides information about your estimate of that parameter before the fit
  - Over- or under-estimate θ before the fit
  - See a best fit value for θ that doesn't match very well with the prefit value
- Quote, for each nuisance parameter, two important quantities
  - **Pull**: the difference of the post-fit and pre-fit values of the parameter, normalized to the pre-fit uncertainty: $pull := \frac{\hat{\theta}-\theta}{\delta\theta}$
  - **Constraint:** the ratio between the post-fit and the pre-fit uncertainty in the nuisance parameter.

- **Pull**: the difference of the post-fit and pre-fit values of the parameter, normalized to the pre-fit uncertainty: $pull := \frac{\hat{\theta}-\theta}{\delta\theta}$
- **Constraint:** the ratio between the post-fit and the pre-fit uncertainty in the nuisance parameter.
- Spot easily possible issues in the fit
  - θ pulled too much may be a hint that our estimate of the pre-fit value was not reasonable
  - θ constrained too much indicates that the data contain enough information to improve the precision in the nuisance parameter with respect to our original estimate, which may or may not make sense.

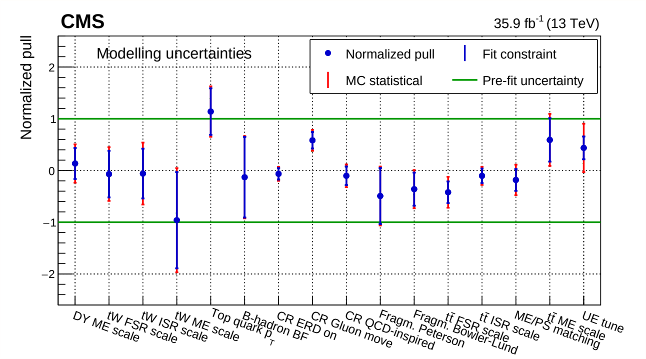


Image collected and cited in doi:10.1016/j.revip.2020.100046, references therein)

UCLouvain
Institut de recherche
en mathématique et physique

- What is more worrying, a small pull with a small constraint, or a large pull with a strong constraint? Question time: Pulls and Constraints
- A pull with very small constraint: $\theta_{prefit} = 0 \pm 1$, $\theta_{postfit} = 1 \pm 0.9$
- The same pull with a strong constraint: $\theta_{prefit} = 0 \pm 1$, $\theta_{postfit} = 1 \pm 0.2$

- What is more worrying, a small pull with a small constraint, or a large pull with a strong constraint? Question time: Pulls and Constraints
- A pull with very small constraint: $\theta_{prefit} = 0 \pm 1$, $\theta_{postfit} = 1 \pm 0.9$
- The same pull with a strong constraint: $\theta_{prefit} = 0 \pm 1$, $\theta_{postfit} = 1 \pm 0.2$

- A way of estimating if a shift is significant is to compare the shift with its uncertainty
- For independent measurements, the compatibility C is

$$C = \Delta\theta / \sigma_{\Delta\theta} = \frac{\theta_2 - \theta_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

- We would conclude that the first case $C = 0.74$, for the second one $C = 0.98$ (larger, still within uncertainty)

Question time, pulls and constraints

- What is more worrying, a small pull with a small constraint, or a large pull with a strong constraint? Question time: Pulls and Constraints
- A pull with very small constraint: $\theta_{prefit} = 0 \pm 1$, $\theta_{postfit} = 1 \pm 0.9$
- The same pull with a strong constraint: $\theta_{prefit} = 0 \pm 1$, $\theta_{postfit} = 1 \pm 0.2$

- A way of estimating if a shift is significant is to compare the shift with its uncertainty
- For independent measurements, the compatibility C is

$$C = \Delta\theta / \sigma_{\Delta\theta} = \frac{\theta_2 - \theta_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

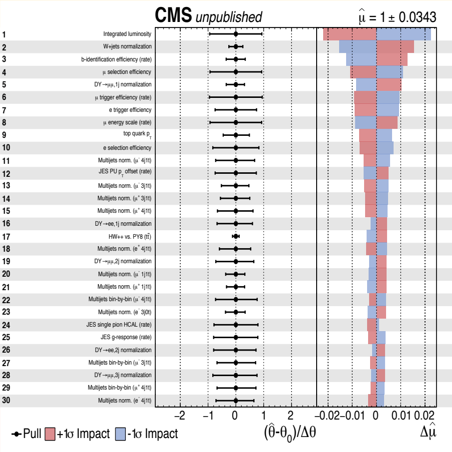
- We would conclude that the first case $C = 0.74$, for the second one $C = 0.98$ (larger, still within uncertainty)
- However, these are not independent measurements!
- The formula is therefore

$$C = \Delta\theta / \sigma_{\Delta\theta} = \frac{\theta_2 - \theta_1}{\sqrt{\sigma_1^2 - \sigma_2^2}}$$

- For the first case, $C = 2.29$, for the second case $C = 1.02$
- The same pull is more significant if there is (almost no) constraint!!!

Impacts

- *Impact* of θ on the post-fit signal strength permits to obtain a ranking of the nuisance parameters in terms of their effect on the signal strength
  - Fix each nuisance parameter to its post-fit value $\hat{\theta}$ plus/minus its pre-fit (post-fit) uncertainty $\delta\theta$ ($\delta\hat{\theta}$)
  - Reperform the fit for μ
  - Compute the impact as the difference between the original fitted signal strength and the refitted signal strength.
- Results on Asimov dataset (replacing the data with the expectations from simulated events) is expected to give “perfect” results



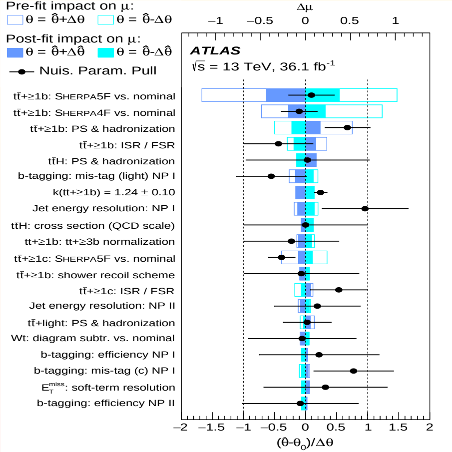
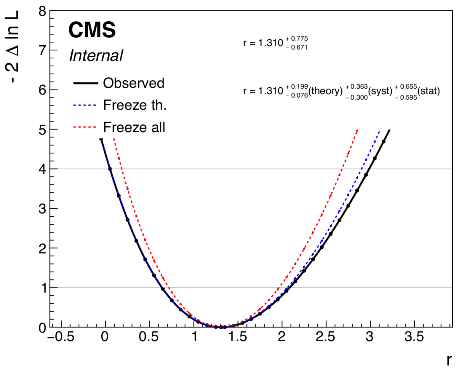


Image collected and cited in doi:10.1016/j.revip.2020.100046, references therein)

Breakdown of systematic uncertainties

- What's the amount of uncertainty that is impotable to a given set of systematic effects?
  - The modern expression of Fisher's formalization of the ANOVA concept
  - *"the constituent causes fractions or percentages of the total variance which they together produce"* (Fisher, 1919)
  - *"the variance contributed by each term, and by which the residual variance is reduced when that term is removed"* (Fisher, 1921)
- Breakdown the contributions
  - Freeze a set of uncertainties θ_i to their post-fit value
  - Repeat the fit to extract a new (smaller) uncertainty on μ
  - Obtain the contribution of θ_i to the overall uncertainty as squared difference betwee the full and reduced uncertainties
  - Statistical uncertainty obtained by freezing all nuisance parameters



Toy data

From sidebands to systematic uncertainties

- Measure a background rate in a sideband, use the estimate in the signal region
- As described, let's model our estimation problem using profile likelihoods

$$\mathcal{L}(\boldsymbol{n}, \boldsymbol{\alpha}^{\boldsymbol{0}} | \mu, \boldsymbol{\alpha}) = \prod_{i \in bins} \mathcal{P}(n_i | \mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j \in syst} \mathcal{G}(\alpha_j^0 | \alpha_j, \delta\alpha_j)$$

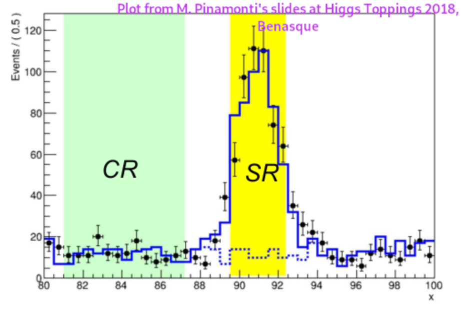
$$\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\boldsymbol{\alpha}}}_\mu)}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\alpha}})}$$

- Sideband measurement

$$L_{SR}(s,b) = Poisson(N_{SR} \mid s+b)$$

$$L_{CR}(b) = Poisson(N_{CR} \mid \tilde{\tau} \cdot b)$$

$$\mathcal{L}_{full}(s,b) = \mathcal{P}(N_{SR} | s+b) \times \mathcal{P}(N_{CR} | \tilde{\tau} \cdot b)$$



- Subsidiary measurement of the background rate:
  - *8% systematic uncertainty on the MC rates*
  - $\tilde{b}$: *measured background rate by MC simulation*
  - $\mathcal{G}(\tilde{b} | b, 0.08)$: our

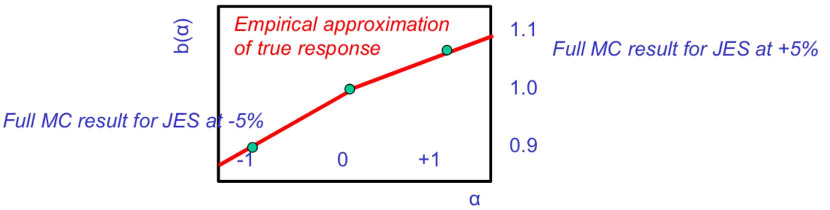
$$\mathcal{L}_{full}(s,b) = \mathcal{P}(N_{SR} | s+b) \times \mathcal{G}(\tilde{b} | b, 0.08)$$

$$\mathcal{L}(\boldsymbol{n}, \boldsymbol{\alpha^0}|\mu, \boldsymbol{\alpha}) = \prod_{i\in bins} \mathcal{P}(n_i|\mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j\in syst} \mathcal{G}(\alpha_j^0|\alpha_j, \delta\alpha_j)$$

$$\downarrow$$

$$\mathcal{L}(\boldsymbol{n}, 0|\mu, \boldsymbol{\alpha}) = \prod_{i\in bins} \mathcal{P}(n_i|\mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j\in syst} \mathcal{G}(0|\alpha_j, 1)$$

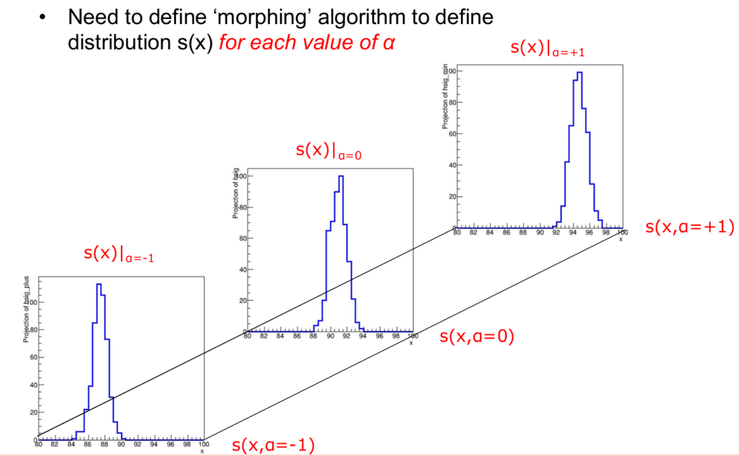
- Subsidiary measurement often labelled *constraint term*
- It is not a PDF in α: $\mathcal{G}(\alpha_j|0, 1) \neq \mathcal{G}(0|\alpha_j, 1)$
- Response function: $\tilde{B}_i(1 + 0.1\alpha)$ (a unit change in α –e.g. 5% JES– changes the acceptance by 10%)



Graphics from W. Verkerke

Interpolation needed between template models

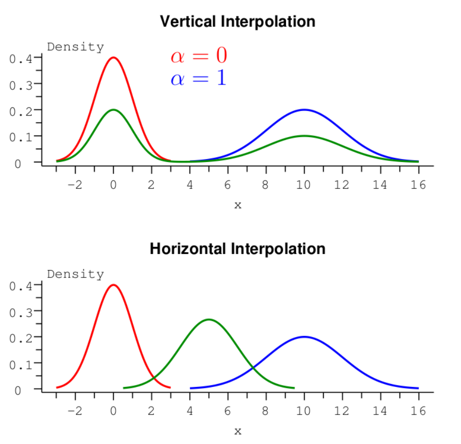
- Conditional density $f(x|\alpha)$ constructed by some means for a discrete set of values $\alpha_1, \dots \alpha_N$
- The exact dependence of $f(x|\alpha)$ on α is unknown
  - In practice $f(x|\alpha_i)$ often nonparametric density estimates in the x space (e.g. histograms)
- Problem: determine $f(x|\alpha)$ for arbitrary α_i
  - Typically α_i within the cloud of $\alpha_1, \dots \alpha_N$, and direct calculation too expensive
  - Need to keep the densities normalized: $\int f(x|\alpha)dx = 1, \forall \alpha$



Graphics from W. Verkerke

Horizontal or vertical morphing?

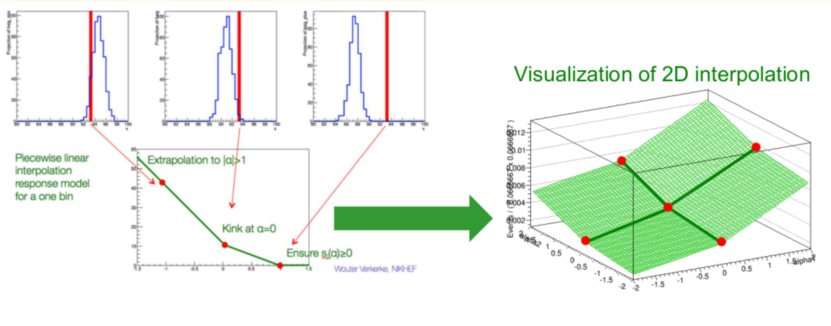
- Vertical interpolation is often not what you want
  - Except some cases, e.g. interpolation of detector efficiency curves



- Vertical interpolation of single-parameter 1D densities:
  $f(x|\alpha) = w_1 f(x|\alpha_1) + (1 - w_1) f(x|\alpha_2)$,
  $w_1 = \frac{\alpha_2 - \alpha}{\alpha_2 - \alpha_1}$, $\alpha \in [\alpha_1, \alpha_2]$
- Horizontal interpolation: identical parameter dependence, but interpolate quantile function
  $q(y|\alpha) = w_1 q(y|\alpha_1) + (1 - w_1) q(y|\alpha_2)$,
  $q(y|\alpha) := F^{-1}(y|\alpha)$
  - Have to solve $q(y|\alpha) = x$ numerically
  - Difficult to evaluate numerically around $y = 0$ and $y = 1$

Horizontal interpolation/morphing in one dimension

- For HEP application and univariate densities, reasonable solution is linear interpolation
  - A.L. Read, Linear interpolation of histograms, NIM A 425, 357 (1999)
  - Can fail dramatically if the change in shape is comparable with or smaller than MC statistical fluctuations
  - Sometimes we may want to avoid adding this new degree of freedom in the model
  - Decoupling rate and shape effects is always possible, even when not neglecting the shape ones)

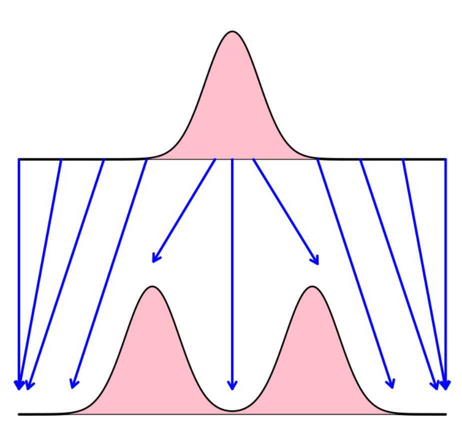


Graphics from W. Verkerke

- The cases $f(\vec{x}|\alpha)$ and $f(\vec{x}|\vec{\alpha})$ remain delicate
- Multivariate parameters: $g(\cdot|\vec{\alpha}) = \sum_{i=1}^{N} w_i(\vec{\alpha}, \vec{\alpha_1}, ..., \vec{\alpha_N}) g(\cdot)\vec{\alpha_i}$
  - $g(\cdot|\vec{\alpha})$ either density function (x) or quantile function (y)
  - Non-negative weights summing up to 1; many techniques (polinomial, local poly, spline best used in 1D)
  - Lack of generality because assumes Euclidean space

What if our metric is not Euclidean?

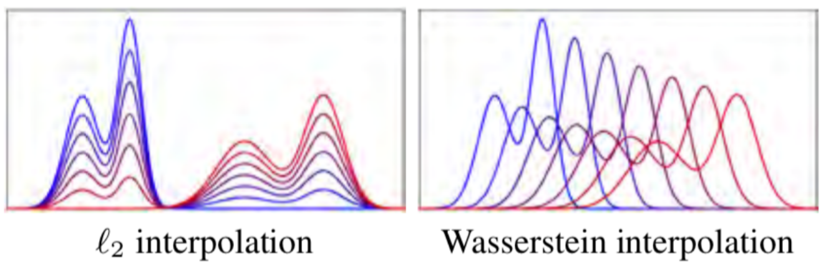
- Given two distributions P_0 and P_1, define an *optimal map* T transforming $X \sim P_0$ into $T(X) \sim P_1$ (Monge, 1781)
- Define a geodesic path between P_0 and P_1 in the space of the distributions, according to a given metric
  - Shape-preserving notion of averages of distributions
  - Distance based on transport along geodesic paths
- Let $X \sim P_0$, and find T by minimizing $\mathbb{E}\left[\parallel X - T(X) \parallel^p \right] = \int \parallel x - T(x) \parallel^p dP_0(x)$
  - Minimization over all T s.t. $T(X) \sim P_1$. Can replace Euclidean distance with any distance
  - The minimizer is called *optimal transport map*



Graphics by Wasserstein

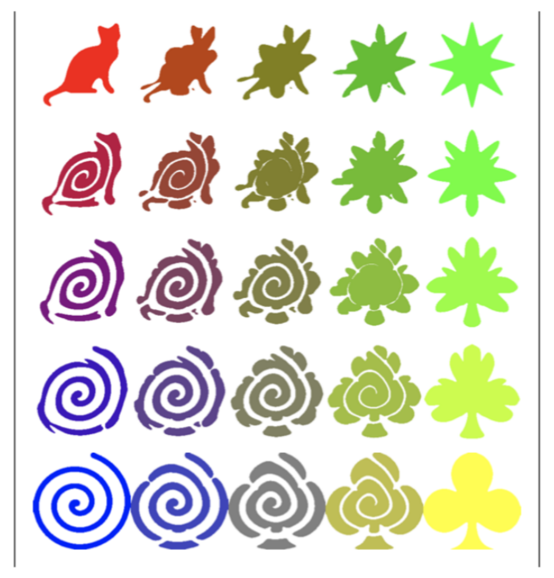
Generalize to arbitrary metric

- Formally a minimization of the weighted average distance:
  $S(f,\vec{\alpha},\vec{\alpha_1},\vec{\alpha_N}) = \sum_{i=1}^{N} w_i(\vec{\alpha},\vec{\alpha_1},\vec{\alpha_N})\left[D\left(f(x|\vec{\alpha}),f(x|\vec{\alpha_i}\right)\right]^p$
- $D(f(x),g(x))$ is a distance (metric functional in the space of distributions)
- Every metric generates an interpolation method (see Chap. 14 of *Encyclopedia of Distances*, Deza and Deza, 4ed., Springer, 2016)
- L^2 distance generates vertical morphing (with $p=2$, $[D(\cdot)]^p$ is the integrated squared error)
- Wasserstein distance generates horizontal morphing (p=1 Earth Mover distance)
  - $W_p(X,Y) := W_p(P_0,P_1) = \left(\int \parallel x - T^*(x) \parallel^p dP_0(x)\right)^{1/p}$, T^* optimal transport map
  - Works well in defining a metric in the space of almost all distributions
  - The set of distributions equipped with Wasserstein distance is a geodesic space (Riemaniann if $p=2$)
  - Given P_0 and P_1 there is always a shortest path (geodesic) between them, and its length is the Wasserstein distance $W(P_0,P_1)$



Graphics from Bonneel, Peyre, Cuturi, 2016

Optimal transport is quite powerful



Graphics from Peyre, Cuturi, 2019

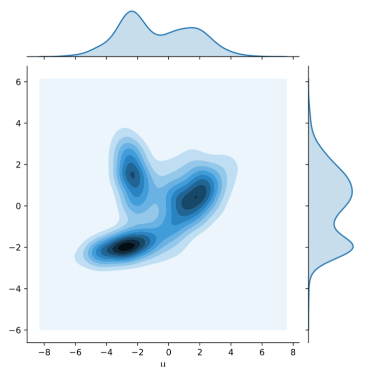
What if a transport map from P_0 to P_1 does not exist?

- Example: $P = \delta_0$ (point mass at 0), $Q = Gaussian$
- Kantorovich relaxation: take the mass at x and split it into small components
- $\mathcal{J}$ set of all joint distributions J for (X, Y) with marginals P and Q (*coupling* between P and Q)
- Find J to minimize $\mathbb{E}_J\left[\parallel X - Y \parallel\right] = \left(\int \parallel x - y \parallel^p dJ(x,y)\right)^{\frac{1}{p}}$
- Wasserstein distance: $W(P,Q) = W(X,Y) = \left(inf_J \int \parallel x - y \parallel^2 dJ(x,y)\right)^{\frac{1}{2}}$

- If an optimal transport T exists, then the optimal J is degenerate and supported on the curve $(x, T(x))$
- Regularization possible by adding term:

$$\mathbb{E}_J\left[\parallel X - Y \parallel\right] = \left(\int \parallel x - y \parallel^p dJ(x,y)\right)^{\frac{1}{p}} + \lambda f(J)$$

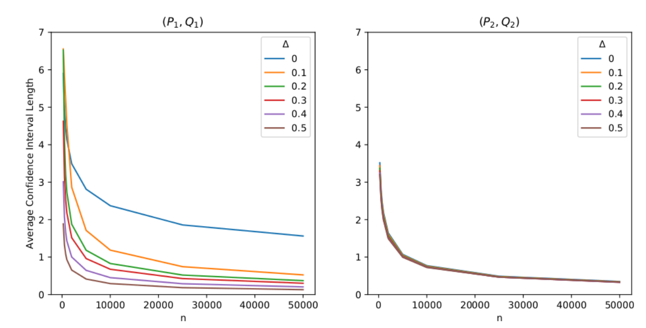
  - $f(J)$ e.g. entropy
  - Fast, and easier inference
  - How to choose λ? Not clear effect of regularization



Graphics from Wikipedia

Uncertainty quantification

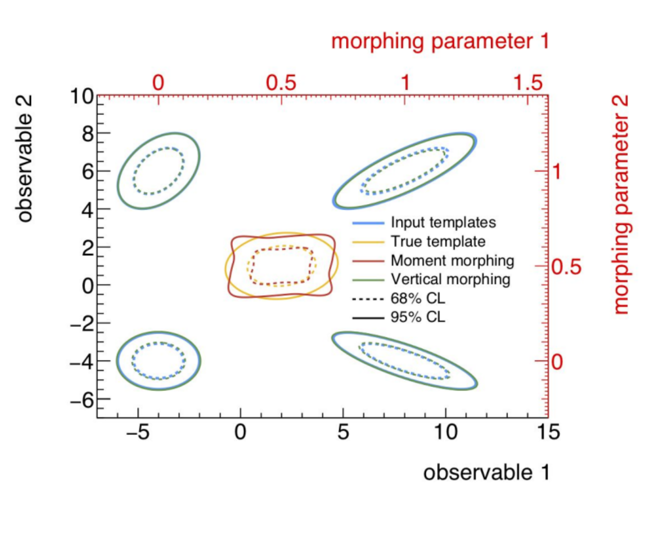
- These methods introduce an uncertainty in the morphed shape determination
- $\hat{T}$ estimate of T based on samples $X_1, ..., X_N \sim P_0$, $Y_1, ..., Y_N \sim P_1$
- Closeness of $\hat{T}$ to T ($\hat{W}(P_0, P_1)$ to $W(P_0, P_1)$) depends on number of dimensions
  $\mathbb{E} \int \| \hat{T}(x) - T(x) \|^2 \, dP_0(x) \approx (\frac{1}{N})^{\frac{1}{d}}$ (curse of dimensionality)
- Getting confidence intervals very hard, solved only for special cases
  - 1D (Munck, Czado, Sommerfeld)
  - MultiD: sliced Wassserstein distance (average W between 1D projections of P_0 and P_1)
  - Under this approximation (weaker metric), can derive confidence regions by a minimax game on the L^r norm of quantile functions of P_0 and P_1 for a fixed confidence level
  - Coverage guaranteed by construction



Graphics from arXiv:1909.07862. Here P_0 is P and P_1 is Q, indices refer to two example cases, $n = 100$

Moment morphing

- Moment morphing: morph standardized densities instead of densities
  - Useful for models with well-behaved first moments (mean and variance)
  - Not as good as horizontal morphing in 1D (inefficient version of it), good approximation in N
  - How to morph the covariance matrix? Many choices available



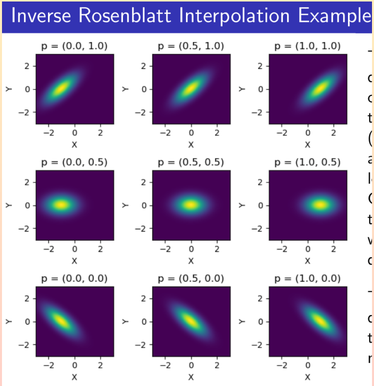
Graphics from Lydia Brenner

- Devise a multi-D equivalent of quantile function: the *Inverse Rosenblatt transformation* (Ann. Math. Statist. 23, 470 (1952).
- The inverse Rosenblatt transformation $x_1 = F_1^{-1}(z_1), x_2 = F_2^{-1}(z_2|z_1)$ uses conditional quantile functions: we know how to interpolate them!
- Computationally intensive (k non-linear equations to be solved numerically, N calls to root-finding, etc)

Let $X = (X_1, \cdots, X_k)$ be a random vector with distribution function $F(x_1, \cdots, x_k)$. Let $z = (z_1, \cdots, z_k) = Tx = T(x_1, \cdots, x_k)$, where T is the transformation considered. Then T is given by

$$z_1 = P\{X_1 \leqq x_1\} = F_1(x_1),$$
$$z_2 = P\{X_2 \leqq x_2 \mid X_1 = x_1\} = F_2(x_2 \mid x_1),$$
$$\vdots$$
$$z_k = P\{X_k \leqq x_k \mid X_{k-1} = x_{k-1}, \cdots, X_1 = x_1\} = F_k(x_k \mid x_{k-1}, \cdots, x_1).$$

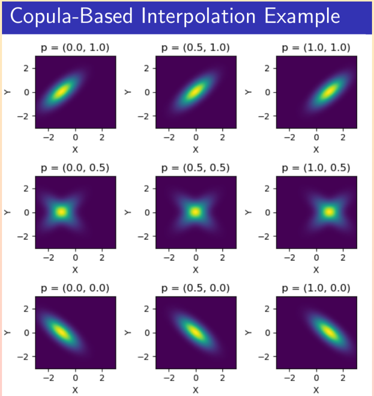
One can readily show that the random vector $Z = TX$ is uniformly distributed over the k-dimensional unit cube.



Graphics by Igor Volobouev

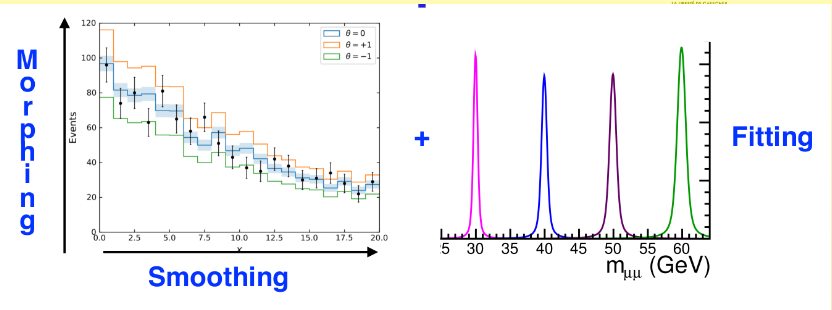
Copula morphing

- Probability integral transforms of marginals of $f(\vec{x})$: $z_1 = F_1(x_1)$,... $z_k = F_k(x_k)$
- Copula density $c(\vec{z})$ is density of the vector of z_k, captures mutual information (and $c(\vec{z})$ uniform if and only if all X_i independent)
- Given the marginal densities $f_i(x) = \frac{dF_i(x)}{dx}$, then $f(\vec{x} = c(F_1(x_1), ..., F_k(x_k)) \prod_{i=1}^{k} f_i(x_i)$
- Now do horizontal morphing on the marginals separately in each variable, then interpolate vertically the copula density
- Much faster than Inverse Rosenblatt transformation
- Results intuitively more "reasonable"

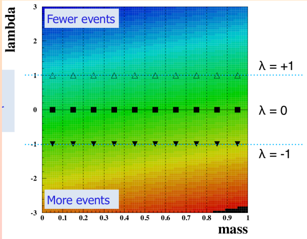


Graphics by Igor Volobouev

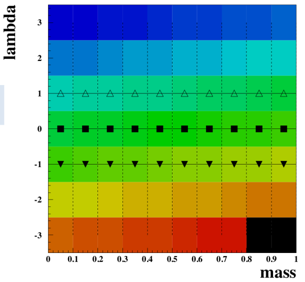
How we tend to call things in CMS



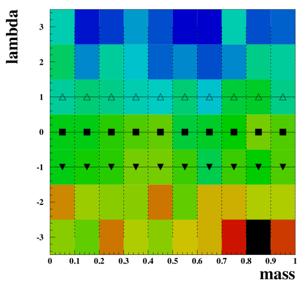
Analytic knowledge on λ, m



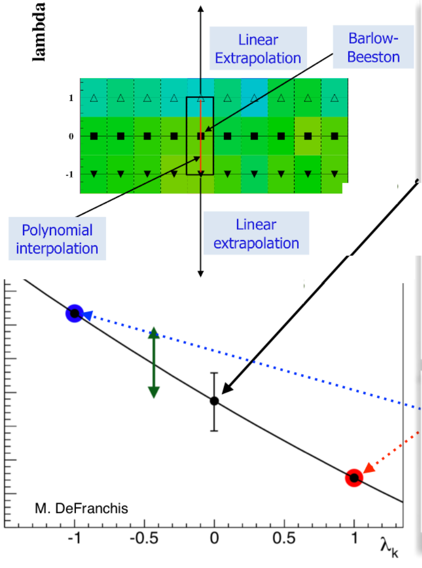
Discretized knowledge on λ, m



Statistical fluctuations



Accounting for various effects: statistical fluctuations



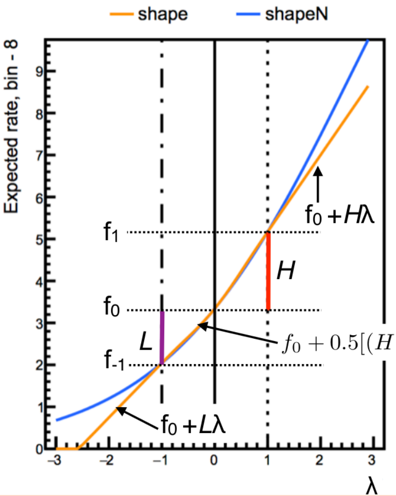
Statistical uncertainty of nominal templates taken into account in Poisson based template fits to data

- ‘Barlow Beeston': one additional nuisance par per contributing template J. Barlow, C. Beeston, CPC 77 (1993) 219-228
- ’Barlow Beeston lite’: one additional nuisance parameter for templates sum →Standard Procedure in CMS
  John Conway, arXiv1103.0354

Statistical uncertainty of ± 1sigma Templates usually neglected → can lead to fake constraints for λ, see https://indico.cern.ch/event/761804/contributions/3160985/attachments/1733339/2802398/Defranchis_template_constraints.pdf

8

Slide by Olaf Behnke



https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/

Shape morphing:

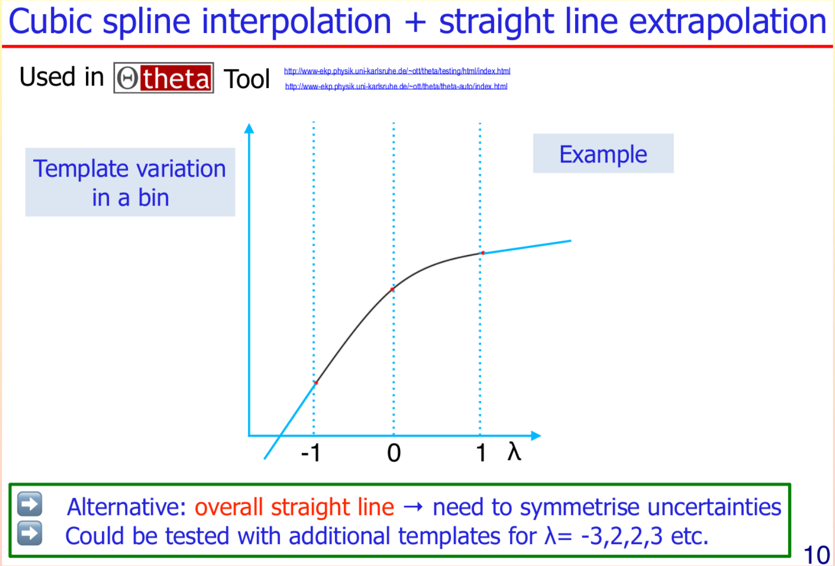
- First normalise templates dividing by the sum over all horizontal (e.g. mass) bins $\rightarrow$ obtain fractions
- shape: morph bin-wise fractions vs λ
- shapeN: morph log of bin-wise fractions vs λ

Interpolation for -1< λ<1:

$$f_0 + 0.5[(H-L)\lambda + 1/8(H+L)[3\lambda^6 - 10\lambda^4 + 15\lambda^2]]$$

Fulfills f''=0 for λ=-1 and λ=1

Slide by Olaf Behnke

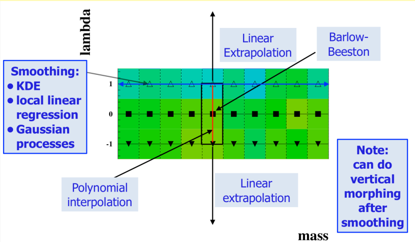


Slide by Olaf Behnke

Horizontal smoothing

UCLouvain
Institut de recherche
en mathématique et physique

- Horizontal smoothing with well-established methods in literature
- Kernel-based methods depend on choice of bandwith
  - Discussed in detail last week (Nick McColl)
- Local linear regression depends on locality window



Kernel Density Estimation (KDE)

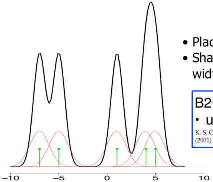
Material © Chad Shafer: https://indico.desy.de/indico/event/22731/session/7/contribution/15/material/0/0.pdf

- Sample n independent points X_i from unknown distribution f
- KDE estimate:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

- Example: Gaussian Kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



- Places a smoothed-out lump over each data
- Shape of 'lumps' is controlled by K(.); their width controlled by h

B2G-18-008:
- use adaptive width $h \sim 1/\text{sqrt}(f(x))$

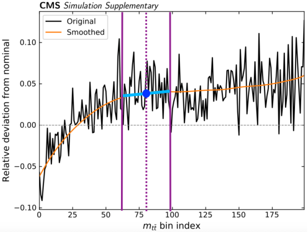
K. S. Cranmer, "Kernel estimation in high-energy physics", *Comput. Phys. Commun.* 136 (2001) 198, doi:10.1016/S0010-4655(00)00243-5, arXiv:hep-ex/0011057.

13

Local Linear Regression (LOWESS)

CMS HIG-17-027

See talk A. Popov https://indico.in2p3.fr/event/19290/contributions/75880



- Use points in sliding window
- Give points near centre larger weights
- Fit straight line
- Move window →
- Connect fitted window centre points

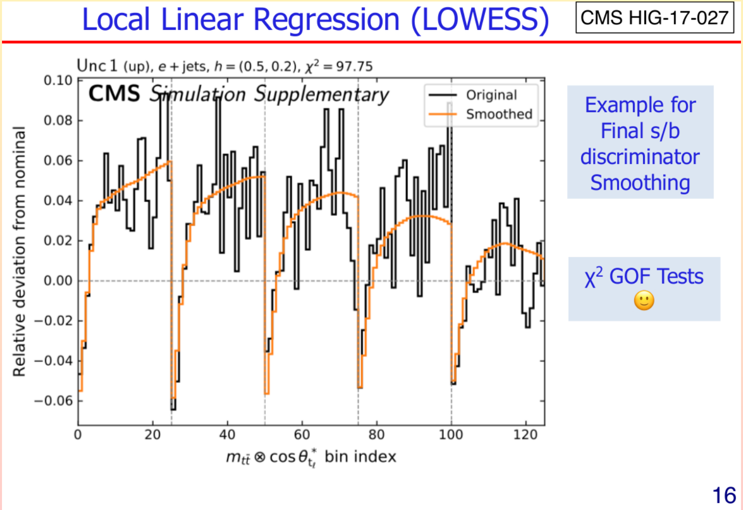
Optimise hyper-pars with cross-validation

15

Slide by Olaf Behnke

Smoothing and Goodness-of-Fit tests

- To compare the smoothed and unsmoothed templates it's tempting to use χ^2
- However, χ^2 not well defined; by construction, smoothing alters number of degrees of freedom
- You have first to treat your smoothing method as a linear filter, and calculate NDoF (in KDE, related to autocorrelation of the kernels used)
  - Somehow related to time series analysis: reduction of NDoF
  - There is literature on this, we can put it in twiki; in the meantime, ask Igor Volobouev ☺

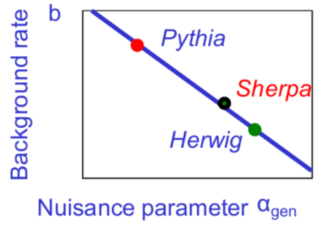


Slide by Olaf Behnke

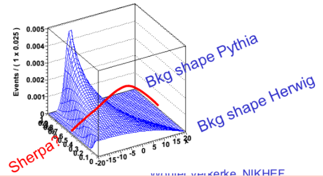
Caveats on modelling theory uncertainties (P.V. at Benasque 2018)

- Cross section uncertainty: easy, assuming a gaussian for the constraint term
  $\mathcal{L}_{full}(s, b) = \mathcal{P}(N_{SR}|s + b) \times \mathcal{G}(\tilde{b}|b, 0.08)$
- Factorization scale: what distribution $\mathcal{F}$ is meant to model the constraint???
  $\mathcal{L}_{full}(s, b) = \mathcal{P}(N_{SR}|s + b(\alpha_{FS}) \times \mathcal{F}(\tilde{\alpha_{FS}}|\alpha_{FS})$
  - "Easy" case, there is a single parameter α_{FS}, clearly connected to the underlying physics model
- Hadronization/fragmentation model: run different generators, observing different results
  - Difficult! Not just one parameter, how do you model it in the likelihood?
  - 2-point systematics: you can evaluate two (three, four...) configurations, but underlying reason for difference unclear
  - Often define empirical response function

- Counting experiment: easy extend to other generators
- There must exist a value of α corresponding to SHERPA

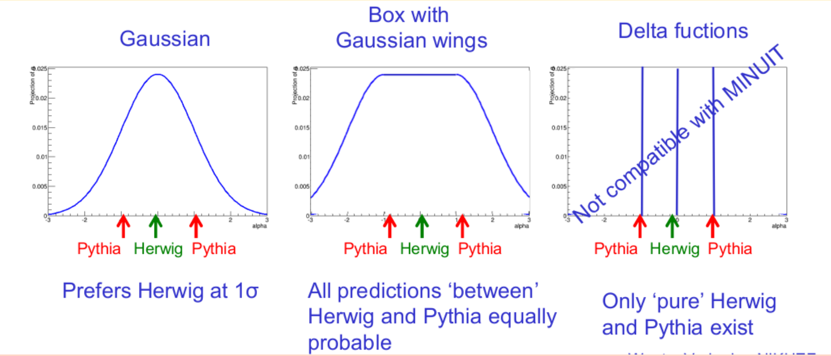


- Shape experiment: ouch!
- SHERPA is in general not obtainable as an interpolation of PYTHIA and HERWIG



Graphics from W. Verkerke

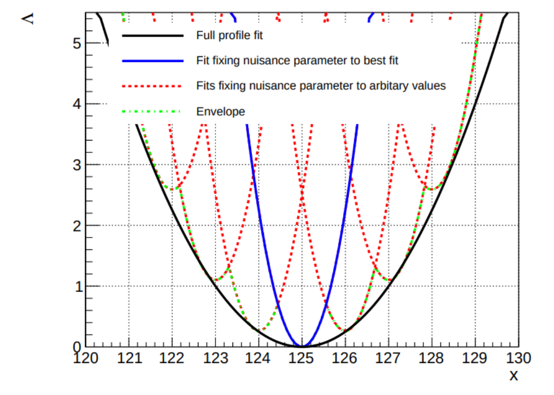
- Attempting to quantify our knowledge of the models
- There is no single parameter, difficult to model the differences within a single underlying model
- Which of these is the “correct” one?



Graphics from W. Verkerke

Solving the delta functions issue: discrete profiling

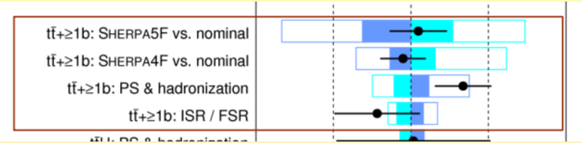
- Label each shape with an integer, and use the integer as nuisance parameter
- Can obtain the original log-likelihood as an envelope of different fixed discrete nuisance parameter values
- How do you define the various shapes?
  - Need many additional generators!
  - Interpolation unlikely to work (*SHERPA is not midway between PYTHIA and POWHEG*)

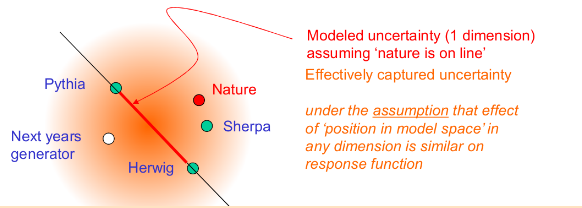


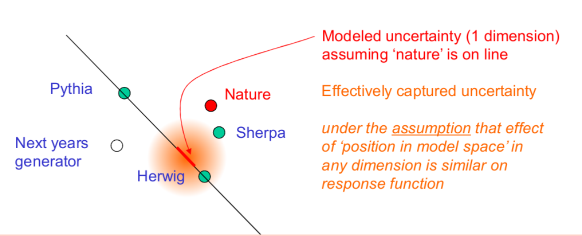
From arXiv:1408.6865

The issue of over-constraining

- How to interpret constraints?
- **Not as measurements**
- Correlations in the fit make interpretation complicated
- Avoid statements when profiling as a nuisance parameter







Graphics from ATLAS and W. Verkerke, as far as I remember

UCLouvain
Institut de recherche
en mathématique et physique

- Closure tests are alternative procedures you can use to check if your measurement is robust
  - E.g. insensitive to systematic effects
  - Usually compare alternative result with nominal result (GoF test) to decide if closure test passed
- **Closure tests are PASS/FAIL tests**
- Correct course of action: if closure test fails, then there is a mistake in the tested procedure, therefore modify/improve the procedre
  - If the alternative procedure highlights e.g. a recalibration to be done, then recalibrate (i.e. use the better procedure)
- Wrong course of action: if closure test fails, add discrepancy as uncertainty
  - The sentence *"The closure test shows a 10% discrepancy, and we consequently assign it as systematic uncertainty"* is pure BS (although you'll sadly find it in many published papers)
- In general, if a closure test fails, always prioritize a mitigation or suppression of the effect by improveming your analysis methods
  - A systematic should be added only as a very very last resort

UCLouvain
Institut de recherche
en mathématique et physique

- numpy
- matplotlib
- mpl_toolkits
- inspect
- iminuit
- pyhf
- scipy
- statsmodels
- itertools
- pandas
- statistics

- Frederick James: Statistical Methods in Experimental Physics - 2nd Edition, World Scientific
- Glen Cowan: Statistical Data Analysis - Oxford Science Publications
- Louis Lyons: Statistics for Nuclear And Particle Physicists - Cambridge University Press
- Louis Lyons: A Practical Guide to Data Analysis for Physical Science Students - Cambridge University Press
- E.T. Jaynes: Probability Theory - Cambridge University Press 2004
- Annis?, Stuard, Ord, Arnold: Kendall's Advanced Theory Of Statistics I and II
- Pearl, Judea: Causal inference in Statistics, a Primer - Wiley
- R.J.Barlow: A Guide to the Use of Statistical Methods in the Physical Sciences - Wiley
- Kyle Cranmer: Lessons at HCP Summer School 2015
- Kyle Cranmer: Practical Statistics for the LHC - http://arxiv.org/abs/1503.07622
- Roberto Trotta: Bayesian Methods in Cosmology - https://arxiv.org/abs/1701.01467
- Harrison Prosper: Practical Statistics for LHC Physicists - CERN Academic Training Lectures, 2015 https://indico.cern.ch/category/72/
- Christian P. Robert: The Bayesian Choice - Springer
- Sir Harold Jeffreys: Theory of Probability (3rd edition) - Clarendon Press
- Harald Crámer: Mathematical Methods of Statistics - Princeton University Press 1957 edition

UCLouvain
Institut de recherche
en mathématique et physique

Backup