# 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, 07–11 December 2020

# Program for today

UCLouvain
Institut de recherche
en mathématique et physique

## Measuring differential distributions

Unfolding

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

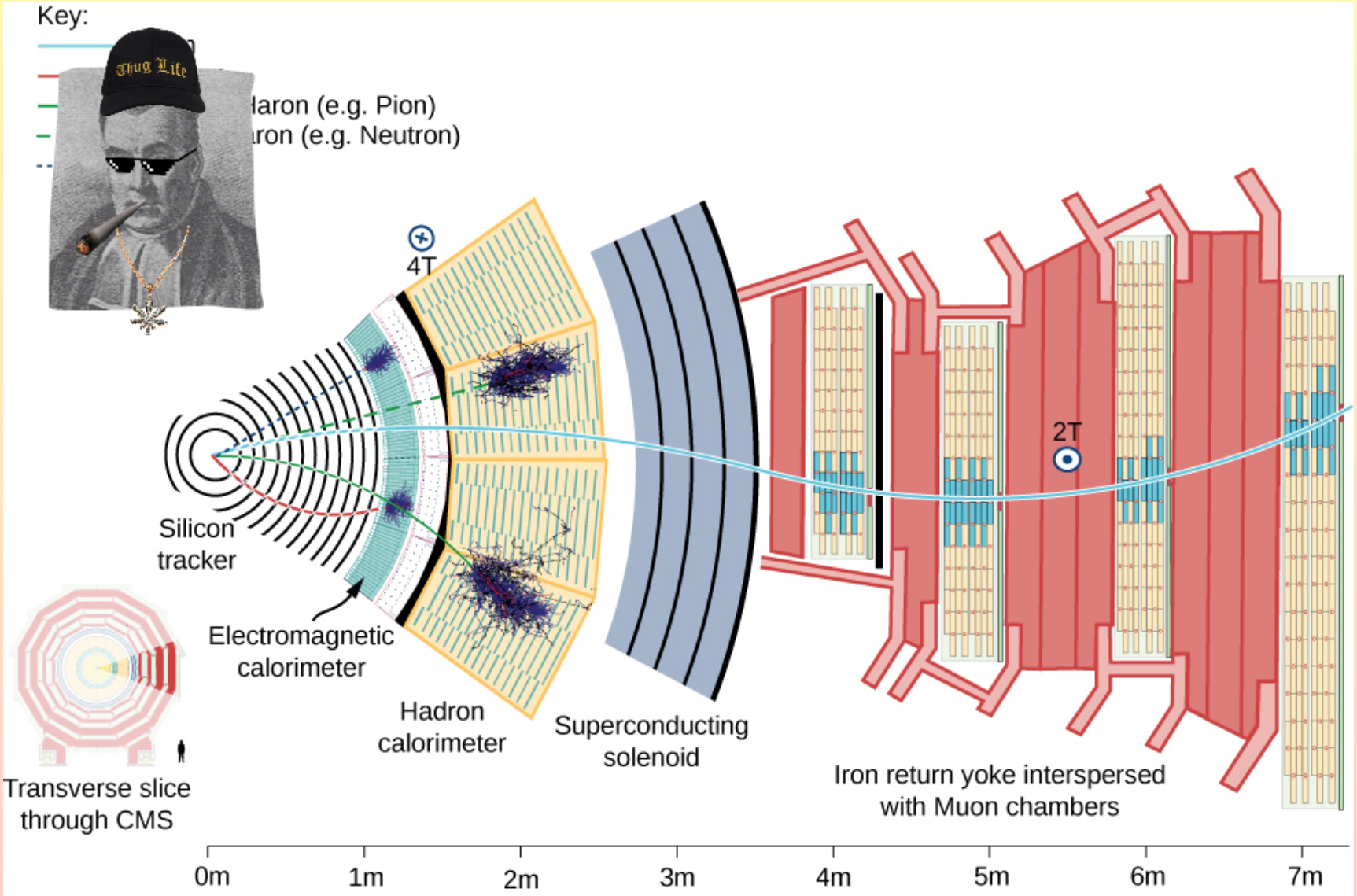
- **Lesson 1 - Fundaments**
  - Bayesian and frequentist probability, theory of measure, correlation and causality, distributions
- **Lesson 2 - Point and Interval estimation**
  - Maximum likelihood methods, confidence intervals, most probable values, credible intervals
- **Lesson 3 - Advanced interval 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

UCLouvain
Institut de recherche
en mathématique et physique

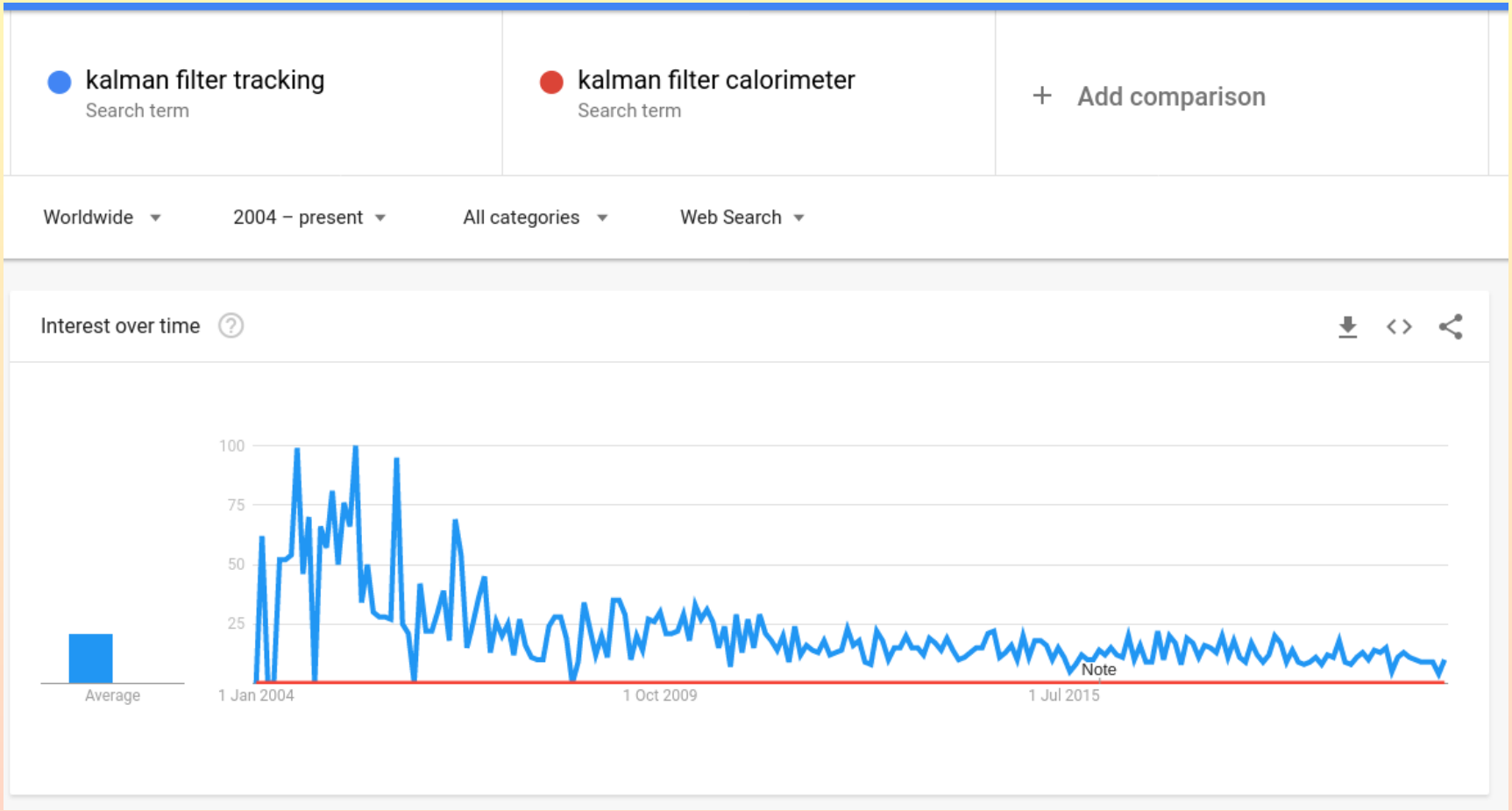
# Commonly-used methods in particle physics

UCLouvain
Institut de recherche
en mathématique et physique

- Question Time!



# Hidden Bayes: The Kalman Filter!



- You have an unknown true trajectory whole evolution is governed by discrete stochastic time steps
- You want to predict the next state $x \in \mathbb{R}^n$ given the previous one
  $x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1}$
- You want to use measurements $z \in \mathbb{R}^n$ to inform your decision $z_k = Hx_k + \nu_k$
- Terms
  - $w_k \sim Gaus(0, Q)$ and $\nu_k \sim Gaus(0, R)$ represent the noise (eventually time dependent) on the process and on the measurement
  - $A$ controls the evolution of $x$ in absence of noise or of driving function $B$
  - $B$ models an optional control input $u$
  - $H$ relates the state to the measurement $z_k$

- Define a priori and a posteriori estimate errors (minus sign in apex = a priori) and cov matrices
  - $e_k^- = x_k - \hat{x}_k^-$ $\quad P_k^- = E[e_k^- e_k^{-T}$
  - $e_k = x_k - \hat{x}_k$ $\quad P_k = E[e_k - e_k^T$
- Compute an a posteriori state estimate $\hat{x}_k$ as a linear combination
  - An a priori estimate $\hat{x}_k^-$ and
  - a weighted difference between $z_k$ (actual measurement) and a prediction $H\hat{x}_k^-$
- Equation: $\hat{x}_k = \hat{x}_k^- + K(z_k - H\hat{x}_k^-)$
  - $(z_k - H\hat{x}_k^-)$ is the *innovation* (or residual)
  - reflect discrepancy between the predicted and measured state
  - $K$ gain, or blending factor minimizing the a posteriori covariance

- The minimization of $K$ is done by inserting the $e_k$ into the equation, computing expectations, deriving w.r.t. $K$ in zero, etc.
- The minimizing $K_k$ is:

$$K_k = \frac{P_k^- H^T}{H P_h^- H^T - R}$$

- $\lim_{R_k \to 0} K_k = H^{-1}$, for smaller error covariances the gain $K$ weights the residual more heavily
- $\lim_{P_k \to 0} K_k = 0$, for smaller a priori estimated covariances, the gain $K$ weights the residual less heavily
- As the measurement error covariance (a priori estimate error covariance) approaches zero, you trust
  - More and more (less and less) your measurement $z_k$
  - Less and less (more and more) your prediction $Hx^-$

- At each time step, time is updated by projecting the state and covariance estimates
  - $x_k^- = Ax_{k-1} + Bu_{k-1}$
  - $P_k^- = AP_{k-1}A^T + Q$
- At each time step, measurement is updated
  - Compute Kalman gain: $K_k = P_k^- H^T (HP_k^- H^T + R)^{-1}$
  - Measure process (obtain $z_k$) and generate a posteriori estimate: $\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-)$
  - Update the a posteriori error matrix $P_k = (I - K_k H)P_k^-$
- When the noises $Q$ and $R$ are constant, both $P_k$ and $K_k$ will stabilize quickly and remain constant
  - And in this case you can even pre-compute them offline
- The Kalman filter uses only the last state of the system to predict the next one
  - Other algorithms such as Wiener one use all the previous estimates to compute the state, computational burden

UCLouvain
Institut de recherche
en mathématique et physique

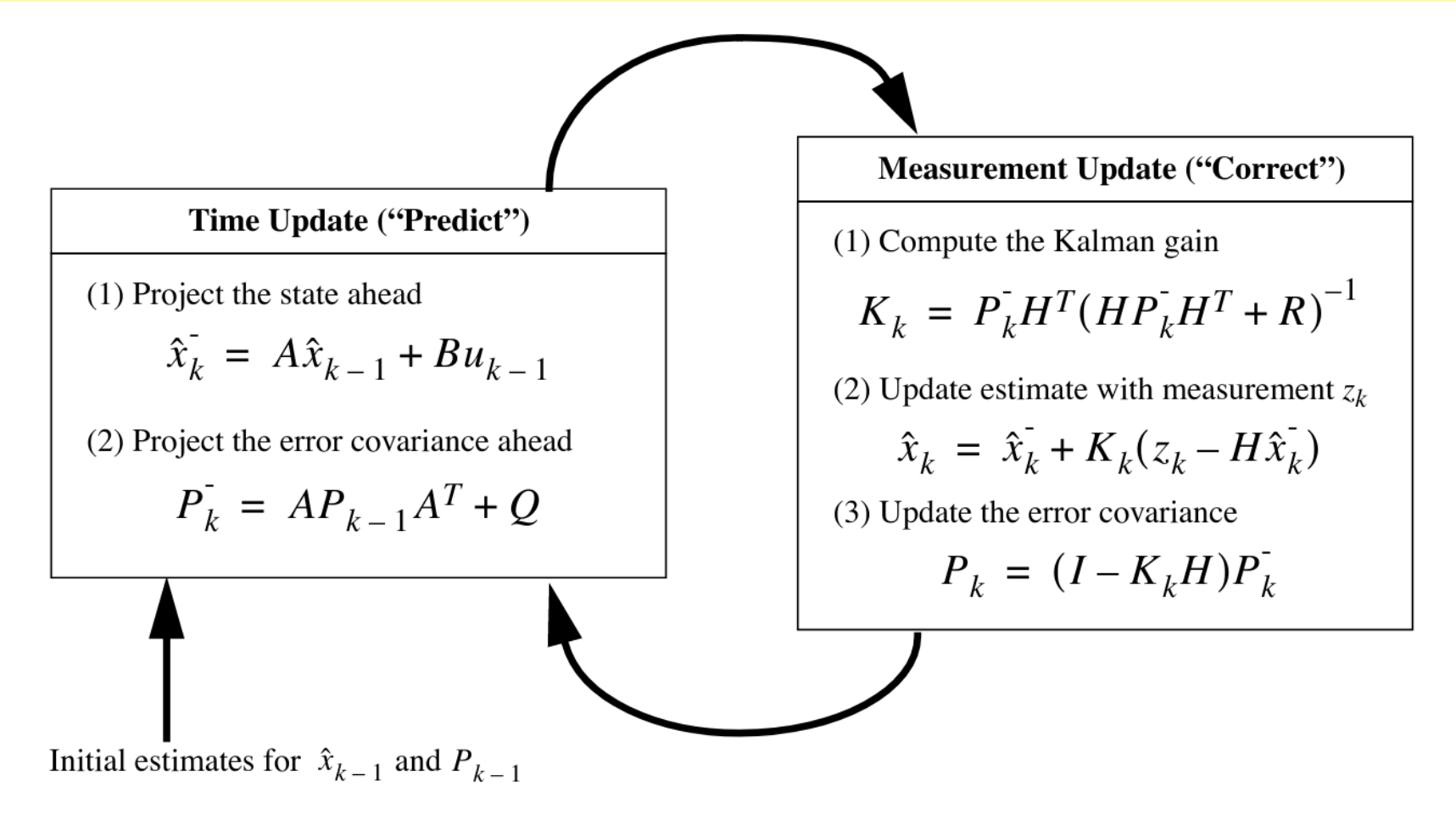



Image from Welch and Bishop, 2006

- If the time or measurement evolution equations are nonlinear, *extended Kalman filter*
  - Linearize around the current estimate using partial derivatives (similar to Taylor expansion)
- This is the form in which it's used for track or vertex reconstruction

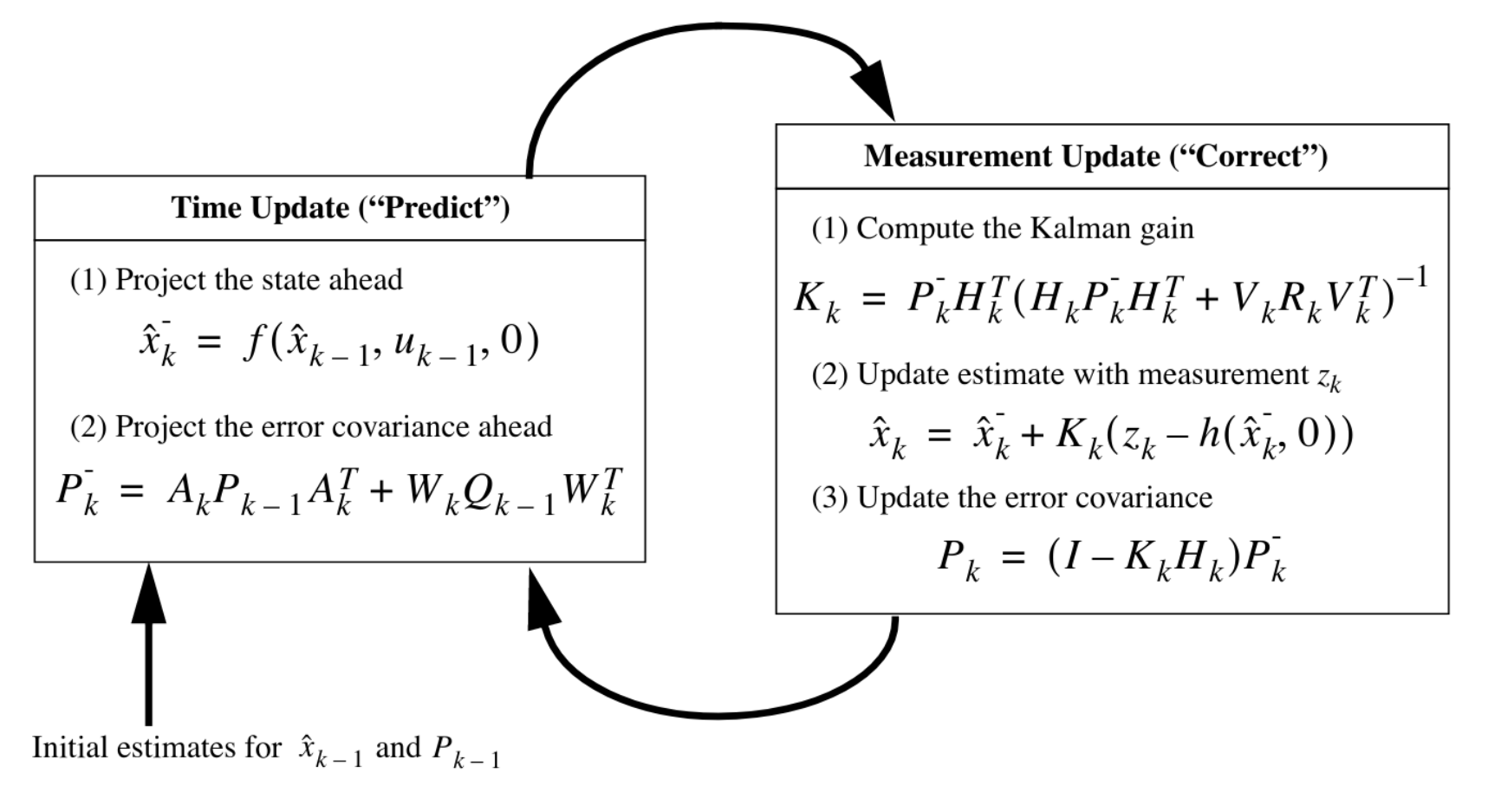



Image from Welch and Bishop, 2006

- This afternoon: code your own Kalman filter
  - If you want animations, make sure you can `import threading`

# Estimate your background with ABCD

- A “simple” version of a background determination in sidebands used since decades: ABCD
  - Assume that the signal strength in B,C,D is so small as to be negligible
- On Tuesday we have seen how to model background measurements in the likelihood function
  - Likelihood approach accounts for signal contamination where basic ABCD assumes zero signal contamination

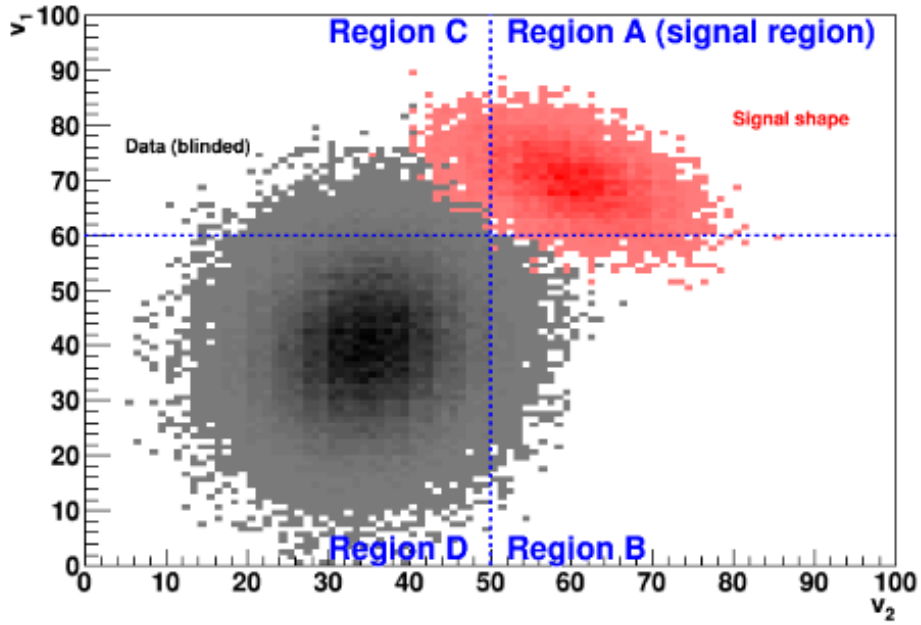



Figure 1: Illustration of a canonical example where the ABCD backgrond estimation method is to be used. The black distribution is the collected data, and the red distribution is the hypothesized shape of the signal that the analysis is targetting.

Image from Buttinger, 2018

- Assume that it's true that

$$\frac{N_C^{bkg}}{N_D^{bkg}} = \frac{N_A^{bkg}}{N_A^{bkg}}$$

- This may be approximately satisfied if the observables defining the regions are sufficiently uncorrelated for background
  - Any correlation in the signal distribution is not relevant to the approach
- If we knew the signal strength in a reliable way, can subtract it in each region to get

$$N_i^{bkg} = N_i - N_i^{sig}, \qquad i = B, C, D$$

- But we don't, and further assume $N_i^{bkg} \simeq N_i, \qquad i = B, C, D$
- We can then estimate background in signal region $A$

$$N_A^{bkg} = \frac{N_C^{bkg}}{N_D^{bkg}} N_B^{bkg} = \frac{N_C}{N_D} N_B$$

- Uncertainties in $N_A^{bkg}$
  - Statistical: standard Poisson on the nominal prediction
  - Systematic: usual uncertainty propagation for the statistical uncertainties in $N_C$, $N_D$, $N_B$

# Validate ABCD

- To validate ABCD, you must have additional regions
- Split B and D
- Apply ABCD to A' B' C' D' to estimate $N_{A'}^{bkg}$, compare with $N_{A'}$

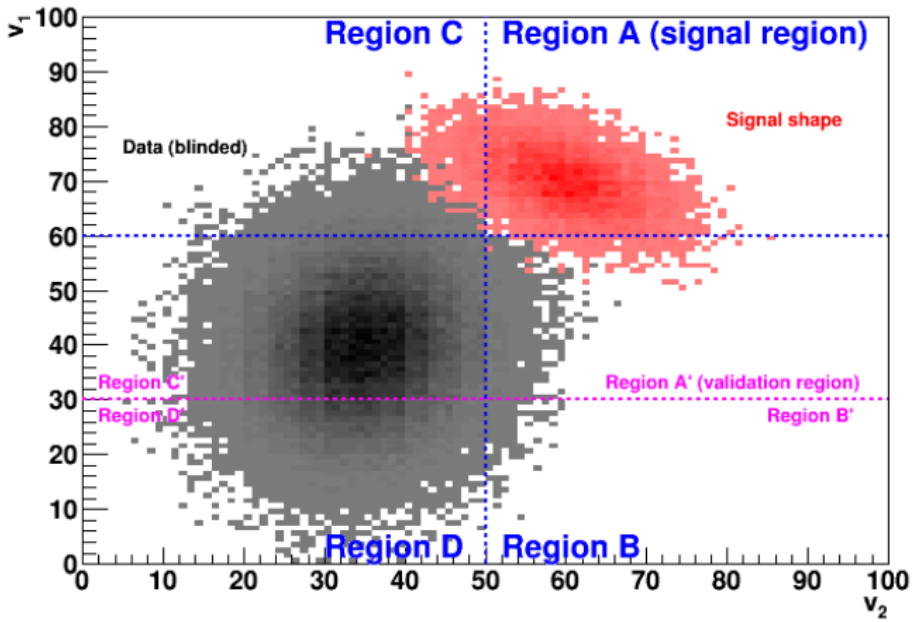



Figure 2: Region B and D of the ABCD plane can be cut into two, to define new regions B',A', D', and C'. Region B' can be used as a validation region in which to test the ABCD method.

Image from Buttinger, 2018

# What if validation horribly fails?

- Crude yet popular approach: take the relative difference between $N_{A'}^{bkg}$ and $N_{A'}$ as an uncertainty in the background estimate
  - BAD PRACTICE (see next-to-next slide) that we had already covered on Monday
  - Induces other issues: must validate the new closure (find new regions) until some closure closes

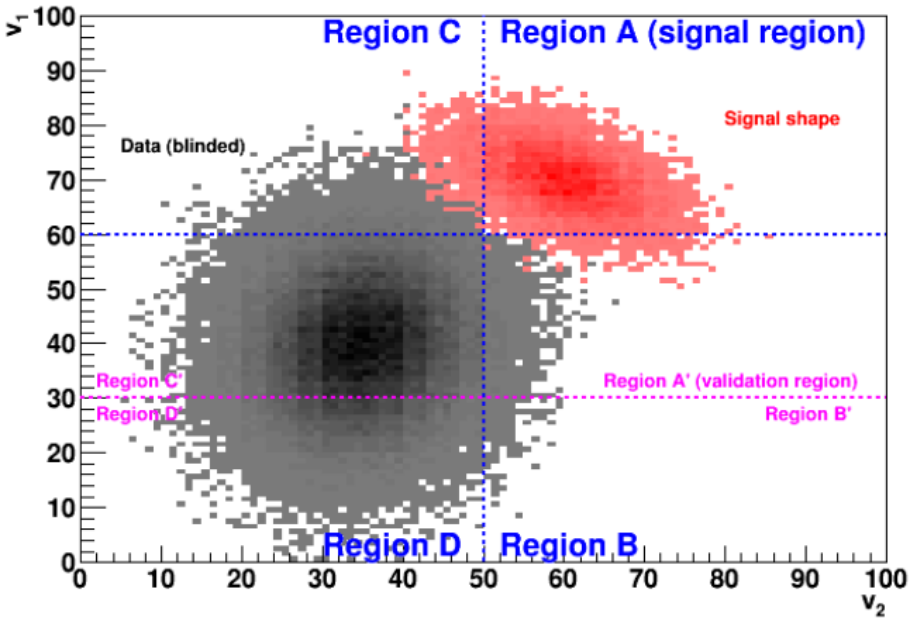



Figure 2: Region B and D of the ABCD plane can be cut into two, to define new regions B',A', D', and C'. Region B' can be used as a validation region in which to test the ABCD method.

Image from Buttinger, 2018

# What if validation horribly fails?

- Better: improve your estimate
  - Build sliding windows in an approximately continuous variable (here, $v_2$)
  - If there is any, forget about it

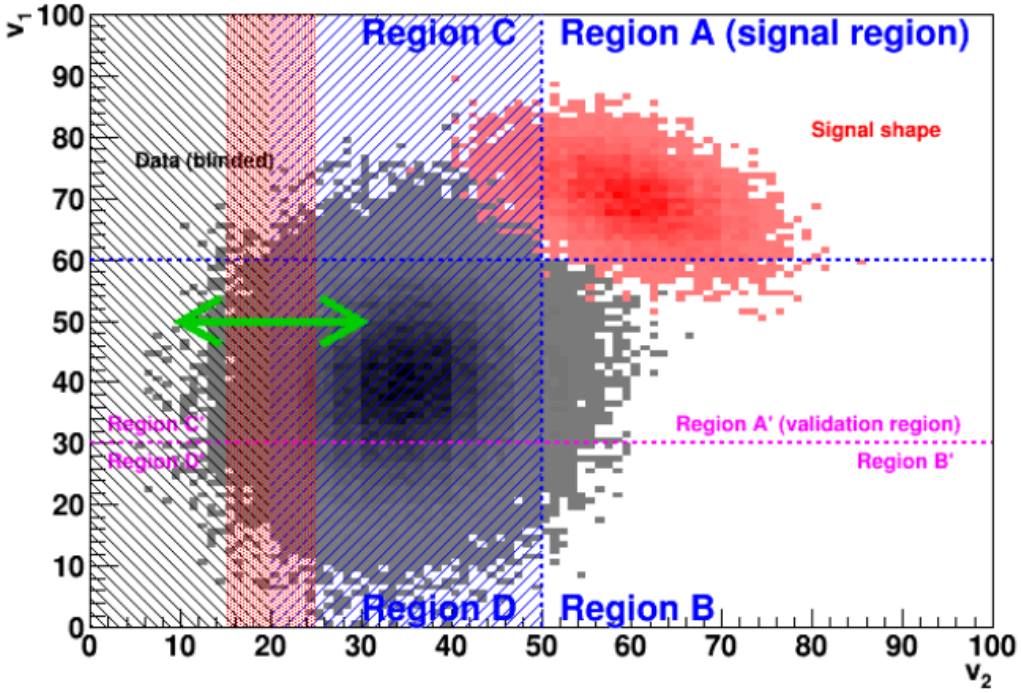
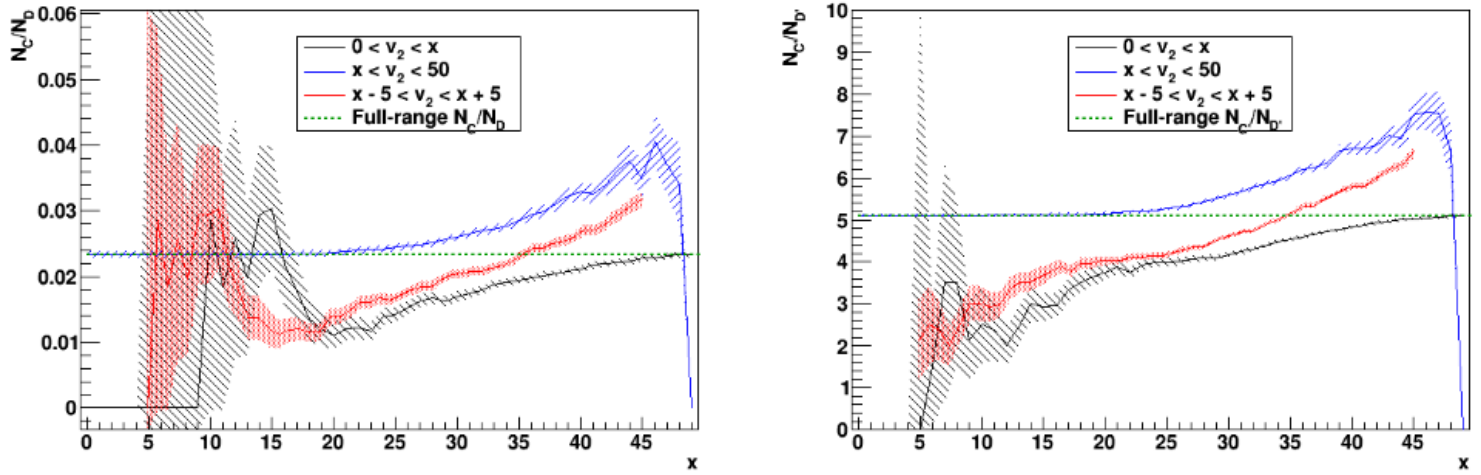



Figure 3: The ratio $N_C/N_D$ (and $N_{C'}/N_{D'}$) can be measured in subregions of region $C$ and $D$ ($C'$ and $D'$) when the $v_2$ variable is sufficiently *continuous*. Making this measurement may show evidence of a trend in the data that is inconsistent with the base assumption that $N_C/N_D = N_B/N_A$ for background events. The shaded black region corresponds to $0 < v_2 < x$, shaded blue is $x < v_2 < 50$ and shaded red is $x - 5 < v_2 < x + 5$, where in all cases $x = 20$. The ratios can be measured as a function of $x$, by varying the subregions as indicated by the green arrows.

Image from Buttinger, 2018

# What if validation horribly fails?

- Better: improve your estimate
  - Build sliding windows in an approximately continuous variable (here, $v_2$)
  - Calculate transfer factors $N_C/N_D$ and $N'_C/N'_D$ as a factor of an approximately continuous variable (here, $v_2$)
  - Correct estimates using the transfer factors (propagating uncertainty in the transfer factors)
- Even better: implement ABCD in your likelihood function (equivalent but automatic propagation of all uncertainties and their correlations)




(a) Regions C and D

(b) Regions C' and D'

Figure 4: The ratios $N_C/N_D$ and $N_{C'}/N_{D'}$ measured as a function of $x$ for three different definitions of sub-regions of $C$, $D$, $D'$, and $C'$. See figure 3 for further explanation of $x$. The dashed green lines indicate the ratio obtained from the full regions.

Image from Buttinger, 2018

- 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

- Overview of the method itself (Pivk, Le Diberder 2004: *sPlot: a Statistical Tool to Unfold Data Distributions*)
  - https://doi.org/10.1016/j.nima.2005.08.106 (or https://sci-hub.tw/10.1016/j.nima.2005.08.106)
  - ROOT implementation at https://root.cern.ch/doc/master/classRooStats_1_1SPlot.html
- Summary of recent discussions and criticism on the method itself and on its practical application
  - Discussions in the Statistics Committee of various experiments
  - Also discussions among experiments
  - Main drawbacks identified by Igor Volobouev and Francisco Matorras
- A simpler alternative formalized by Louis Lyons
  - Labelled *AtoZ*, but in use since a couple decades in various experiments

## sPlot in a nutshell

- Explore a data sample constituted by two or several sources of events
- Assume the events are characterized by two set of variables
  - *Discriminating variables*: the distribution of all the sources of events is known
  - *Control variables*: the distribution of some sources of events is either unknown or considered as such
- "Reconstruct the distribution of the control variable, independently for each of the various sources of events, without making use of any a priori knowledge on this variable"
- "use the knowledge available for the discriminating variable to be able to infer the behavior of the individual sources of events with respect to the control variable."
  - Usually using a MLE estimate using the shape of the discriminating variable
- "An essential assumption for the s Plot technique to apply is that the control variable is uncorrelated with the discriminating variable."
  - First problem: you should and substitute *(in)dependent of* rather than *(un)correlated with*

Literal quotes from https://doi.org/10.1016/j.nima.2005.08.106

- Two (independent?) variables, with different mixtures

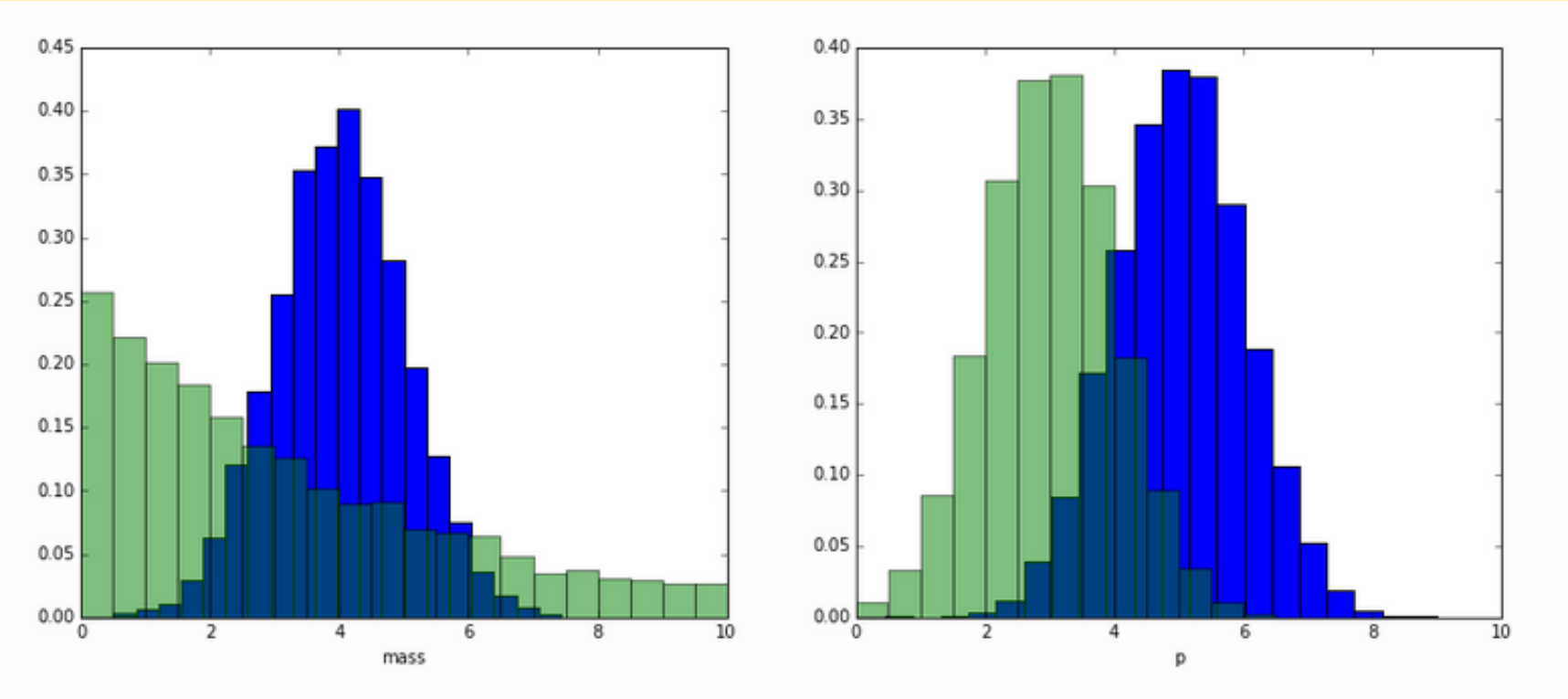



Image by agorozhnikov

- Actually observe the mixture (unknown labels)

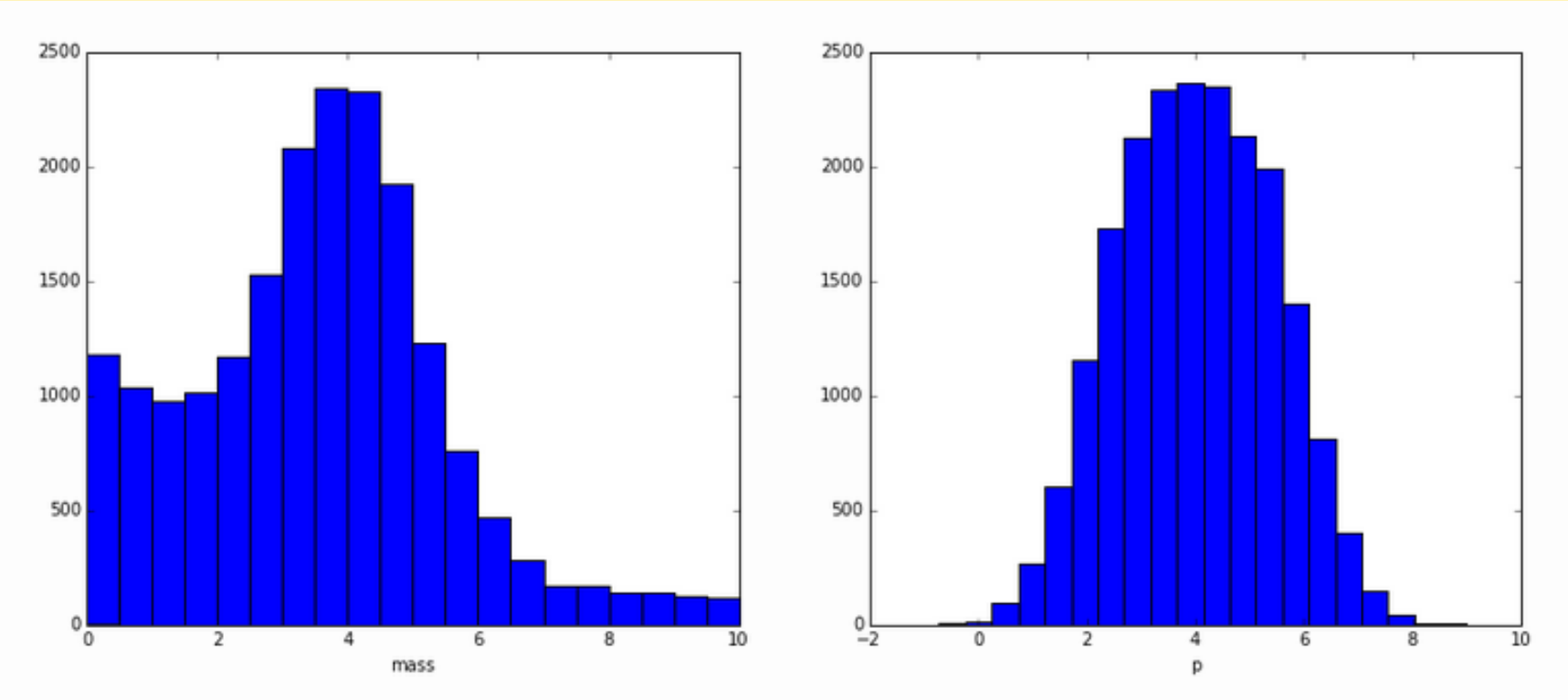



Image by agorozhnikov

UCLouvain
Institut de recherche
en mathématique et physique

- If you know the distribution of each class, can fit to obtain probabilistic estimate of the per-bin fraction

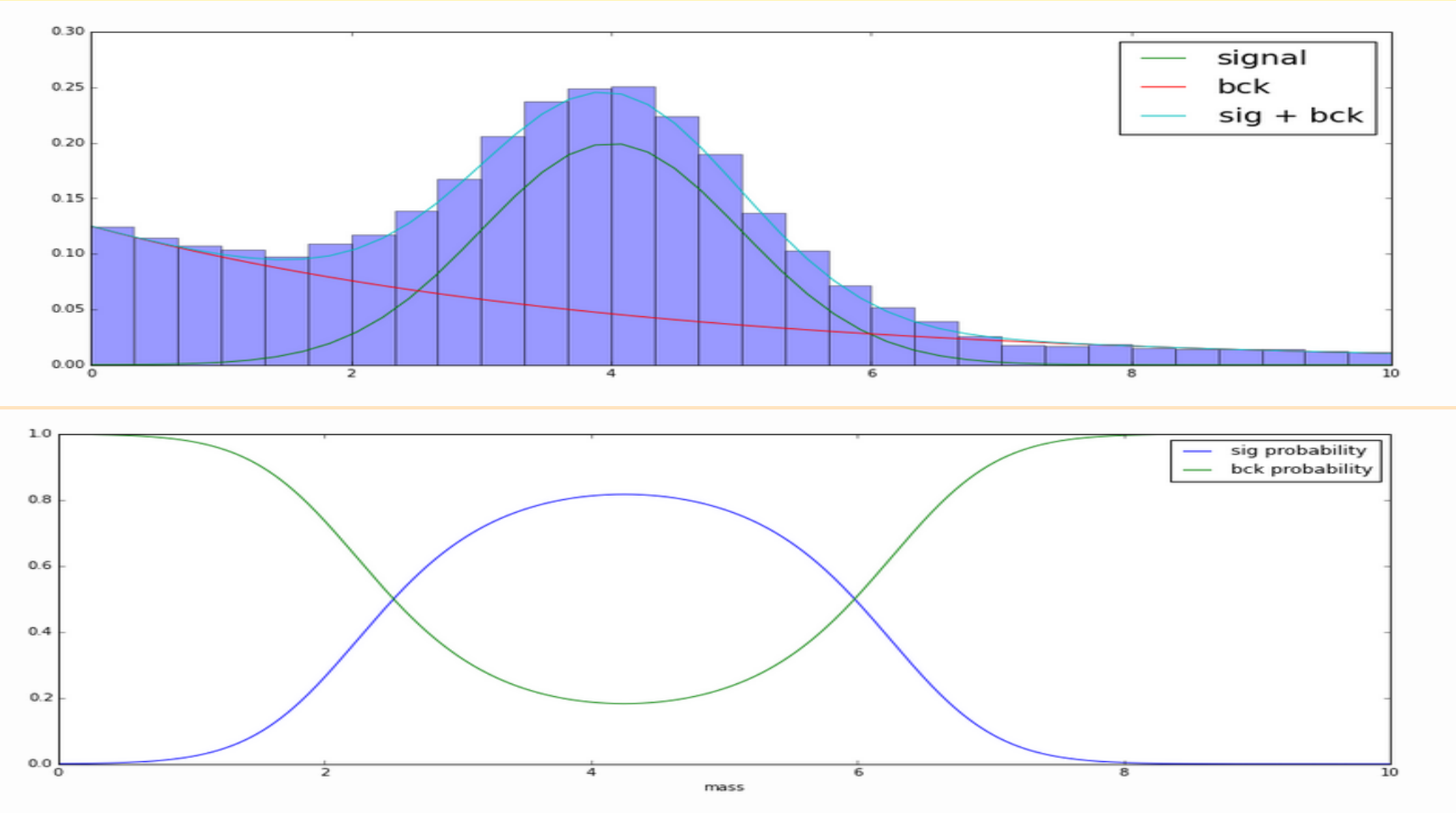



Image by agorozhnikov

UCLouvain
Institut de recherche
en mathématique et physique

- Convert the probabilities (above) to weights to be applied to each event (below) to compensate for the contribution of the other classes

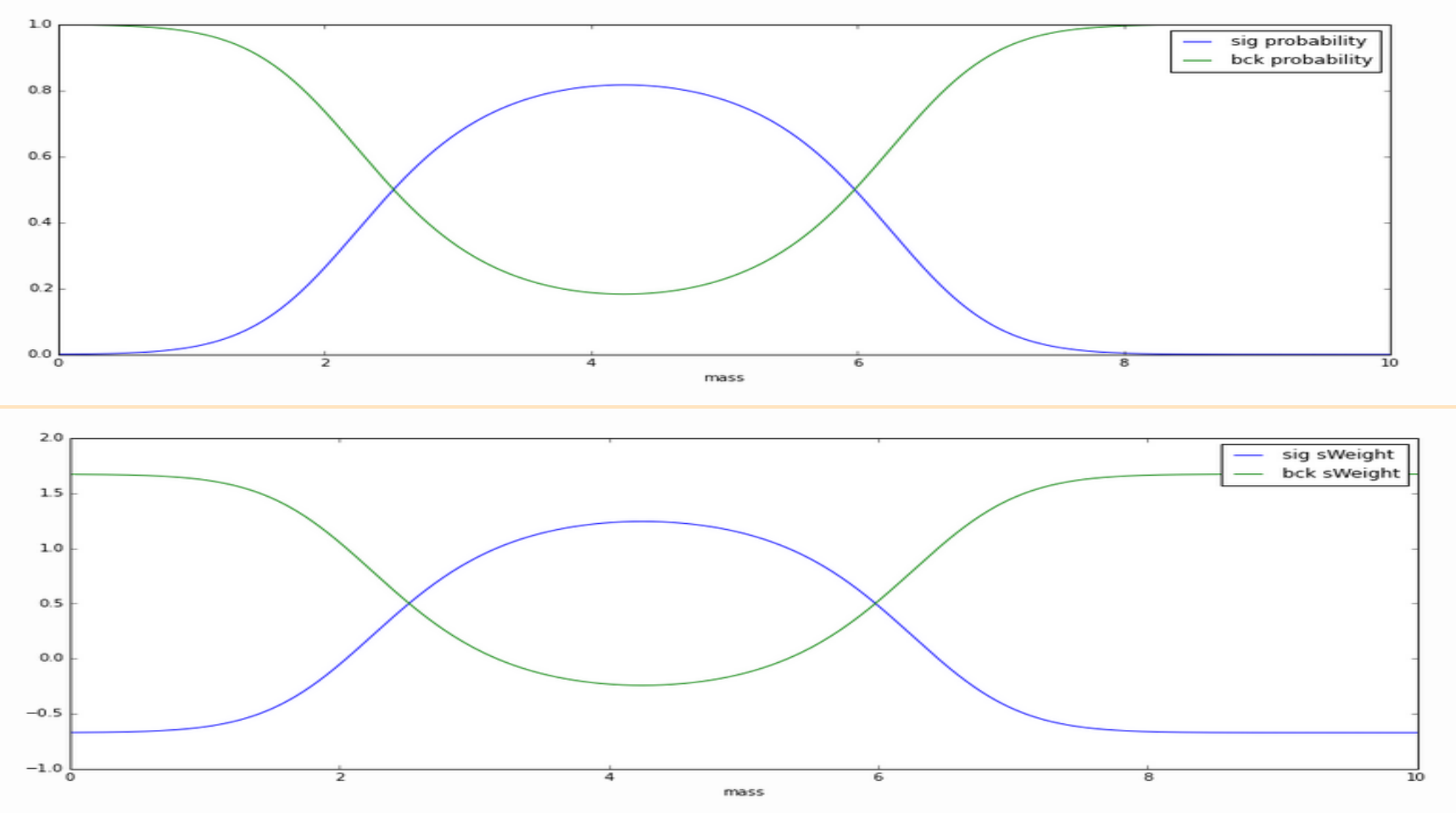



Image by agorozhnikov

- Reconstruct control distribution both for signal and background based on the weights computed from the discriminating distribution

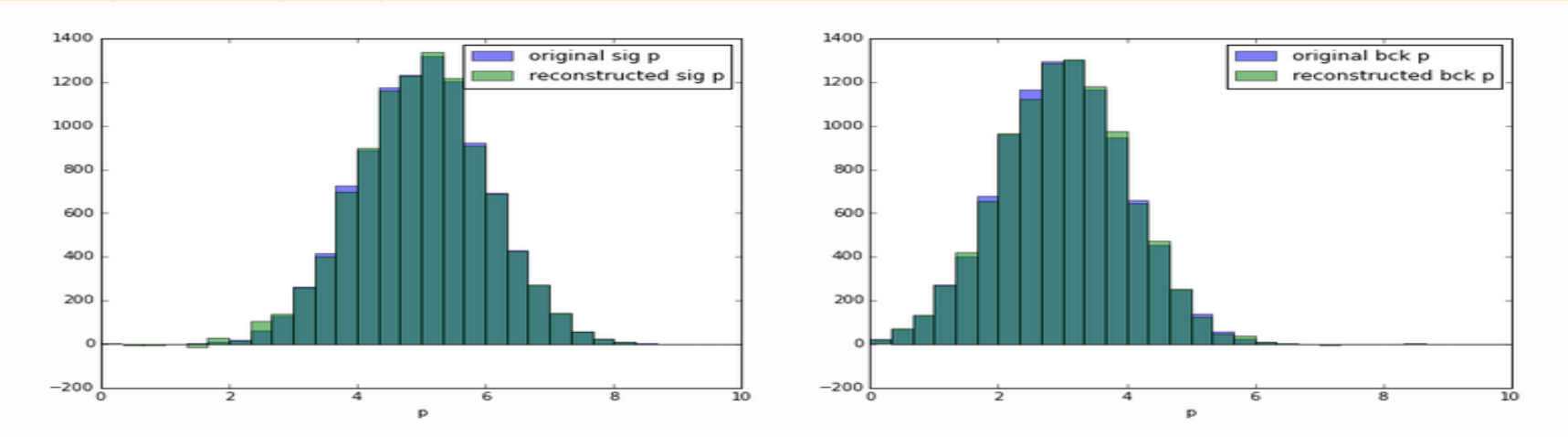



Image by agorozhnikov

- This works if control and discriminating variables are *independent* **within each class**
- If the variables are correlated (hence not independent) then bad things happen
- Reminder: variables might be uncorrelated but still dependent (remember linear correlation vs mutual information?)

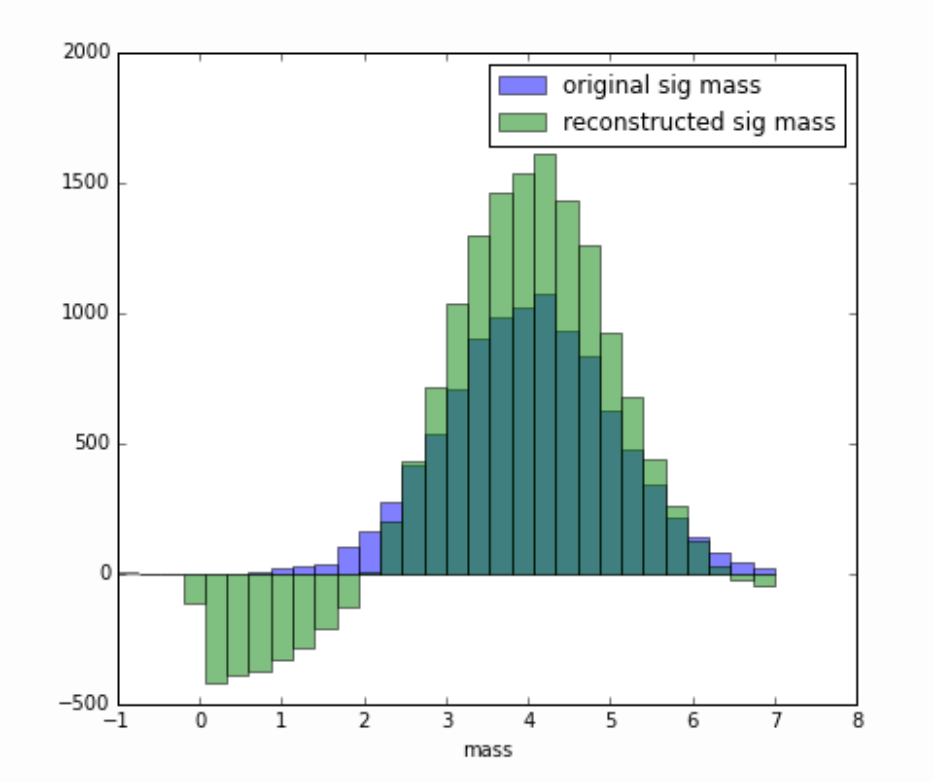



Image by agorozhnikov

- Extended likelihood: $ln\mathcal{L} = \sum_{e=1}^{N} ln\left\{ \sum_{i=1}^{N_s} N_i f_i(y:e) \right\} - \sum_{i=1}^{N_s} N_i$
  - $N$ number of events in the data sample
  - $N_s$ number of species of events populating the data sample (e.g. $N_s = 2$ if one sig and one bkg)
  - $N_i$ number of expected events for the $i$-th species
  - $f_i$ is the p.d.f. of the discriminating variables for the $i$-th species
  - $f_i(y_e)$ value taken by the p.d.f. for the event $e$, associated with a set of $y_e$ for teh set of discriminating variables)
  - $x$ is the set of control variables (by definition, they don't appear in the extended likelihood
- $ln\mathcal{L}$ is a function of the $N_s$ yields $N_i$ and of eventual implicit parameters designed to tune the $f_i$s on the data sample
- Determine $N_i$ and the implicit parameters by maximizing the likelihood

- The target is reconstructing the distribution of the control variables
- Validate by goodness-of-fit test of the MLE (not really convincing, it does not tell anything about the result for the control variables)
- If the distribution of the control variable is known for at least one source, then can compare the expected distribution to the one extracted by the method
  - Hence the name of control variable
  - Can even use a discriminating variable $y_i$ which does not improve the fit as a control variable $x$
  - Can even go wild, taking out a discriminating variable, recomputing the MLE, and use the excluded discriminating variable as control variable
- Going in signal-enriched region and compare the distribution of $x$ with a MonteCarlo simulation is discouraged
  - Can be used only if signal and background have significantly different shapes
  - The cuts for the signal-enriched fraction results in a reduced subsample of the data distribution, with larger statistical fluctuations
- Reconstruct true distribution of $x$ for the $n$-th species, $\mathbf{M}_n(x)$ from the sole knowledge of the p.d.f. $f_i$ of the $y$ variables

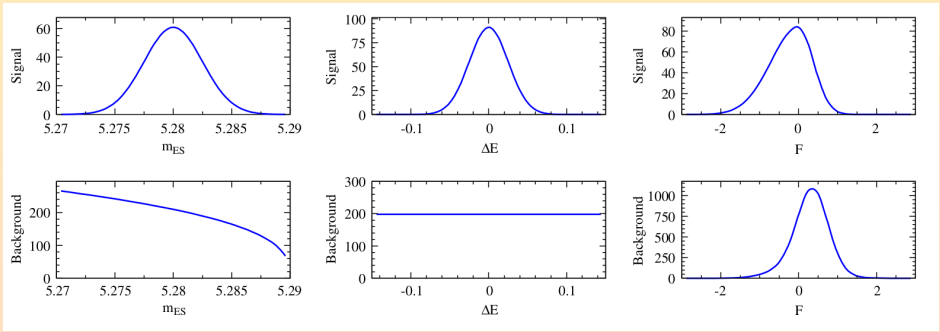
# $x$ is part of the $y$ set: the not-so-useful inPlot technique

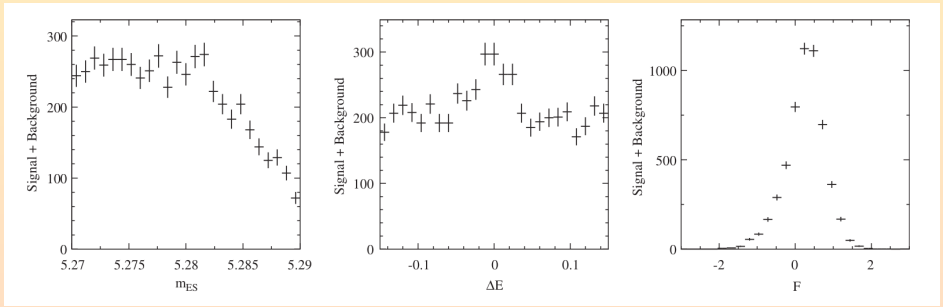
- $x$ is fully determined, $x = x(y)$, cannot be used as control variable
  - Authors say *fully correlated*, but what they should have writtens is *dependent*. The same mistake all across the paper
- Determine the yields $N_i$ for all species
- Reweight each event by a weight depending on $f_i$ and $\hat{N}_i$:
  $\mathcal{P}_n(y_e) = \frac{N_n f_n(y_e)}{\sum_{k=1}^{N_s} N_k f_k(y_e)}$
- Build the $x$-distribution $\tilde{M}_n$ by histogramming weighted events (summing $N_{\delta x}$ events lying inside the bin with center $\bar{x}$ and width $\delta x$)
  $N_n \tilde{M}_n(\bar{x})\delta x := \sum_{e \in \delta x} \mathcal{P}_n(y_e)$
- On average, the sum over the events in the bin can be substituted by
  $\int dy \sum_{j=1}^{N_s} N_j f_j(y) \delta(x(y) - \bar{x})\delta x$
- On average, substituting expected $N_i$ with fitted $N_i$,
  $\langle N_n \tilde{M}_n(\bar{x}) \rangle = \int dy \sum_{j=1}^{N_s} N_j f_j(y) \delta(x(y) - \bar{x})\delta x \mathcal{P}_n(y_e) = N_n \int dy \delta(x(y) - \bar{x}) f_n(y) =: N_n \boldsymbol{M}_n(\bar{x})$
- Sum over the events of the weights $\mathcal{P}_n$ provides estimate of the $x$ distribution of events of $n$-th species.
- Drawback: because $x = x(y)$, the p.d.f. of $x$ enters the weights $\rightarrow$ quality of fit cannot be assessed easily (biases when $f_i(y)$ not accurate)
  - Events in the tail of the implicit $M_n$ would enter $\tilde{M}_n$ with very small weight $\rightarrow$ bias towards the assumed $M_n$

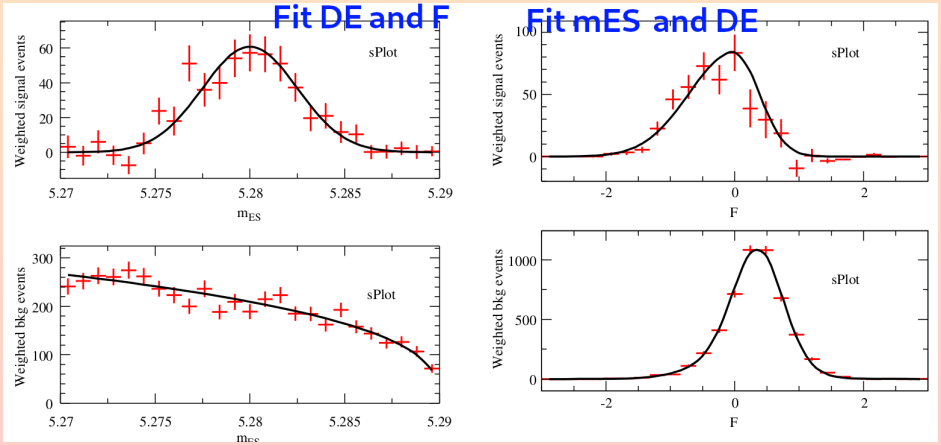
- "More precisely, the two sets of variables x and y are assumed to be uncorrelated: hence, the total PDFs $f_i(x,y)$ all factorize into products $\boldsymbol{M}_i(x)f_i(y)$
  - This holds only when $x$ and $y$ are independent; if they are uncorrelated they are not necessarily independent.
  - Mistake spread all across the paper; read the paper by substituting (in)dependent for (un)correlated
- Using the naive weight from the previous slide,
  $\langle N_n\tilde{M}_n(\bar{x})\rangle = \int\int dydx \sum_{j=1}^{N_s} N_j\boldsymbol{M}_j(x)f_j(y)\delta(x(y)-\bar{x})\mathcal{P}_n = N_n\sum_{j=1}^{N_s}\boldsymbol{M}_j(\bar{x})\left(N_j\int dy\frac{f_n(y)f_j(y)}{\sum_{k=1}^{N_s}N_kf_k(y)}\right)$
- This is different from $N_n\boldsymbol{M}_n(\bar{x})$ because the integral on $dy$ does not equal $\delta_{jn}^{cronecker}$
  - Unless $y$ is totally discriminant, but this would make the whole point moot; you could just apply cuts to select a pure sample of events of the $n$-th species
- Correction term related to the inverse of the covariance matrix (and to the Fisher information) which is minimized by the fit
  $\boldsymbol{V}_{nj}^{-1} = \frac{\partial^2(-\mathcal{L})}{\partial N_n\partial N_j} = \sum_{e=1}^{N}\frac{f_n(y_e)f_j(y_e)}{(\sum_{k=1}^{N_s}N_kf_k(y_e))^2}$
- Again, can substitute on average the sum with an integral, $\langle\boldsymbol{V}_{nj}^{-1}\rangle = \int dy\frac{f_n(y)f_j(y)}{\sum_{k=1}^{N_s}N_kf_k(y)}$
- ...and the variance can be used in the expression for $\tilde{M}$, yielding
  $\langle\tilde{M}_n(\bar{x})\rangle = \sum_{j=1}^{N_s}\boldsymbol{M}_j(\bar{x})N_j\langle\boldsymbol{V}_{nj}^{-1}\rangle$
- The distribution of interest can be recovered as: $N_n\boldsymbol{M}_n(\bar{x}) = \sum_{j=1}^{N_s}\langle\boldsymbol{V}_{nj}\rangle\langle\tilde{M}_j(\bar{x})\rangle$

# One-slide summary of the technique

- Assume the control variable $x$ does not belong to $y$ and is independent on $y$
- Compute the covariance-weighted sWeight ${}_s\mathcal{P}_n(y_e) := \frac{\sum_{j=1}^{N_s} \boldsymbol{V}_{nj} f_j(y_e)}{\sum_{k=1}^{N_s}} N_k f_k(y_e)$
- Obtain the distribution of the control variable $x$ from the sPlot histogram $N_n\, {}_s\tilde{M}_n(\bar{x})\delta x := \sum_{e\in\delta x} {}_s\mathcal{P}_n(y_e)$
- This reproduces on average the true distribution, $\langle N_n\, {}_s\tilde{M}_n(x)\rangle = N_n \boldsymbol{M}_n(x)$
  - If $x$ and $y$ are not independent, then this expression cannot be compared directly with the true distribution of the $n$-th species
  - MonteCarlo simulation of the whole procedure to obtain the expected distribution to compare with




Plots from https://doi.org/10.1016/j.nima.2005.08.106

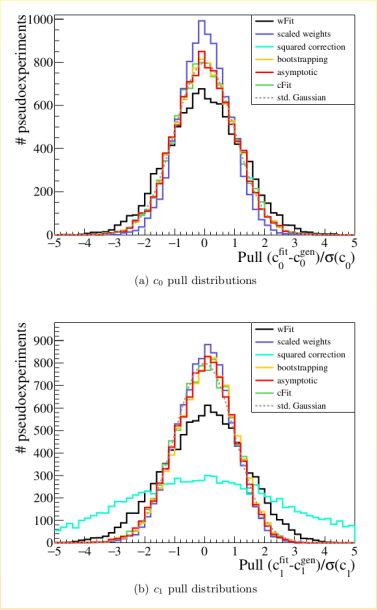
## Pitfall: independence of $x$ and $y$

- The factorization of $f_i(x, y)$ into $\boldsymbol{M}_i(x) f_i(y)$ relies on *independence* of the variables.
- If not satisfied, then must use $f_i(y|x)$ rather than the marginal $f_i(y)$, obtaining results biased in an a-priori unknown manner
- Independence is much stronger than lack of correlation: very restrictive condition
- Checking for correlation is a first step, but should check for independence!
- The solution is indeed to test for indepencence between the variables
  - Can be done only in sidebands (cannot use $S + B$ events, because independence in $S$ and $B$ separately does not imply independence in $S + B$)
- Necessary-but-not-sufficient for independence: rank-correlation-coefficient based tests (Kendall's or Spearman's)
  - Based on whether pairs of observations have the same rankings in the two variables
  - Spots only pairwise dependence
- Empirical copula
  - Multivariate cdf with uniform marginals, useful to check for conditional independence
  - Used pairwise between $x$ and $y$ (not within $y$ or $x$ variables)
  - Ideally between one $x$ and many $y$, but not yet available
- I argue that mutual correlation can be used with profit in this case

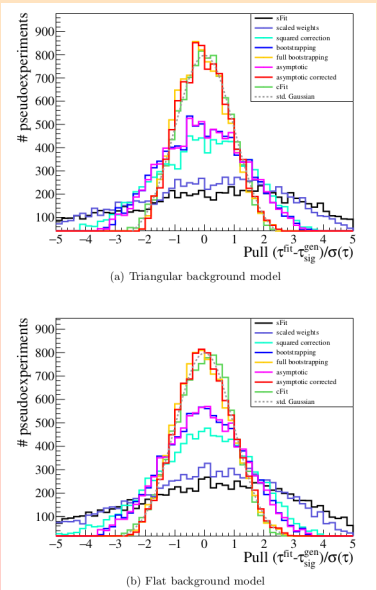
# Pitfall: coverage of the uncertainties

- The paper claims that asymptotically the statistical uncertainty is $\sigma[N_n\, {}_s\tilde{M}_n(\bar{x})\delta x] := \sqrt{\sum_{e\in\delta x}({}_s\mathcal{P}_n)^2}$
  - Demonstration is provided based on the definition of variance, $\left\langle \left( \sum_{e\in\delta x} {}_s\mathcal{P}_n \right)^2 \right\rangle - \left\langle \sum_{e\in\delta x} {}_s\mathcal{P}_n \right\rangle^2$
- Equations are incorrect for finite samples: they ignore the correlations between the weights introduced by the covariance-weighted version of the weights
  - $V_{nj}$ are random variables estimated in data, subject to correlated statistical uncertainties
- Confidence intervals from the second derivative of MLE do not yield correct coverage in presence of event weights
  - Noted by Langenbruch in https://arxiv.org/abs/1911.01303, but his proposed solutions (right) do not address the previous problem
- Possible solutions
  - Rederive variance propagating uncertainties on the weight corrections
  - Use resampling or toys

Example of acceptance effects




Example of background subtraction




Plots from arXiv:1911.01303

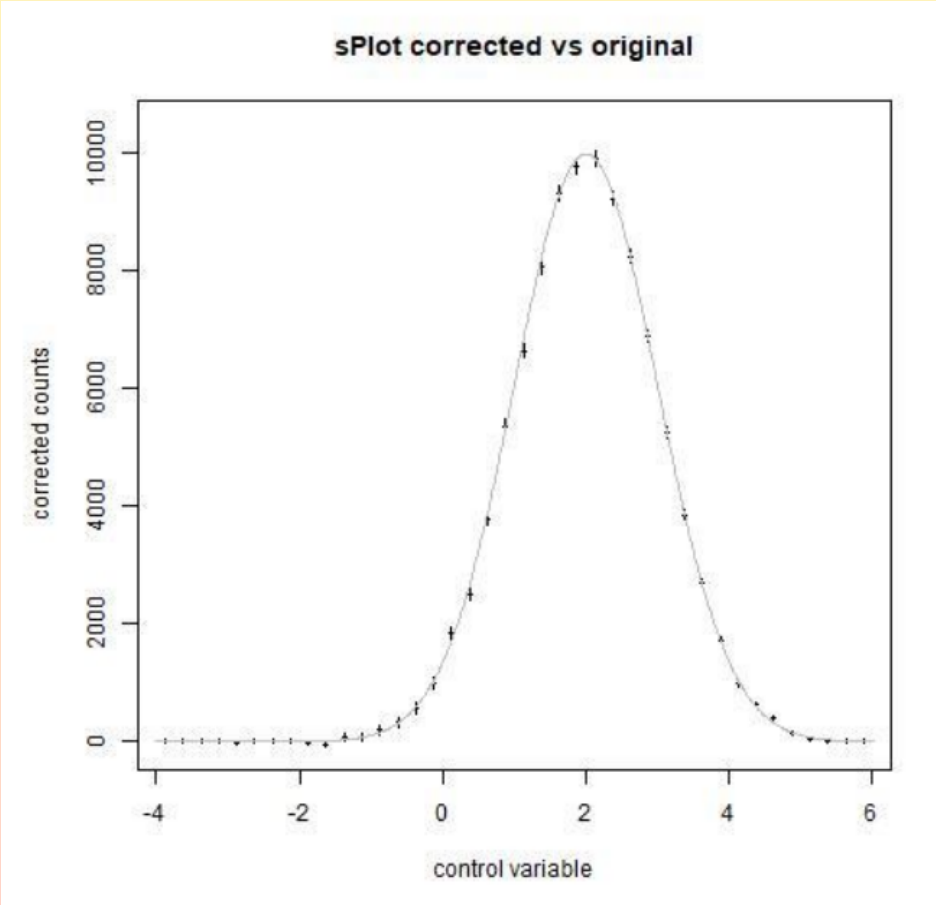
# Pitfall: inference

- Correlated weights induce correlations in the sPlot bin contents
- Any inference from these histograms should take those correlations into account
  - Not addressed in sPlot paper
  - Assumption of lack of correlation cannot be done
- Can be avoided by proper calculation of the bin covariance matrix (either via proper propagation from $\boldsymbol{V}_{nj}$ or toys/resampling)
  - Provided that this bin covariance matrix is then used for any subsequent inference

- Some naive toy example by Francisco Matorras to illustrate the entity of the issues
- Two samples (signal and background)
- One control variable ("mass", plotted) and one discriminant variable ("other mass" fitted and which defines the weights)
- Bivariate normal pdf, $S \sim Gaus(2,1) \times Gaus(2,1)$, $B \sim Gaus(0,1) \times Gaus(0,1)$, $\frac{S}{S+B} = 0.1$
- Mimic sPlot analysis, repeating for many toys to check for bias and coverage
  - Calculate weights
  - Calculate yield / signal fraction as sum of weights
  - Plot the distribution for control variable using sWeights
  - Extract mass and fractions from fit of the sWeighted distributions
- No pretense of generality (bias and coverage will in general exhibit different issues in other scenarios), but indicative of the typical issues with bias and coverage

# No correlation for variables neither in signal nor in background

- No bias (assumption of independence is satisfied)
- Uncertainties don't cover correctly (uncertainties don't account for correlations between the weights)



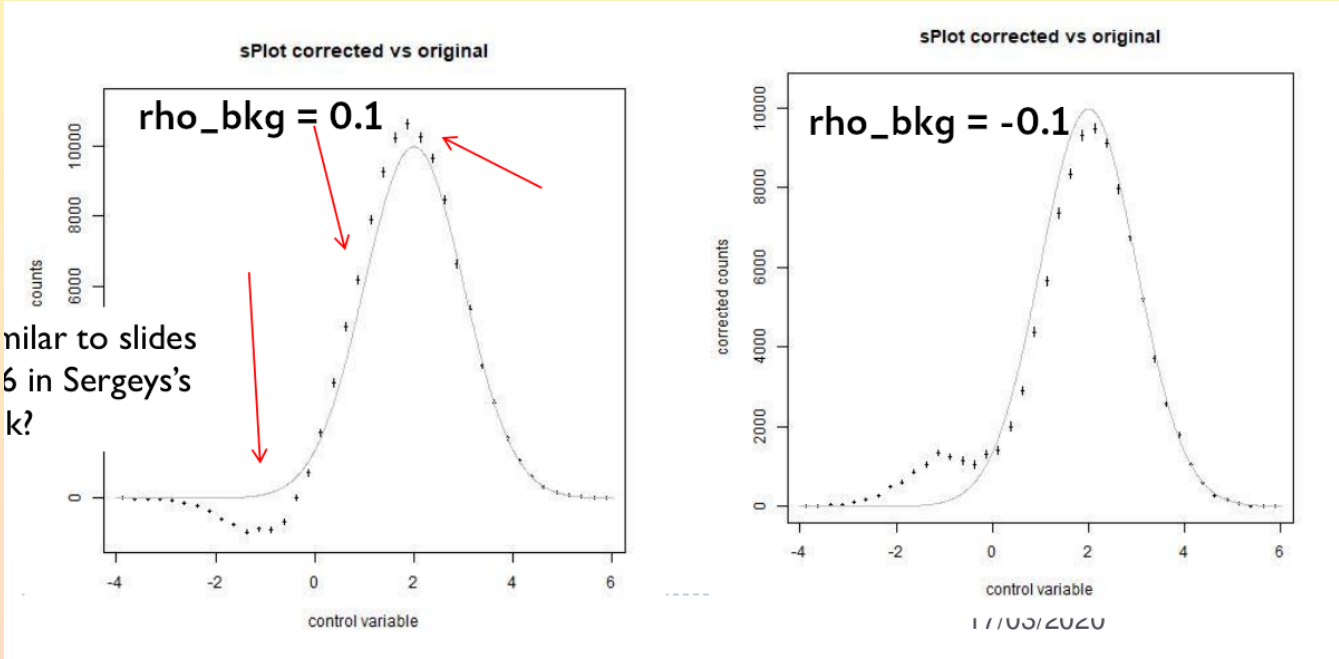

| ρsig=0,ρbk=0 | f (truth=0.1) | Coverage | Mass (truth=2) | Coverage |
|---|---|---|---|---|
| Pure weights/ "fit" | 0.0997±0.0005 | 0.70 | 2.002±0.007 | 0.31 |
| Fit around the peak | 0.0999±0.0005 | 0.68 | 1.998±0.005 | 0.48 |

Note: shown errors represent errors on the mean of the parameter, from the spread of the different pseudoexps. Coverage is calculated from the central value and uncertainty of each of the pseudoexps

Plots by Francisco Matorras

# Small ($\pm 0.1$) correlation for variables in background

- Bias! (assumption of independence is not satisfied)
- Uncertainties don't cover correctly (uncertainties don't account for correlations between the weights, plus non-independence of variables)
  - Similar results for $\rho_{bkg} = 0.1$

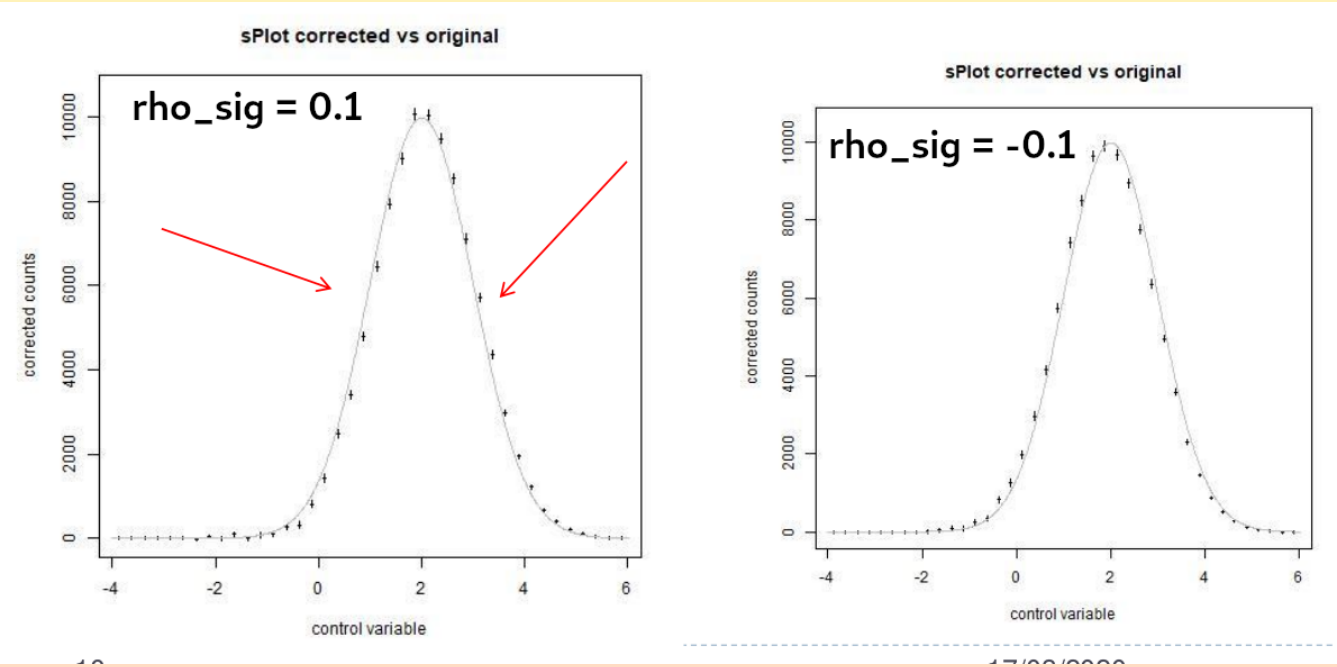



| ρsig=0,ρbk=-0.1 | f (truth=0.1) | Coverage | Mass (truth=2) | Coverage |
|---|---|---|---|---|
| Pure weights/ "fit" | 0.0995±0.0004 | 0.78 | 1.814±0.007 | 0.01 |
| Fit around the peak | 0.0991±0.0004 | 0.12 | 2.072±0.005 | 0.17 |

Plots by Francisco Matorras

# Small ($\pm 0.1$) correlation for variables in signal

- Bias! (assumption of independence is not satisfied)
- Uncertainties don't cover correctly (uncertainties don't account for correlations between the weights, plus non-independence of variables)
  - Similar results for $\rho_{sig} = -0.1$




| ρsig=0.1,ρbk =0 | f (truth=0.1) | Coverage | Mass (truth=2) | Coverage |
|---|---|---|---|---|
| Pure weights/ "fit" | 0.1003±0.0005 | 0.69 | 2.068±0.007 | 0.27 |
| Fit around the peak | 0.1000±0.0005 | 0.71 | 2.055±0.005 | 0.26 |

Plots by Francisco Matorras

- Similar to ABCD method, but extended to many regions (bins)
  - Hence the name: $A..Z$
- Already in use since decades
  - It's a formalization with fancy name, documented in a recent talk
- Solve the plugged-in equations $\hat{s}_{ij} + \hat{b}_{ij} = d_{ij}$ for the estimated number of signal and background events (can extend to multiple backgrounds)
  - $M$: bins in $x$. $N$: bins in $y$
  - $s_{ij}$: observed number of sig events in bin $(x_i, y_j)$. Corresponding estimated number: $\hat{s}_{ij}$
  - $b_{ij}$: observed number of bkg events in bin $(x_i, y_j)$. Corresponding estimated number: $\hat{b}_{ij}$
  - $d_{ij} = s_{ij} + b_{ij}$ observed data yield in bin $(x_i, y_j)$
  - $p_j^s$: probability for a sig event to fall into the $j$-th bin of $y$
  - $p_j^b$: probability for a bkg event to fall into the $j$-th bin of $y$
- Equation can be rewritten as $p_j^s n_i^s + p_j^b n_i^b = d_{ij}$, where $n_i^s$ and $n_i^b$ are estimates of the total yields for sig and bkg in the distribution of the parameter of interest
  - $M \times N$ equations, $2M$ unknown parameters
  - $N = 2$: analytical solution (as in ABCD)
  - $N > 2$: likelihood fit or $\chi^2$ fit: for each bin $i$ of $x$, fit $y$ for $n_i^s$ and $n_i^b$
- Covariance matrices calculated easily (parameters are linear combinations of $d_{ij}$
  - Must still check for independence of the variables in sidebands, but at least if they are independent you should get correct coverage with simple formulas

- sPlot is a method to reconstruct the distribution of a control variable of unknown p.d.f., using only the knowledge of the p.d.f. s of a discriminating variable
  - Relies on the assumption of independence between the (set of) control variables and the (set of) discriminating variables
  - Provides also an estimate of the statistical uncertainty on the estimated distributions
- Three issues highlighted by the StatComm (details in https://indico.cern.ch/event/895770/
- Assumption of independence
  - If not satisfied, leads to unpredictable biases
  - Many analyses don't check the assumption
  - The remaining ones check that the variables are uncorrelated, but this is necessary but not sufficient for independence!
  - Possible solution: check for independence! (rank correlation coefficient, or copulas) (can be done only in sidebands)
- Statistical uncertainties computed by the method are wrong
  - $V_{nj}$ are random variables estimated in (the same) data $\rightarrow$ correlated statistical uncertainties
  - Confidence intervals from second derivative of likelihood don't yield correct coverage for weighted events anyways
  - Possible solutions: rederive the expression for the uncertainty by propagating things correctly, or use resampling/toys
- Inference
  - Correlated weights induce correlations in sPlot bin contents
  - Inference based on those histograms should take these correlations into account
  - Possible solutions: proper calculation of bin covariance matrix (either by propagating things properly, or using resampling/toys)

- Toy examples show that significant issues can arise even in very simple cases
  - Bias can arise even for very small correlations between the variables
  - Significant (under)coverage can arise even when no correlations between the variables!
- A simpler, template-based method (essentially a generalization of ABCD) formalized by Louis but in use since decades
  - For each bin of $x$, likelihood (or $\chi^2$) fit to the discriminating variable $y$
- A colleague's dramatic conclusion

> - The current use of sPlot is fraught with problems. If used indiscriminately, sPlot will produce biased estimates with incorrect uncertainties. At the same time, the users of sPlot will believe that they have done nothing wrong. This is the worst situation imaginable in scientific practice.
> - Perhaps, solutions mitigating sPlot problems can be developed (e.g., along the lines outlined in this talk). Before that happens, however, the use of sPlot for producing publication results should be discouraged.

# Always check your assumptions

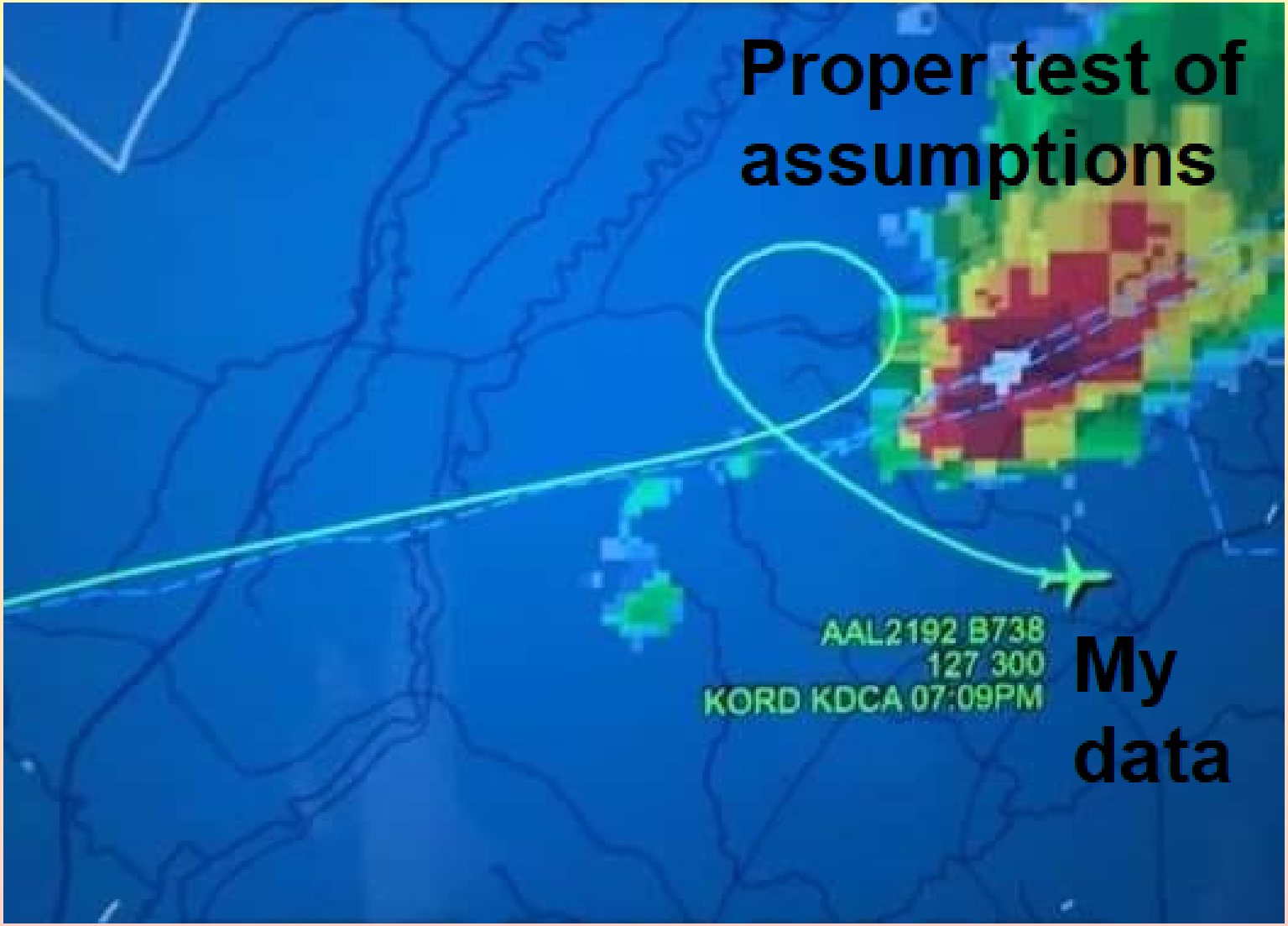
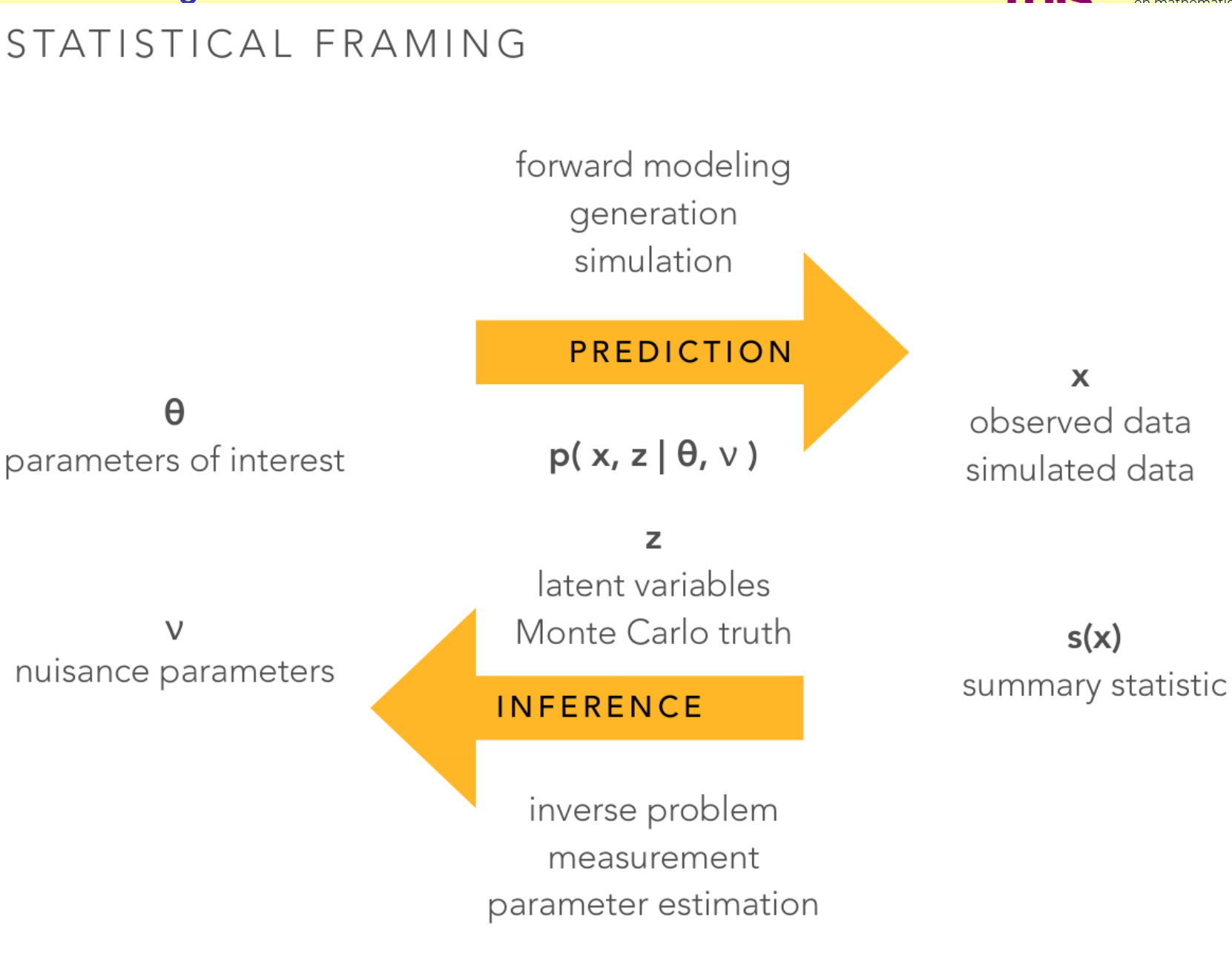



Image from the Statistical Statistics Memes Facebook Page




From Kyle Cranmer's PhyStat seminar

# The likelihood is a landmark for most statistical inference methods



- (Profile) maximum likelihood fits
  - Point estimate by maximizing the likelihood function
  - Interval estimate by intersection with likelihood function
  - Feldman-Cousins (build intervals using the likelihood ratio for proper ordering of probability elements)
  - Test of hypothesis using ratios of the likelihood of the two hypotheses
- Bayesian methods rely on computing posterior distribution $p(\theta|\vec{x}) \propto p(\vec{x}|\theta)\pi(\theta)$
  - Need the likelihood to build the posterior: usually resample posterior (Markov Chain MonteCarlo)
  - Point and interval estimates from posterior shape
  - Test of hypotheses from ratio of marginal likelihoods (Bayes Factor)

- We want to obtain the posterior for $\lambda$ given data $X$ and model $M$ you need to obtain the evidence by integration

$$P(\lambda|X,M) = \frac{P(X|\lambda,M)\pi(\lambda|H)}{P(X|H)} = \frac{P(X|\lambda,M)\pi(\lambda|H)}{\int P(X|\lambda,M)\pi(\lambda|H)d\lambda}$$

- We could perform MonteCarlo integration to evaluate $I = E_f[h(X)] = \int_{\mathcal{X}} h(h)f(x)dx$
  - $f$ must be a closed form (otherwise ABC, see later!)
  - Sample from $f(x)$ to approximate $I$ with empirical average

$$\bar{h}_m = \frac{1}{m}\sum_{j=1}^{m} h(x_j)$$

  - with variance

$$\bar{v}_m = \frac{1}{m-1}\sum_{j=1}^{m}[h(x_j) - \bar{h}_m]^2$$

  - For large $m$

$$\frac{\bar{h}_m - E_f[h(X)]}{\sqrt{v_m}} \sim Gaus(0,1)$$

# Monte Carlo integration, when possible, will have a stable variance

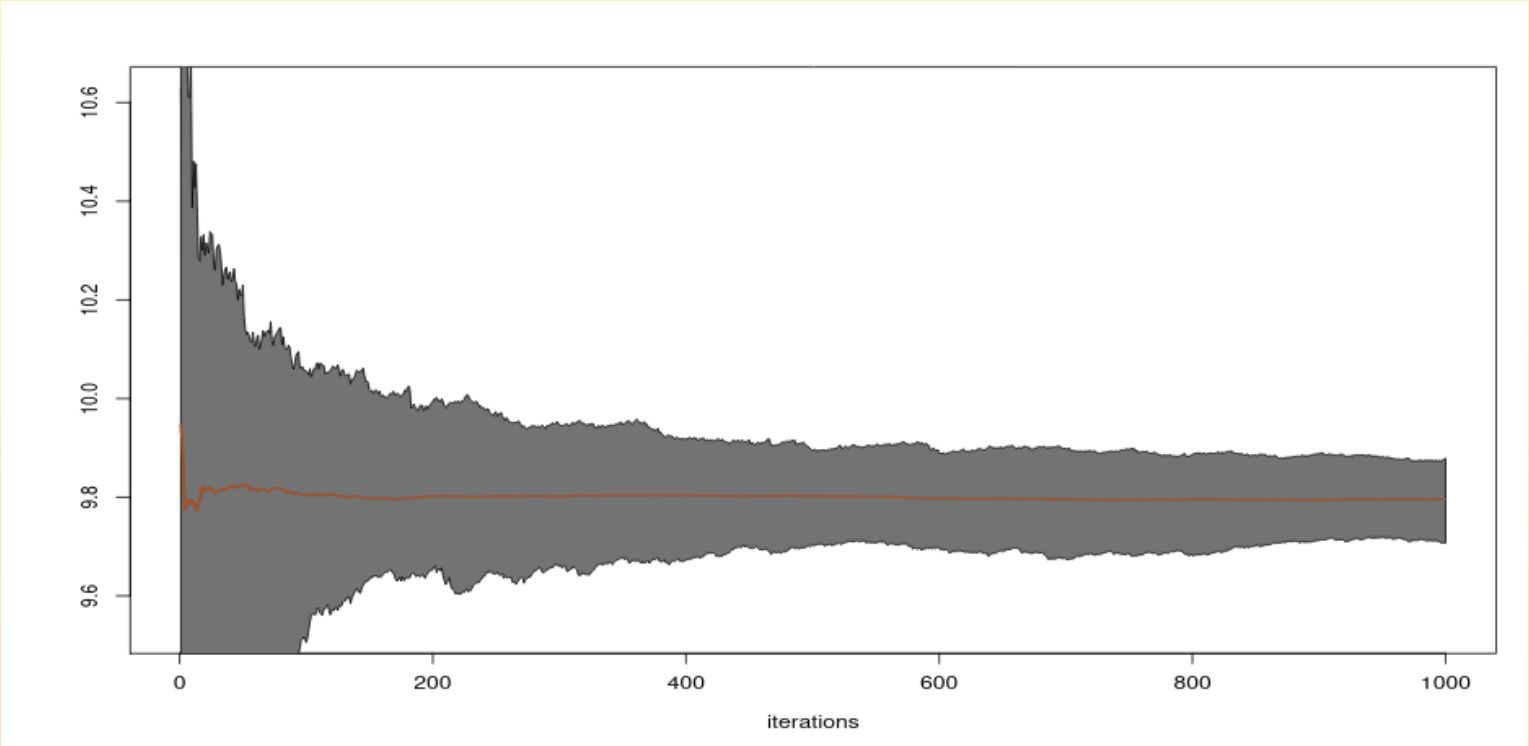



Image from Christian P. Robert

- For integrating the tails, need to simulate huge amounts of events
- Can mitigate by importance sampling: modify the probability of generating events in problematic regions (e.g. tails)
- Sample from $g(x)$
- Estimate integral from modified empirical average

$$\bar{h}_m = \frac{1}{m} \sum_{j=1}^{m} \frac{f(x_j)}{g(x_j)} h(x_j)$$

- Convergence of the estimator still guaranteed

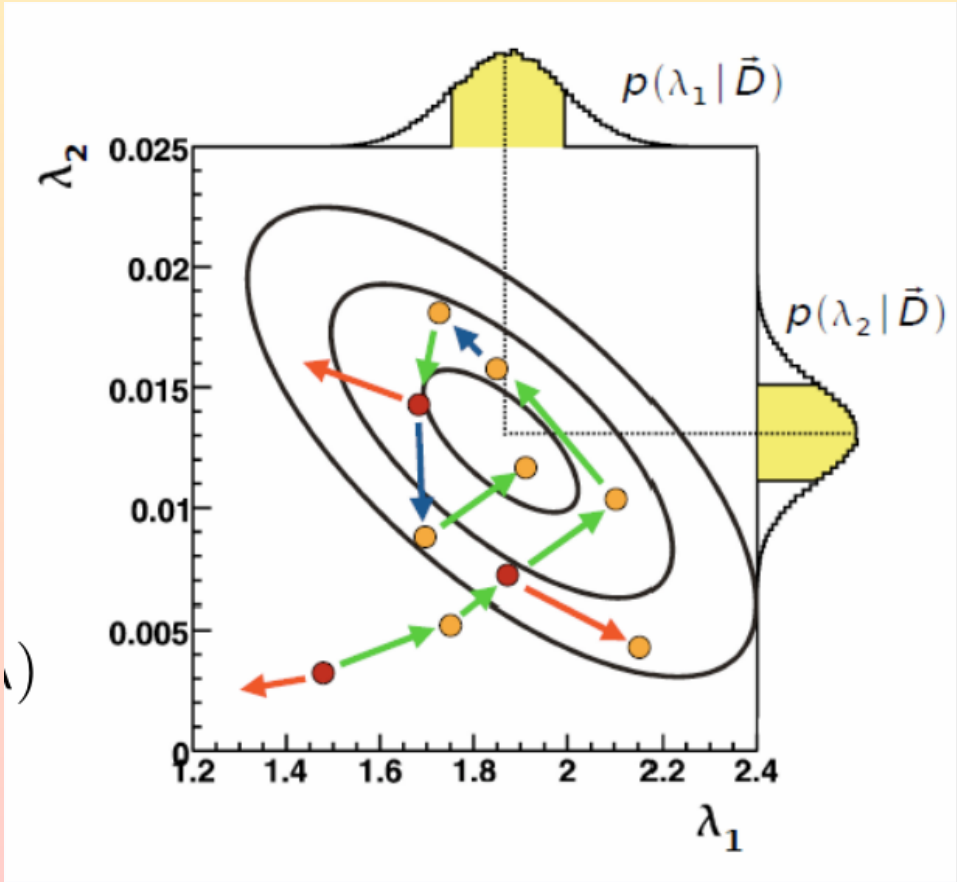
- Integration difficult/impossible in high dimension ($\vec{\lambda}$)
- Replace integration with a path along the function

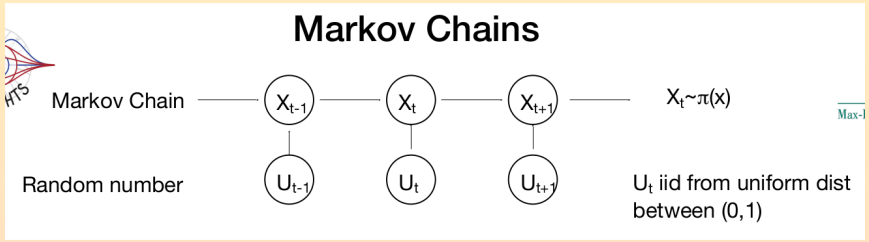
- Markov chain: a sequence of random variables whose distribution evolves over a “time” variable as a function of the previous realization
- Describe the chain via a *transition kernel* defining the migration probabilities from a state to any other state
  - Discrete: **stochastic** matrix $K_{xy} = P(x_n = y | x_{n-1} = x)$
  - Continue: conditional density $\mathcal{K}$ via an integral
- Given $K$, a sequence $X_n$ of random variables is a Markov Chain if for any $t$

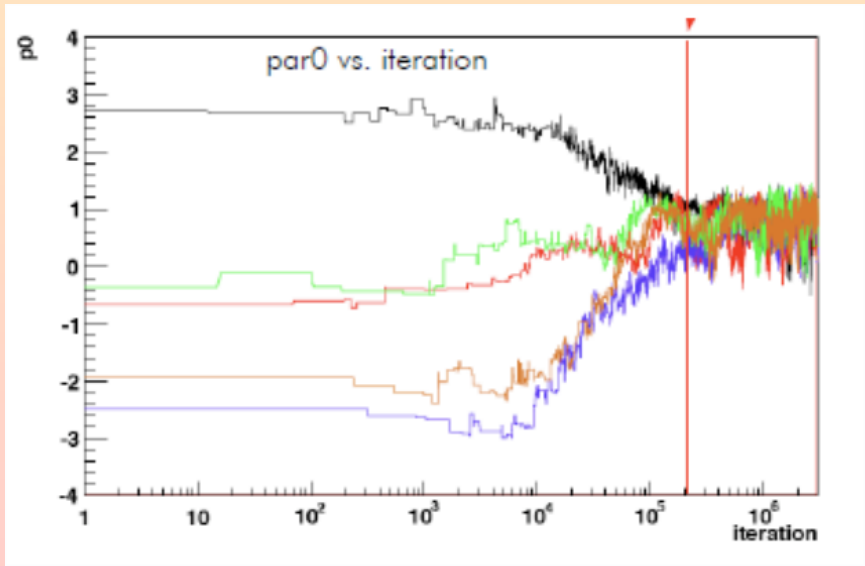
$$P(x_{k+1} \in A | x_0, x_1, ..., x_k) = P(x_{k+1} \in A | x_k) = \int_A \mathcal{K}(x_k) dx$$

## Map posterior probability by sampling (Allen Caldwell's seminar and 1808.08051)

- Markov process (probability of transition depends on current state only) on a finite phase space: Markov chain
- For recurrent, irreducible, aperiodic chains, Basic Limit Theorem guarantees after many iterations result independent on initial state
- Metropolis-Hastings: accept/reject algorithm, accepting transition if it goes towards more probable state
  - Burn-in: multiple chains to check for stationarity
- Still computationally costly!

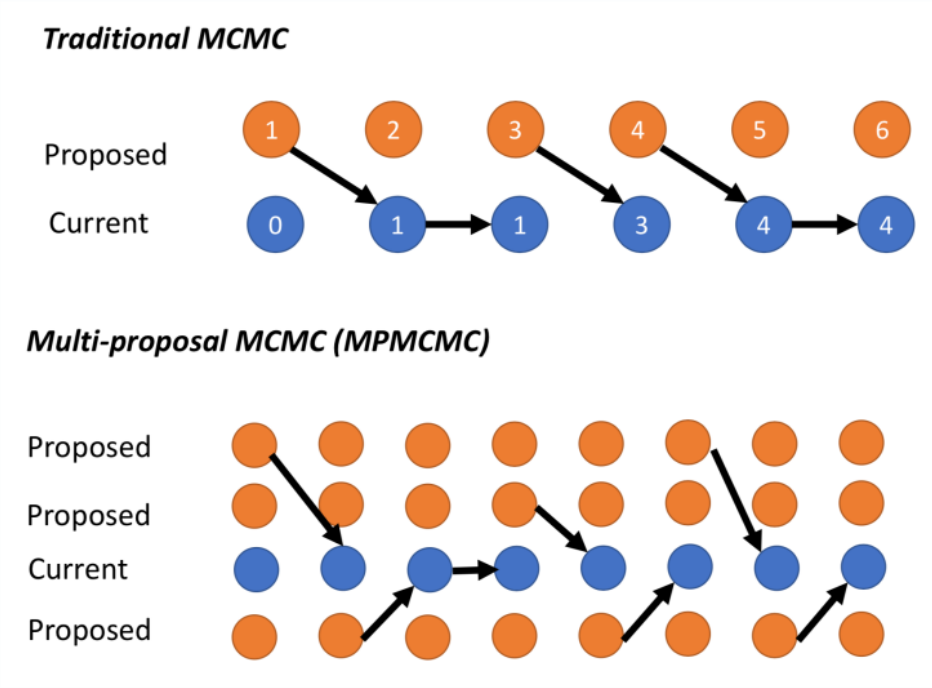


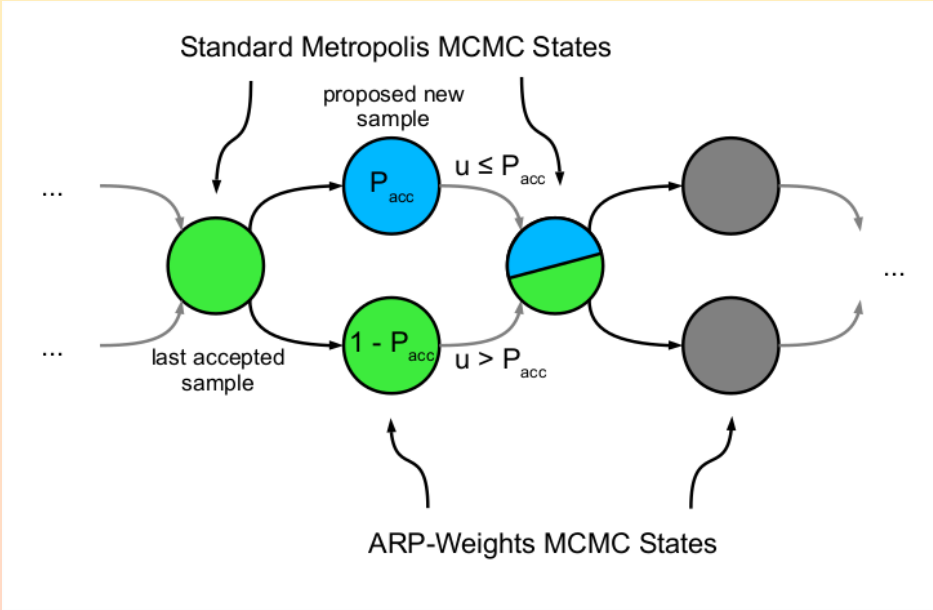



# Improving the sampling (Allen Caldwell's seminar and 1808.08051)

- Strategies to improve the sampling
  - Inspired by physics (not shown here): Hamiltonian MCMC!



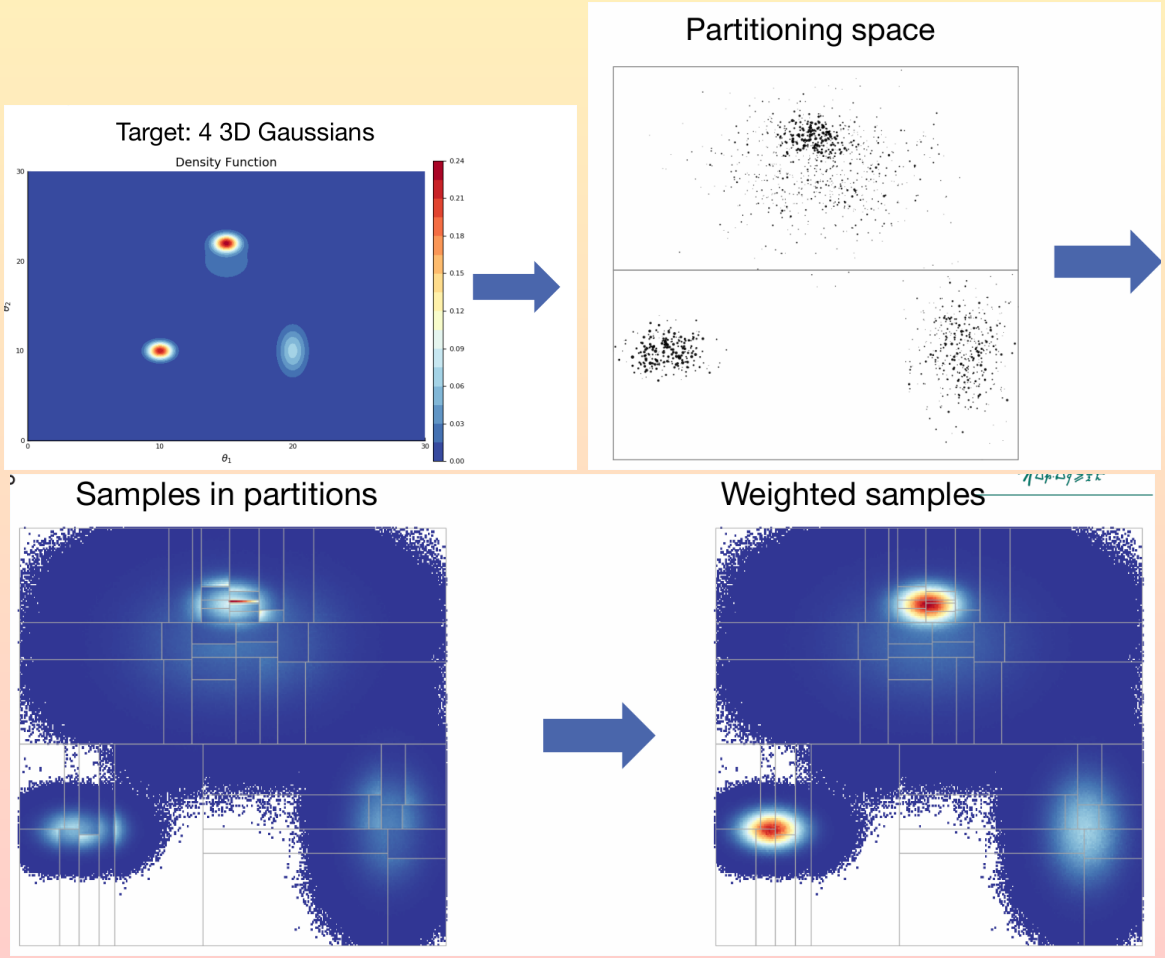
# Accelerating MCMC computations 1808.08051

UCLouvain
Institut de recherche en mathématique et physique

- Improve sampling either via efficient sampling (e.g. Hamiltonian MC, etc) or via massive parallelization: AHMI
  - Adaptive Harmonic Mean Integration (AHMI) good with multimodal posteriors
  - Partition space to obtain regions with moderate spread of values, can then integrate by using harmonic mean to reweight by volume

$$E\left[\frac{1}{f(\lambda)}\right]_{\hat{f}_\Delta(\lambda)} = \int_\Delta \frac{1}{f(\lambda)} \cdot \frac{f(\lambda)}{I_\Delta} d\lambda = \frac{V_\Delta}{I_\Delta}$$

$$\hat{I}_i = \frac{\hat{I}_\Delta}{\hat{r}} = \frac{f_{\min} N_{\Omega_i} V_\Delta}{\sum_{\in \Delta} f_{\min}/f(\lambda_i)}$$



- In HEP we often don't have access to the likelihood
- Monte Carlo generators are used to generate samples distributed according to a given likelihood function $x \sim p(x|\theta)$
- The likelihood sometimes is intractable

$$p(x|\theta) = \int dz p(x,z|\theta) = \int dz p_x(x|\theta,z) \prod_i p_i(z_i|\theta, z_{<i})$$

- Latent variables
  - Matrix element
  - Parton shower
  - Detector...

- Sample from the density using accessory variables $\rightarrow$ Approximate Bayesian Computation
- Smart ways of integrating the generating function $\rightarrow$ Matrix Element Method
- Bypass the integration by learning the generating function $\rightarrow$ Surrogate models by learning the likelihood (ratio)
- Active learning: learn surrogates by alternating simulation and inference stages (not treated today)
  - Train autoregressive flows on simulated data in order to learn a model of the likelihood in the region of high posterior density
- Learn the structure of the data, together with the density, for intractable likelihoods $\rightarrow$ manifold learning
- Do inference by finding a "smart", optimal summary statistic $\rightarrow$ INFERNO
- Many more (not discussed today)

# “Classical” likelihood-free inference: ABC

- Approximate Bayesian Computation
  - Almost Bayes Can
- “ABC is a recent computational technique that only requires a generative model, i.e., being able to sample from the density $f(\cdot|\theta)$”
  - Developed originally (and still being developed) in population genetics (Griffith et al., 1997; Tavaré et al., 1999)


- Sometimes the likelihood function $f(\vec{x}|\theta)$ is not available, or there are latent variables making its computation impossible or very costly (high-dimensional $\vec{z}$)

$$f(\vec{x}|\theta) = \int_{\mathcal{Z}} f(\vec{x}, \vec{z}|\theta) d\vec{z}$$

## When the likelihood is not in closed form

- Want to compute the posterior: $p(\theta|\vec{x}) \propto f(\vec{x}|\theta)\pi(\theta)$
- Use a likelihood-free rejection technique to obviate the lack of closed form for the likelihood

## The ABC accept/reject algorithm (Tavaré et al., 1997)

- Let $\vec{x} \sim f(\vec{x}|\theta)$ be an observation, under the prior $\pi(\theta)$
- Keep jointly simulating
  - $\theta' \sim \pi(\theta)$
  - $\vec{z} \sim f(\vec{z}|\theta')$
- Until the auxiliary variable $\vec{z}$ is equal to the observed value, $\vec{z} = \vec{x}$

- It works!

$$\begin{aligned} f(\theta_i) &\propto \sum_{\vec{z}\in\mathcal{D}} f(\vec{z}|\theta_i)\pi(\theta_i)\mathbb{1}_{\vec{x}}(\vec{z}) \\ &\propto f(\vec{x}|\theta_i) \\ &= \pi(\theta_i|\vec{x}) \end{aligned}$$

## ABC is Approximative

- If $\vec{x}$ is a continous random variable, replace equality $\vec{z} = \vec{x}$ with a tolerance condition defined via a distance metric $\rho$:

$$\rho(\vec{x}, \vec{z}) < \epsilon$$

- The output of the algorithm will be (Pritchard et al., 1999)

$$P_\theta\Big\{\rho(\vec{x}, \vec{z}) < \epsilon\Big\} \propto \pi\Big(\theta\Big|\rho(\vec{x}, \vec{z}) < \epsilon\Big)$$

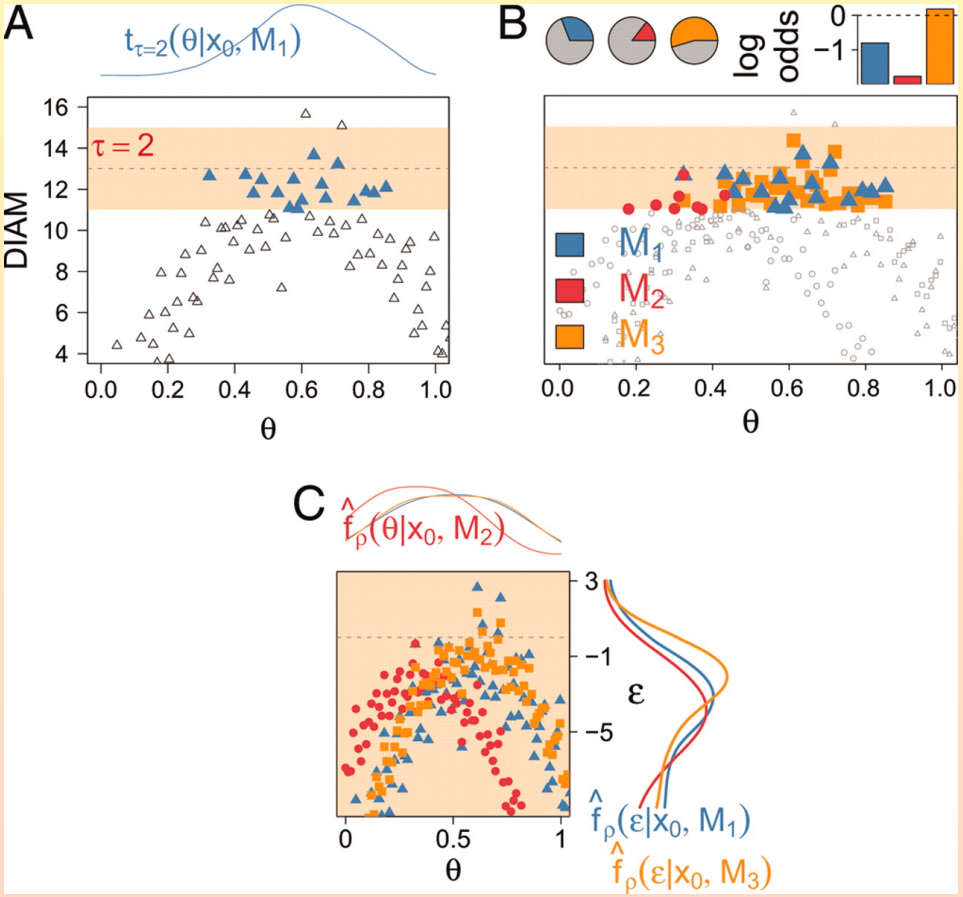
- Will converge to the posterior for $\epsilon \to 0$

- Distances in the space of a statistic $\eta(\cdot)$
  - Faster
  - $\eta(\cdot)$ not necessarily a sufficient statistic

**for** $i = 1$ **to** $N$ **do**
  **repeat**
    generate $\theta'$ from the prior $\pi(\cdot)$
    generate $\vec{z}$ from the likelihood $f(\cdot|\theta')$
  **until** $\rho\Big\{\eta(\vec{z}), \eta(yvec)\Big\} \leq \epsilon$
  set $\theta_i = \theta'$
**end for**

# Simulating from the prior

- Poor efficiency in simulating from the prior
- Various proposals to solve this
  - Modify proposal to sample more efficiently in the vicinity of $\vec{x}$
  - View as conditional density estimate, allow larger $\epsilon$
  - Many many more... (ABC-NP, ABC-NCH, ABC-kNN, ABC-MCMC, ABC-PMC, ABC-SMC...)
- ABC$_\mu$: **Augment** the likelihood by including unknown term accounting for $\epsilon$ in the inferential framework (Ratmann et al., 2009)

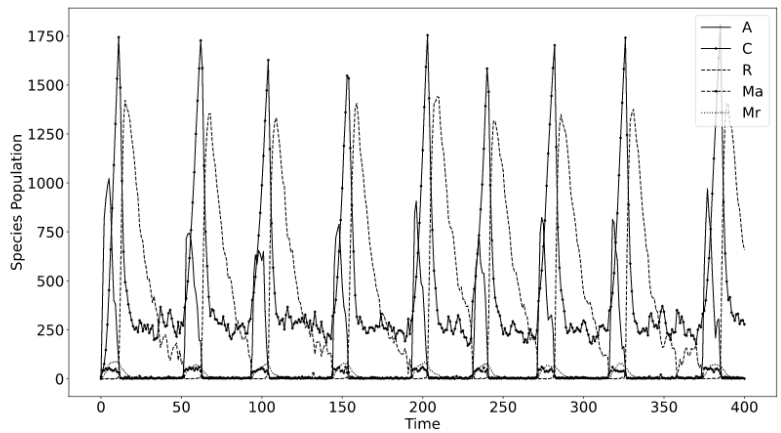


# Choice of summary statistic

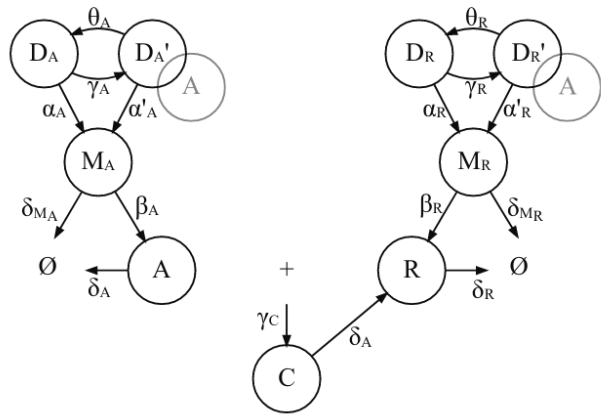
- Largely debated field
  - Sufficient/non sufficient
  - How to choose an optimal statistic
  - Won't enter into detail
- Highlight: use CNNs to learn the summary statistic (Åkesson, 2020)
  - Mean posterior estimation error vastly reduced

| Architecture | Train Time | No. of Parameters | |
|---|---|---|---|
| | | Total | Trainable |
| DNN | 26s | 457, 167 | 450, 767 |
| $\text{PEN}_{10}$ | 1m 53s | 385, 727 | 383, 327 |
| CNN | 4m 41s | 492, 415 | 490, 015 |

| Param. | Neural network architectures | | | Traditional statistics | |
|---|---|---|---|---|---|
| | CNN | $\text{PEN}_{10}$ | DNN | All | AS |
| $\alpha_A$ | **0.392** | 0.402 | 0.639 | 1.009 | 1.378 |
| $\alpha'_A$ | **0.512** | 0.532 | 0.744 | 0.975 | 0.904 |
| $\alpha_R$ | **0.920** | 0.938 | 0.990 | 1.785 | 1.936 |
| $\alpha'_R$ | **0.758** | 0.790 | 0.890 | 0.997 | 0.931 |
| $\beta_A$ | **0.505** | 0.523 | 0.836 | 1.352 | 1.289 |
| $\beta_R$ | **0.490** | 0.514 | 0.691 | 1.161 | 1.214 |
| $\delta_{MA}$ | **0.494** | 0.507 | 0.769 | 1.815 | 1.652 |
| $\delta_{MR}$ | **0.403** | 0.425 | 0.566 | 0.783 | 0.964 |
| $\delta_A$ | **0.255** | 0.256 | 0.594 | 0.869 | 0.885 |
| $\delta_R$ | **0.366** | 0.399 | 0.774 | 0.942 | 0.984 |
| $\gamma_A$ | **0.867** | 0.886 | 0.972 | 1.211 | 1.219 |
| $\gamma_R$ | **0.786** | 0.806 | 0.922 | 1.384 | 1.261 |
| $\gamma_C$ | **0.608** | 0.644 | 0.899 | 0.787 | 0.935 |
| $\theta_A$ | **0.601** | 0.627 | 0.907 | 1.088 | 1.119 |
| $\theta_R$ | **0.819** | 0.833 | 0.931 | 1.175 | 1.122 |
| *mean* | **0.585** | 0.605 | 0.808 | 1.155 | 1.186 |



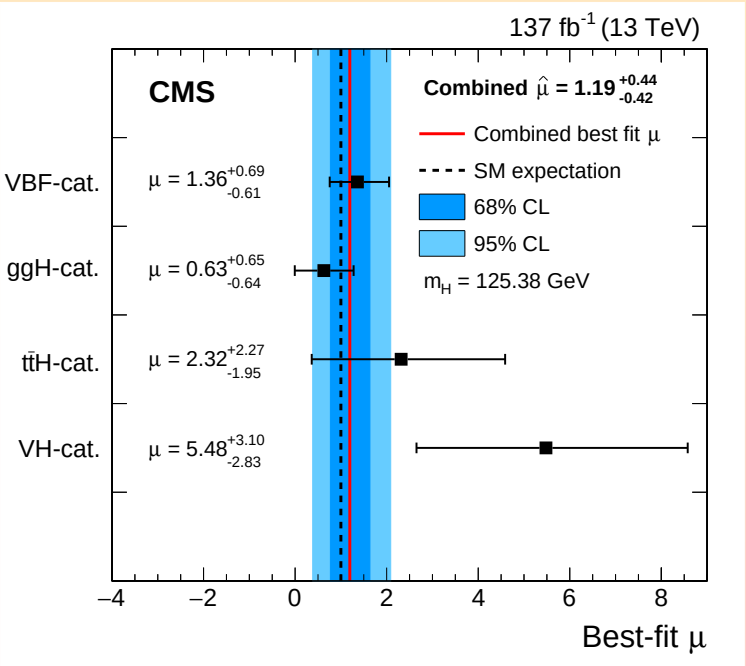

(a) Trajectories of genetic oscillator

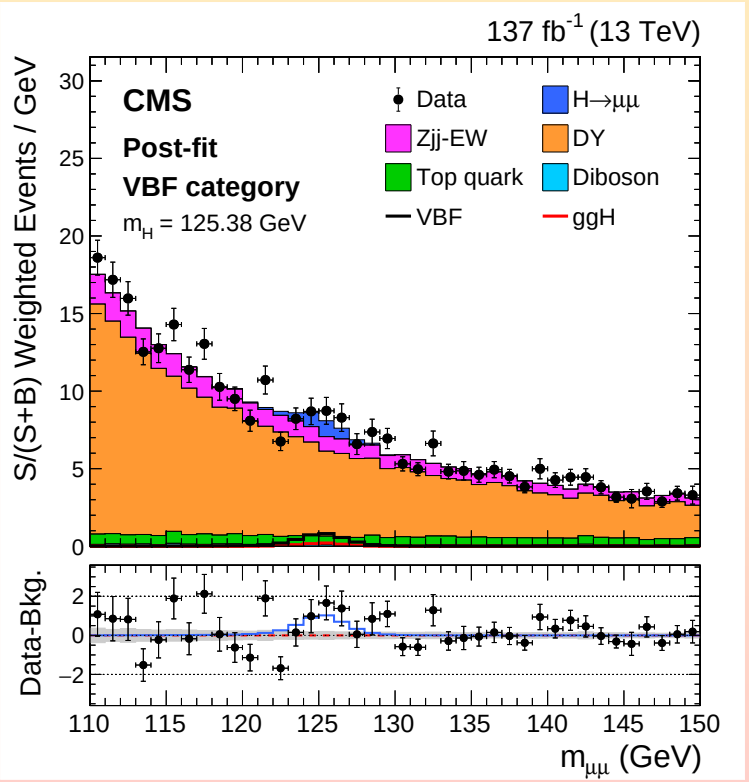



(b) Biochemical network

# How do we approach likelihood intractability in HEP?

- We usually find “by instinct” powerful summary statistics
- Select a few problem-driven observables (e.g. mass, decay angles, other kinematical variables)
- Try to lower the dimensionality of the summary statistic
  - At high dimensionality, need too many simulated events to populate phase space
- Sample the summary statistic and estimate the likelihood via density estimation...
- Or simply use **histograms**!
  - Given an histogram, write down the likelihood as product of the poisson likelihoods in each bin
- Find MLE, construct confidence limits using likelihood ratio, etc




Brehmer, Cranmer 2020. Plots from HIG-19-006

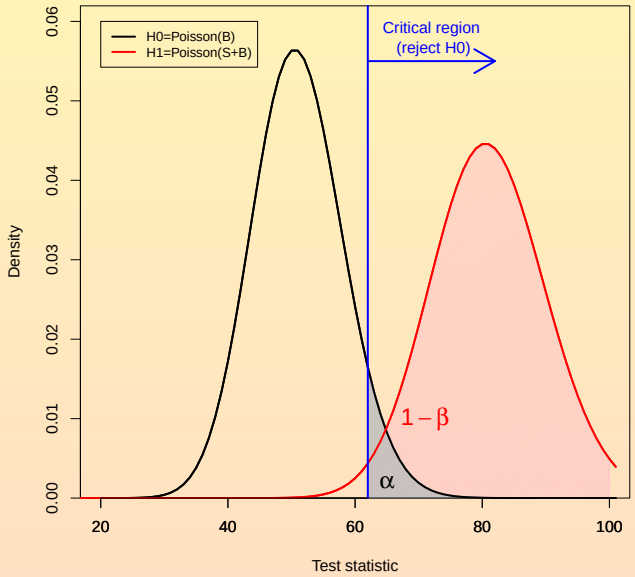
# From histograms to likelihood

- Count the amount of events in a search region
  - Usually assume they follow a Poisson distribution
- Define a function of the data (*test statistic*)
  - Can be the counts themselves: look for excesses
  - May be the (profile) likelihood

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

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

- Draw the *null* hypothesis $H_0$
  - Usually the physics we already accept as standard of the art
- Draw the *alternative* hypothesis $H_1$
  - Effect we hope to observe
- Define a critical region for rejecting the null hypothesis
- Look where does the observed value of the test statistic lie
- Likelihood ratio between the two hypothesis as the most powerful statistic (in 1D)




- Systematic uncertainties induce variations in the number of events in the search region
  - We account for them in our statistical procedures at the hypothesis testing stage
- Often machine learning techniques are employed to optimize the analysis at early stages: systematic uncertainties not accounted for in the optimization

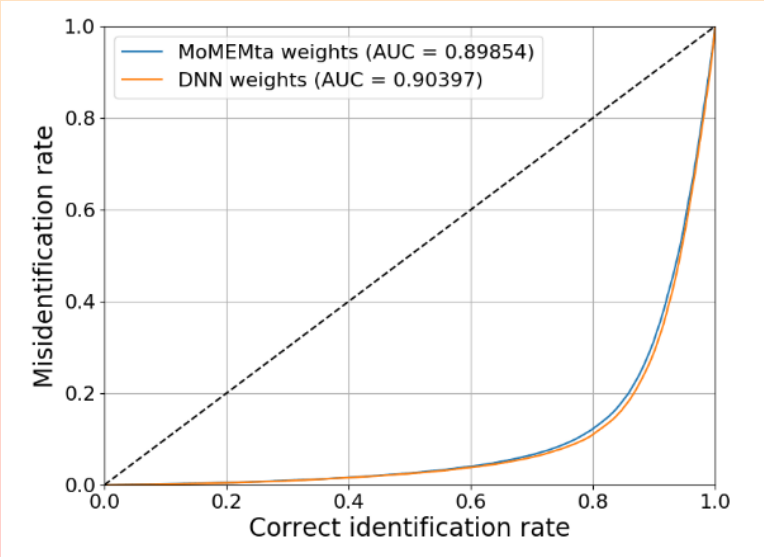
Image from P. Vischia, *XXX*
(textbook to be published by Springer in 2021)

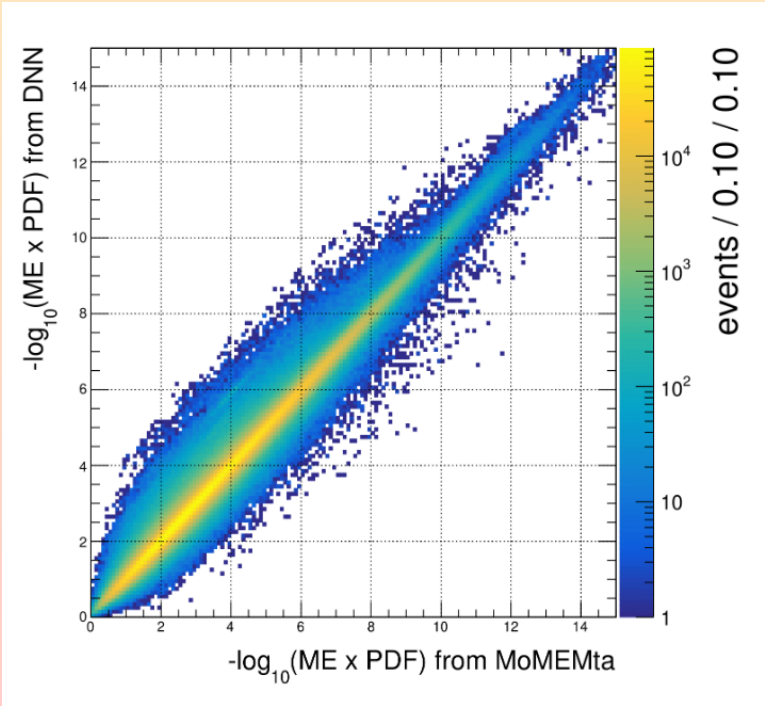
# The Matrix Element Method

- Approximate the likelihood from the precise model that includes shower and detector effects to a tractable transfer function $\hat{p}_{tf}(x|z_p)$
  - Marginal distribution simplified (no need to integrate over many microscopic interactions)
  - Approximated to a conveniente Gaussian pdf
  - Integrand is tractable, integral is over a much lower-dimensional space

$$\hat{p}_{MEM}(x|\theta) = \int \mathrm{d}z_p \; \hat{p}_{tf}(x|z_p)\, p(z_p|\theta) \sim \frac{1}{\sigma(\theta)} \int \mathrm{d}z_p \; \hat{p}(x|z_p)\, |\mathcal{M}(z_p|\theta)|^2 \,,$$

- Feasible to compute, even though till sometimes expensive
  - Also, jets are not well modelled by simple transfer functions
- Recent advancements (Florian Bury's IML2020 Workshop talk)
  - Learn MEM weights (MoMEMta: 3000 years, DNN: $\sim$10 hours)
  - Learn Matrix Element itself (DNN time comparable with MadGraph, and some fluctuations in the integrand)



- Monte Carlo generators are used to generate samples distributed according to a given likelihood function $x \sim p(x|\theta)$
  - Sometimes likelihood intractable because it depends on latent variables
    $p(x|\theta) = \int dz p(x,z|\theta) = \int dz p_x(x|\theta,z) \prod_i p_i(z_i|\theta, z_{<i})$
    (matrix element, parton shower, detector...)
- When the likelihood is intractable, inverting the problem to obtain $p(\theta|x)$ is impossible or requires huge amount of generated events or observations
  - Approximate Bayesian Computation (ABC): generate events by sampling from a prior $\pi(\theta)$, accept/reject algorithm to build the posterior
  - Probabilistic programming systems: use samples from generative model to train a tractable surrogate model
  - Tomorrow we will see advanced cases based on Machine Learning!

# Measuring differential distributions

- Combination of 2016, 2017, and 2018 results
  - Inclusive and differential measurement in reconstructed-level (folded) space
  - Differential measurement in generator-level (unfolded) space

# One-slide preview of combination of results

- When possible, combine the data
- When you cannot, then perform measurement simultaneously (e.g. combined likelihood fit)
- If you don't have access to all the data, then you have to combine results
  - Several variants of $\chi^2$, with some known pitfalls (e.g. for BLUE)
  - For all methods, combining highly correlated estimates require careful assessment of correlations
    - Assuming 100% is not always "conservative"

- Measure $N$ times the same quantity: estimates $\theta_i$ and uncertainties $\sigma_i$.
- Maximum Likelihood Estimate and variance are:

$$\hat{\theta}_{ML} = \frac{\sum_{i=1}^{N} \frac{\theta_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

$$\frac{1}{\sigma^2_{\hat{\theta}_{ML}}} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$$

- Variance linked to the amount of information about $\theta$ contained in the data set
- Usually we have access to an estimate $\hat{\sigma}_{\hat{\theta}_{ML}}$ of $\sigma_{\hat{\theta}_{ML}}$
- By construction, the best value lies inside the range of the estimates, $min(\theta_i) < \hat{\theta}_{ML} < max(\theta_i)$
- Works only with symmetric uncertainties
  - Typically assuming a Gaussian approximation of the likelihood)
  - Expression of the Central Limit Theorem
- Intrinsic difference between averaging and most probable value
  - Averaging results in average value and variance that propagate linearly
  - Taking the mode (essentially what MLE does) does not add up linearly!
- With asymmetric uncertainties, always combine the likelihoods (better if refitting all the data in an individual simultaneous fit)

## 1 parameter, N correlated measurements

- BLUE (Best Linear Unbiased Estimator)[1]

$$\hat{\theta} = \sum_{i=1}^{N} w_i \theta_i \,, \qquad \sigma_{\hat{\theta}} = \boldsymbol{w}^T \boldsymbol{V} \boldsymbol{w} \,, \qquad \boldsymbol{w} = \frac{\boldsymbol{V}^{-1} \mathbb{1}_{1\times N}}{\mathbb{1}^T_{1\times N} \boldsymbol{V}^{-1} \mathbb{1}_{1\times N}} \,, \qquad \sum_{i=1}^{N} w_i = 1$$

- Equivalent to a $\chi^2$ method, but also assigns a weight to each measurement
- Correlation between measurements can be partial or full
- Weight can be used to rank the contribution of each measurement: other options may be less desirable (doi:10.1140/epjc/s10052-014-2717-6)
  - Don't use Relative Importance RI (if doing partial combination first, $w_i$ invariant but RI not invariant)

| Weight type | | $\geq 0$ | $\Sigma_i = 1$ | Consistent with partial combination |
|---|---|---|---|---|
| BLUE coeff. | $w_i$ | ✗ | ✓ | ✓ |
| RI | $\lvert w_i \rvert / \sum_{i=1}^{n} \lvert w_i \rvert$ | ✓ | ✓ | ✗ |
| IIW | $\mathcal{W}_i$ | ✓ | ✗[*] | ✓ |
| MIW | $\mathcal{M}_i$ | ✓ | ✗ | ✓ |

Table by Luca Lista

[1] Aitken, Proc. Roy. Soc. Edinburgh 55(1935)42, Lyons et al. NIM A270(1988)110

- For two measurements $\theta_1$ and $\theta_2$,

$$\hat{\theta} = \frac{\theta_1(\sigma_2^2 - \rho\sigma_1\sigma_2) + \theta_2(\sigma_1^2 - \rho\sigma_1\sigma_2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}, \qquad \sigma_{\hat{\theta}} = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}$$

- Common uncertainty $\sigma_C = \rho\sigma_1\sigma_2$ highlights similarity to weighted average (if coefficients are positive, otherwise difficult interpretation)
  - $\hat{\theta} = \frac{\theta_1/\sigma_1'^2 + \theta_2/\sigma_2'^2}{1/\sigma_1'^2 + 1/\sigma_2'^2}$ Weighted average
  - $\sigma_{\hat{\theta}} = \frac{1}{1/\sigma_1'^2 + 1/\sigma_2'^2} + \sigma_C^2$
  - Sort of analysis of variance by extracting uncorrelated contribution to uncertainty as $\sigma_i'^2 = \sigma_i^2 - \sigma_C^2$, and common correlated one as $\sigma_C^2$
- Negative coefficients are usually hint of high-correlation regime
  - If $\rho > \sigma_1/\sigma_2$, increasing $\rho$ further decreases uncertainties and induces negative weights
  - Uncertainty strongly depends on $\rho$: estimating $\rho$ becomes crucial (and delicate)
  - Assuming $\rho = 1$ is "conservative" only if the uncorrelated contributions to the total uncertainty dominate: otherwise, it is actually aggressive
- Different analyses of the same data are likely to result in highly-correlated results: don't try to combine them

- $\theta_1 = -1.0 \pm 0.3$
- $\theta_2 = +1.0 \pm 0.4$
- $\hat{\theta}$ outside of range for large $\rho$
- $\sigma_{\hat{\theta}}$ dramatically reduced for large $\rho$
- Peelle's Pertinent Problem (PPP)

https://indico.cern.ch/event/904488/contributions/3814643/

| $\rho$ | $\hat{\theta}$ | $\sigma_{\hat{\theta}}$ | $\sigma_{\hat{\theta}}/\hat{\theta}$ |
|---|---|---|---|
| 0.0 | -0.280 | 0.058 | -0.206 |
| 0.1 | -0.310 | 0.063 | -0.204 |
| 0.2 | -0.347 | 0.068 | -0.197 |
| 0.3 | -0.393 | 0.074 | -0.187 |
| 0.4 | -0.455 | 0.079 | -0.173 |
| 0.5 | -0.538 | 0.083 | -0.154 |
| 0.6 | -0.660 | 0.087 | -0.132 |
| 0.7 | -0.854 | 0.090 | -0.105 |
| 0.8 | -1.207 | 0.089 | -0.074 |
| 0.9 | -2.059 | 0.080 | -0.039 |

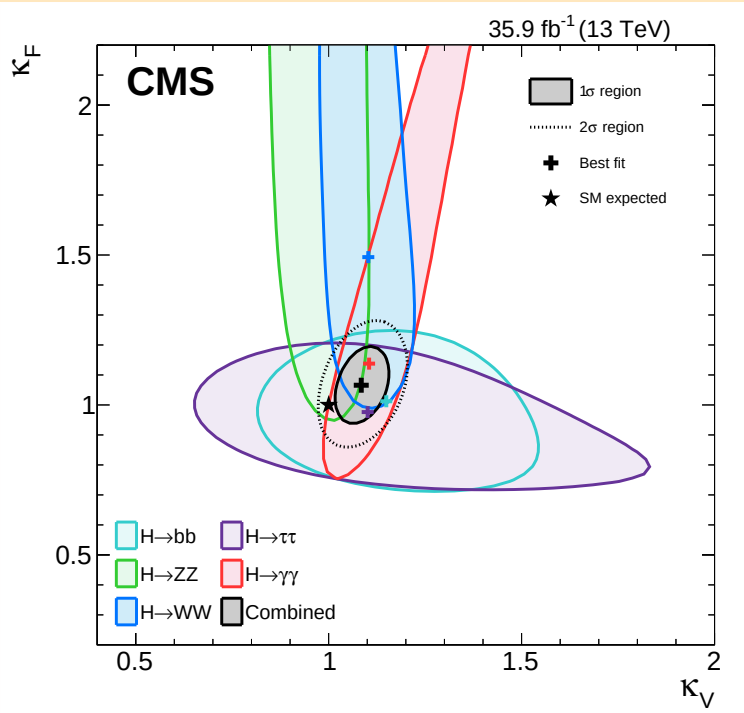
- Absolute fixed gaussian uncertainties are assumed
- Biased results in general when the estimated uncertainty depends on the numerical value of the parameter (i.e. function of the sample size, typically $\propto \sqrt{N}$)
- Biased results in particular for multiplicative factors (uncertainties relative to the assumed central value) like luminosity
- Mitigation via Iterative BLUE
  - Recompute uncertainties iteratively by rescaling them each time to the combined point estimate
  - Lyons et al, Phys Rev D41 982 (1990); Lista, NIM A764 82 (2014) and A773 87 (2015)
- In these cases (e.g. for any Poisson counting experiment) it's better to avoid the bias by avoiding BLUE

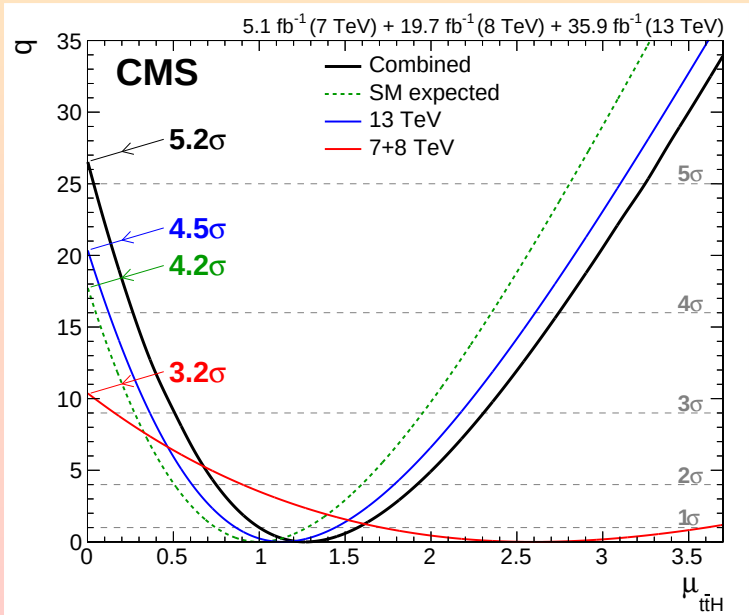
## Whenever possible, perform some kind of combined analysis

- Combine the data, if possible
- Otherwise, combine the individual likelihoods
  - Should regard BLUE as a backup solution in case the analyses don't provide the individual likelihood functions

$$\mathcal{L}(\boldsymbol{X}_1, ..., \boldsymbol{X}_N|\boldsymbol{\theta}) = \prod_i \mathcal{L}_i(\boldsymbol{X}_i|\boldsymbol{\theta}_i)$$

- $\boldsymbol{\theta}_i$ are vectors of a mix of experiment-specific and common parameters
- Accomodates any correlation scheme, supports asymmetric uncertainties
  - In high-correlations regime, as with BLUE the combined estimate might not lie inside the range of the individual ones
- Once you have the combined likelihood, you can then use it to fit for $\boldsymbol{\theta}$ simultaneously to all the data




Plots from Higgs coupling combination (HIG-17-031) and ttH observation (HIG-17-035)

- E.g. straight line fit $y = ax + b$
  - Correlation within measurement (i.e. among paramers)
  - No correlation across different measurements
- $\chi^2$ can be used to obtain the best combined value
  - Uncertainties on the best combined values $a_{best}$ and $b_{best}$ can be much smaller than individual uncertainties
  - The best combined values can lie outside of the range of the individual measurements
  - Can profile over a parameter considered a nuisance (e.g. $b$) to obtain profiled likelihood for the other (e.g. $\mathcal{L}_{prof}(a)$)
- Don't combine the profile likelihoods!
- Profile the combined likelihood!

## k parameters, N measurements

- Generalization for arbitrary number of parameters and arbitrary correlations within and across measurements
- Includes differential measurements
- $\chi^2$ approach, with non-zero off-diagonal elements in the covariance matrix

# When neither BLUE nor likelihood are feasible

- Combined likelihood cannot be performed if the individual likelihoods are not available
- BLUE cannot be always safely used
  - Requires symmetric absolute uncertainties and yields a biased otherwise
  - Result and uncertainty strongly depend on accurate estimate of correlations (cannot assume 100%)
- $\chi^2$ fits!

- Convino (J. Kieseler, 10.1140/epjc/s10052-017-5345-0)
- Equivalent to a direct likelihood combination
  - Assumes likelihood can be approximated as sum of $\chi^2$ contributions
  - Need only public material (Hessian, additional uncertainties, central values)

$$-2\ln L = \chi^2 = \sum_{\alpha} \left(\chi^2_{s,\alpha} + \chi^2_{u,\alpha}\right) + \chi^2_{p\cdot}$$

Measurement - results - - constraints - - correlations -

Correlation assumptions between measurements and priors

- HAverager (from HERAAverager, now used in ATLAS. https://github.com/HAverager/HAverager)
- Explicit $\chi^2$ minimization
- Proper treatment of correlations
- Iterative approach for relative uncertainties
- Supports asymmetric uncertainties

$$\chi^2(\vec{m},\vec{b})_{exp} = \sum_i \frac{(m_i - \mu_i - \sum_j f_i(b_j))^2}{\Delta_i^2} + \sum_j b_j^2.$$

$$f_i(b_j) = \Gamma_i^j b_j + \omega_i^j b_j^2,\ \omega_i = \frac{\Gamma_i^{j+} + \Gamma_i^{j-}}{2}.$$

- Combined likelihood
  - RooStats https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome and HistFactory https://cds.cern.ch/record/1456844
  - Theta http://www-ekp.physik.uni-karlsruhe.de/ ott/theta/theta-auto/
  - Higgs Combination Tool http://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/
- $\chi^2$-based tools (when full likelihood not available, but correlation scheme complicated / asymmetric uncertainties
  - HAverager https://github.com/HAverager/HAverager (FORTRAN with Python interface)
  - Convino https://github.com/jkiesele/Convino/
  - Other tools accounting for constraints (also non-linear) linked at https://twiki.cern.ch/twiki/bin/viewauth/CMS/StatComCombination
- BLUE and Iterative BLUE
  - Python: http://agiamman.web.cern.ch/agiamman/blue/
  - C++: http://blue.hepforge.org/

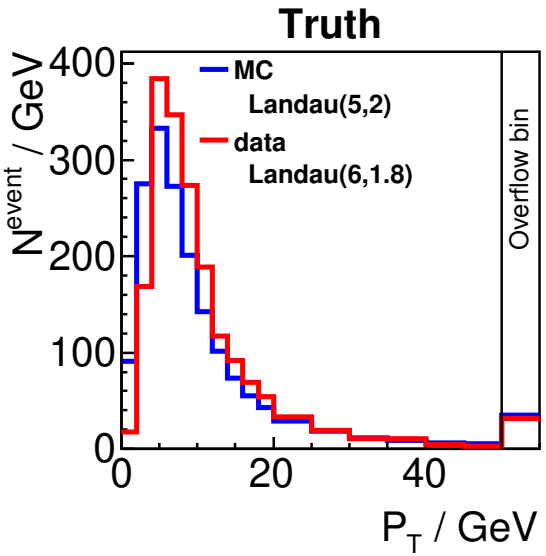
# One-slide preview of combining unfolded results

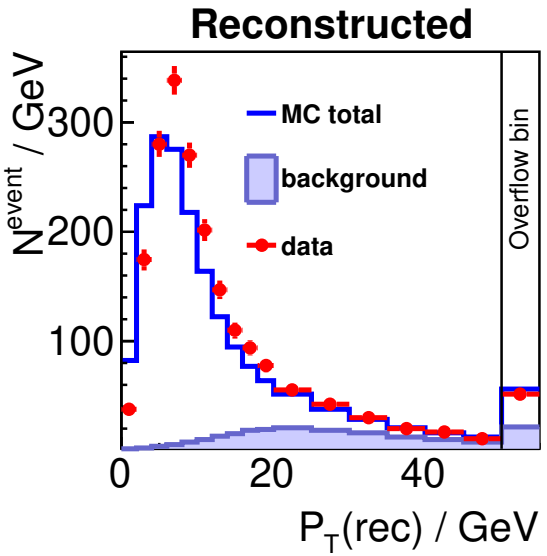


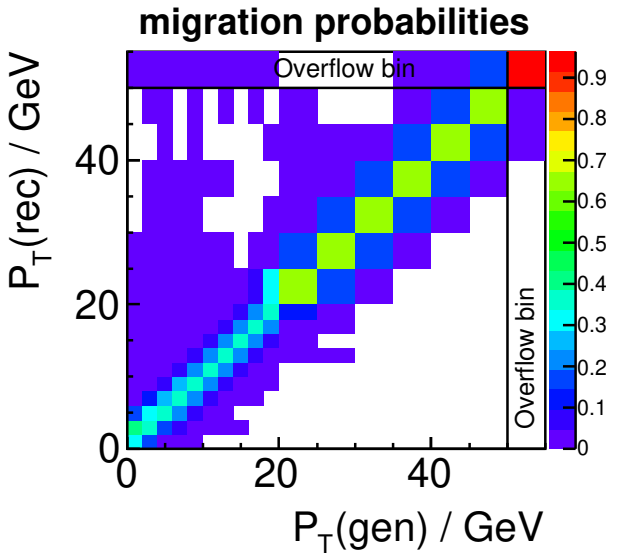
- Do you need to combine?
- If data quality (response matrix) is similar across the years, can add the data, $N = N_{2016} + N_{2017} + N_{2018}$ and unfold $N$
  - Result more stable (direct increase of statistics)
  - To be avoided if data quality is different
- If not, then unfold simultaneously the predictions for the three years (combined likelihood)
- If you don't have access to data/likelihoods, combine the three unfolded estimates
  - Don't combine regularized results (otherwise it's like applying three times the regularization)

# Unfolding: the problem

- Given the observations $y$, find a transformation to the corresponding vector $\theta$ in theory space
  - Model the detector response as a matrix of transition probabilities
  - Subtract the expected background counts: $y = n - b$
  - Invert the response to convert experimental data to distributions in the theory space
  - Inference: compare unfolded result with different models in the theory space
- **Best solution: fold any theory you want to test and make comparisons in the experimental data space**
  - Sometimes unpractical (data preservation: computing constraints or future format incompatibilities)



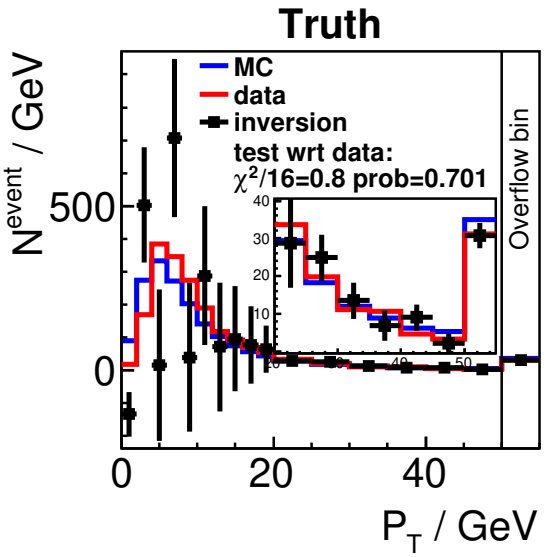

Plots by Stefan Schmitt, ArXiv:1611.01927

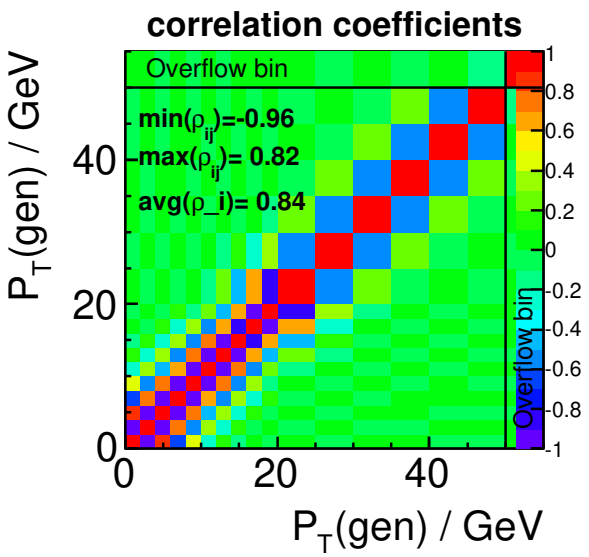
# Unfolding: naïve solutions

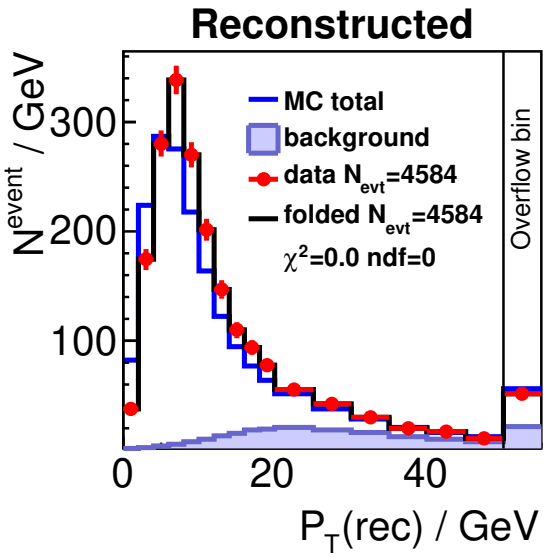
- Bin-by-bin correction factors $\hat{\theta}_i = (y_i - b_i)\frac{N_i^{\text{gen}}}{N_i^{\text{rec}}}$; disfavoured
  - Heavy biases due to the underlying MC truth
  - Yields the wrong normalization for the unfolded distribution
- Invert the response matrix $\hat{\boldsymbol{\theta}} = \boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b})$
  - Only for square matrices, $N_{reco} = N_{gen}$, but always unbiased
  - Oscillation patterns (small determinants in matrix inversion)
  - Patterns also seen as large negative $\rho_{ij} \sim -1$ near diagonal
  - Result is correct within uncertainty envelope given by $\boldsymbol{V}_{\boldsymbol{\theta\theta}}$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

determinant



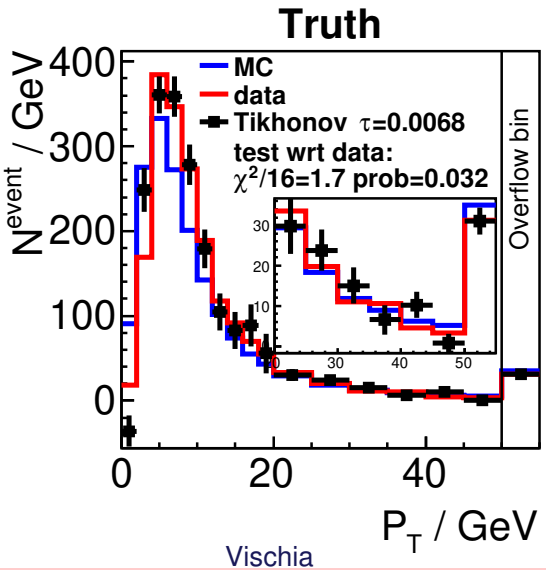



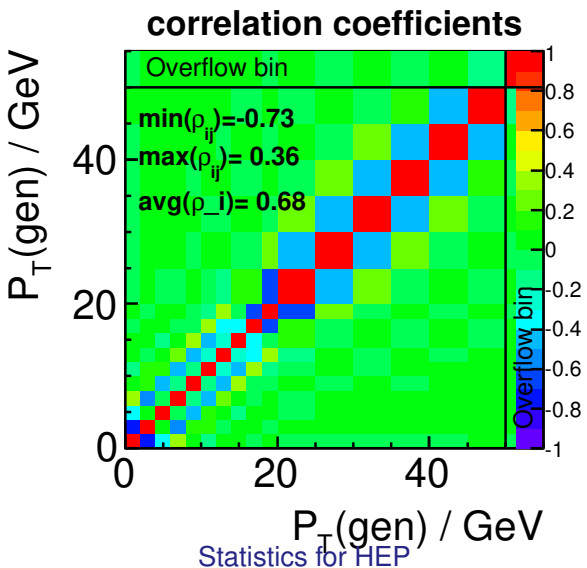

Cartoon from https://www.mathsisfun.com/algebra/matrix-inverse.html, plots by Stefan Schmitt, ArXiv:1611.01927

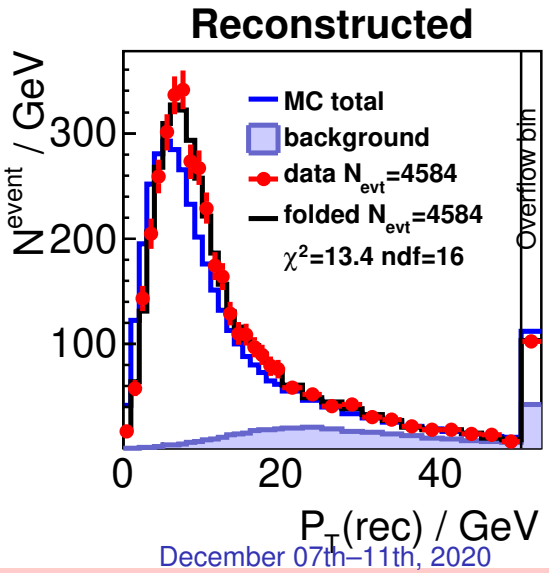
# $\chi^2$ fit with optional Tiknonov regularization (TUnfold)

$$\mathcal{L}(\mathbf{x},\lambda) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3,$$
$$\mathcal{L}_1 = (\mathbf{y}-\mathbf{A}\mathbf{x})^T \mathbf{V_{yy}} (\mathbf{y}-\mathbf{A}\mathbf{x}),$$
$$\mathcal{L}_2 = \tau^2 (\mathbf{x}-f_b\mathbf{x_0})^T (\mathbf{L}^T\mathbf{L})(\mathbf{x}-f_b\mathbf{x_0}),$$
$$\mathcal{L}_3 = \lambda(Y - \mathbf{e}^T\mathbf{x}),$$
$$Y = \sum_i y_i,$$
$$e_j = \sum_i A_{ij}.$$

- $\mathbf{y}$: observed yields
- $\mathbf{A}$: response matrix
- $\mathbf{x}$: the unfolded result
- $\mathcal{L}_1$: least-squares minimization ($V_{ij} = e_{ij}/e_{ii}e_{jj}$ correlation coefficients)
- $\mathcal{L}_2$: regularization with strength $\tau$ ($\mathbf{L}$: suppress deviation or their derivatives)
- Bias vector $f_b\mathbf{x_0}$: reference with respect to which large deviations are suppressed
- $\mathcal{L}_3$; area constraint (bind unfolded normalization to the total yields in folded space, useful for Poisson counts deviating from gaussian $\chi^2_{Neyman}$ ansatz)

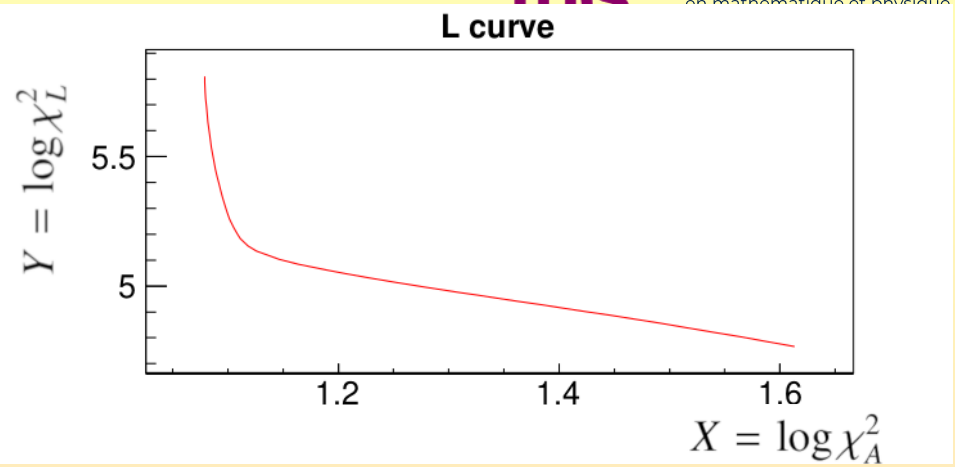


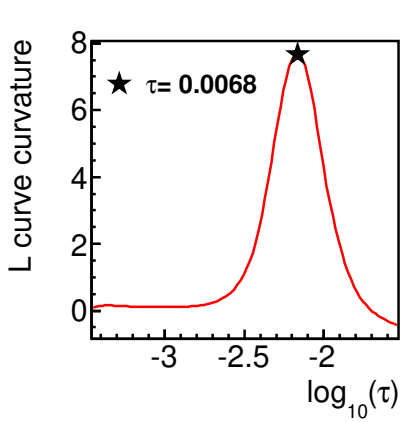



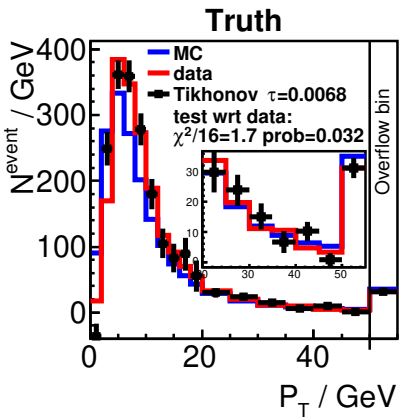
# TUnfold: choosing the best regularization

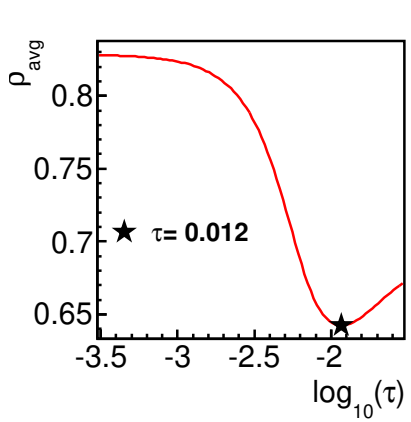
$$\chi^2_{\text{TUnfold}} = \chi^2_A + \tau^2 \chi^2_L$$
$$\chi^2_A = (\boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{b} - \boldsymbol{y})^{\mathsf{T}} (\boldsymbol{V_{yy}})^{-1} (\boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{b} - \boldsymbol{y})$$
$$\chi^2_L = (\hat{\boldsymbol{x}} - \boldsymbol{x_B})^{\mathsf{T}} \boldsymbol{L}^{\mathsf{T}} \boldsymbol{L} (\hat{\boldsymbol{x}} - \boldsymbol{x_B})$$

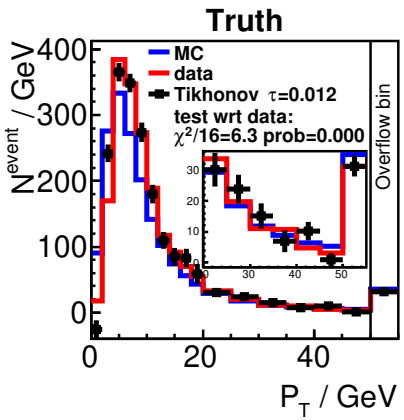



- Compute condition number from SVD of the response matrix
  - If small, $\mathcal{O}(10)$, problem is well-conditioned $\rightarrow$ **no regularization**
  - If large $\mathcal{O}(10^5)$, problem is ill-conditioned, and regularization will likely help
- Choose regularization strength $\tau$ corresponding to maximum curvature of L-curve
- Or minimize the global $\rho_{\text{avg}} = \frac{1}{M_x} \sum_{j=1}^{M_x} \rho_j$
  - Often results in stronger regularization than L-curve




Plots by Stefan Schmitt, ArXiv:1611.01927

## Bottom Line test

- Theory/data comparison should not be more discriminative in unfolded space than in folded space
  - Bob Cousins https://arxiv.org/abs/1607.07038
- Compare data-MC $\chi^2_{unfold}$ and $\chi^2_{smeared}$
- Compare $\Delta\chi^2_{unfold}$ and $\Delta\chi^2_{smeared}$ between models (should not invert hierarchy of "preferred" model)
- Strong regularization can bias too much to the bias vector (usually the MC prediction)
  - If e.g. $\chi^2_{unfold} << \chi^2_{smeared}$ maybe regularization too strong
  - If $\chi^2_{unfold} >> \chi^2_{smeared}$, unfolding bias maybe underestimated

Material from talk M. Weber, PHYSTAT 2011

| MC Generator | $\chi^2$ values between data and smeared mc | $\chi^2$ values between unfolded data and Gen mc |
|---|---|---|
| PYTHIA6 | 421 | 398 |
| HERWIG++ | 211 | 200 |
| MADGRAPH | 2590 | 2570 |
| ALPGEN | 3860 | 3860 |

- $\chi^2$ approximation of the likelihood
  - $N_{reco} = N_{gen}$: matrix inversion $\lambda = A^{-1}y$
  - $N_{reco} > N_{gen}$: $\chi^2 = [\mathbf{y} - \mathbf{A}\lambda]^T \mathbf{V_y}^{-1} [\mathbf{y} - \mathbf{A}\lambda]$
  - $\chi^2$ minimization can also work for $N_{reco} \leq N_{gen}$, but needs strong regularization or problematic anyways (reguire reference non-regularized run which will not converge)
- Systematic uncertainties: repeat full procedure with each varied response matrix A
  - Currently need to symmetrize uncertainties
- Internally maps multidimensional variables to one-dimensional vectors (can therefore regularize in multidimensional phase space)
- Simultaneous unfolding: minimize a combined $\chi^2$ to unfold in a single step the background-subtracted yield vectors for all three years
- **Caveat:** don't unfold separately with regularization and then combine the results, because you would be counting the regularization term three times!

$$y = \begin{pmatrix} y_{16} \\ y_{17} \\ y_{18} \end{pmatrix}; \mathrm{V_y} = \begin{pmatrix} V_{16,16} & V_{16,17} & V_{16,18} \\ V_{16,17} & V_{17,17} & V_{17,18} \\ V_{16,18} & V_{17,18} & V_{18,18} \end{pmatrix}; \quad \mathrm{K} = \begin{pmatrix} K_{16} \\ K_{17} \\ K_{18} \end{pmatrix}$$

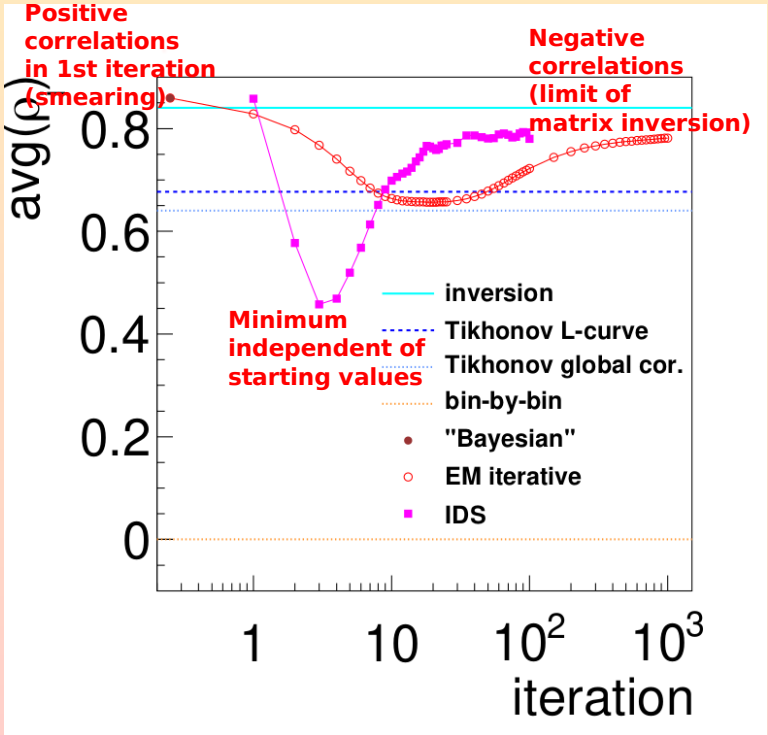
Image by Olaf

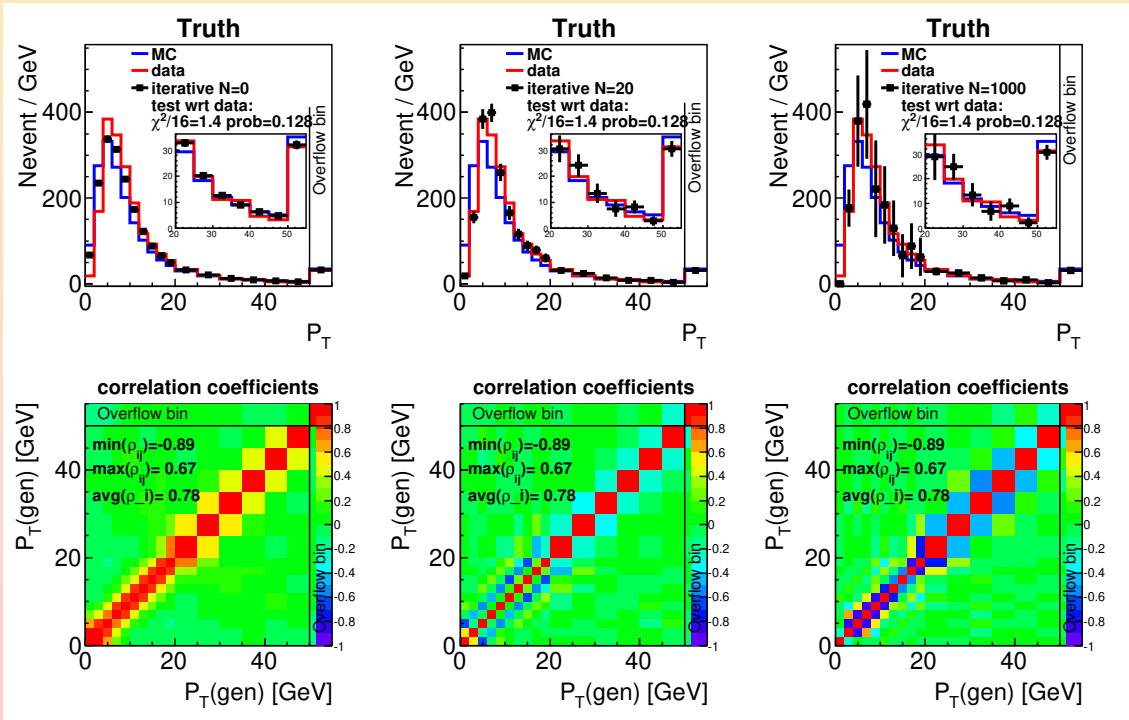
# D'Agostini (Iterative) Unfolding

- Use Bayes theorem to invert the problem
- Iterative improvement over the result of a previous iteration;

$$x_j^{(n+1)} = x_j^{(n)} \sum_{i=1}^{M} \frac{A_{ij}}{\epsilon_j} \frac{y_i}{\sum_{k=1}^{N} A_{ik} x_k^{(n)} + b_i}$$

  - It converges (slowly, $N_{iter} \sim N_{bins}^2$) to the MLE of the likelihood for independent Poisson-distributed $y_i$
  - Not necessarily unbiased for correlated data (does not make use of covariance of input data $V_{yy}$)
- Intrinsically frequentist method
  - for $N_{iter} \to \infty$ converges to matrix inversion, if all $\hat{x}_j$ from matrix inversion are positive
- In HEP many people don't iterate until convergence
  - Fixed $N_{iter}$ is often used; the dependence on starting values provides regularization
- Don't use software defaults!!! (e.g. some software has $N_{iter} = 4$)
  - Minimizing the global $\rho$ is a good objective criterion, but there are others (Akaike information, etc)

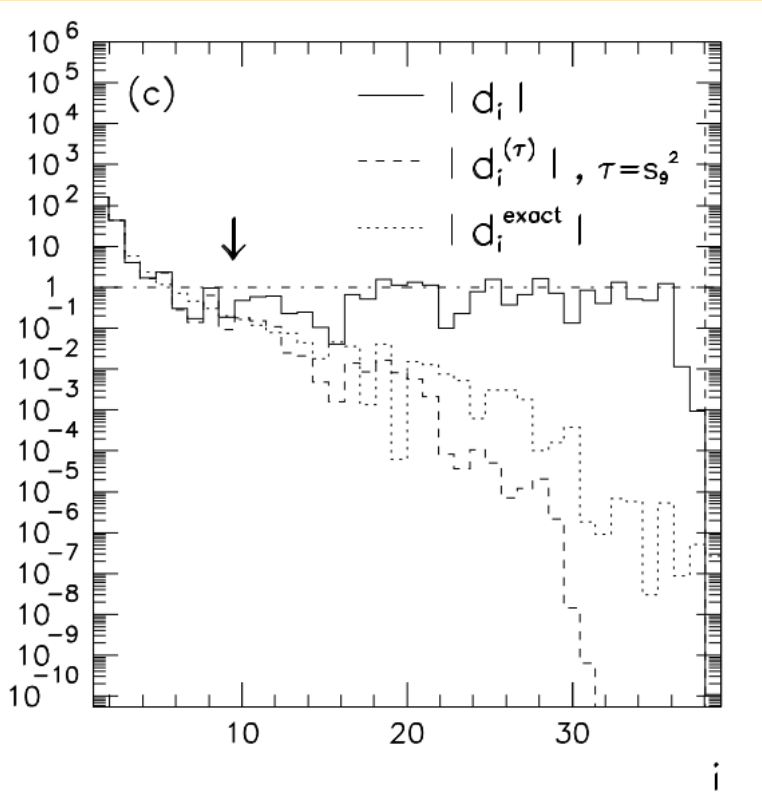



Plots by Stefan Schmitt, ArXiv:1611.01927

## Other methods (in RooUnfold)

- Bin-by-bin (severely discouraged)
- Iterative (D'Agostini)
- Singular Values Decomposition (similar to chi2 tunfold, but less flexible): regularization by eigenvalues of the response matrix
  - Höcker, Kartvelishvili https://doi.org/10.1016/0168-9002(95)01478-0
  - Requires $N_{reco} = N_{gen}$
  - For smooth distributions, only the first few $k$ SV are statistically significant
  - Use the last significant (pull incompatible with zero) SV to define regularization strength
- RooUnfold (Adye, https://arxiv.org/abs/1105.1160) per-se discouraged (at the moment, it does not support custom bias vector)



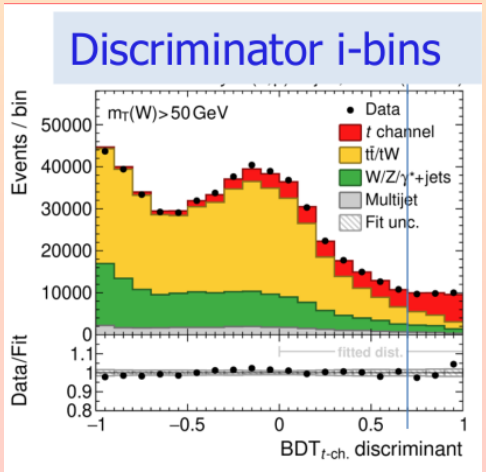

Plot from https://doi.org/10.1016/0168-9002(95)01478-0

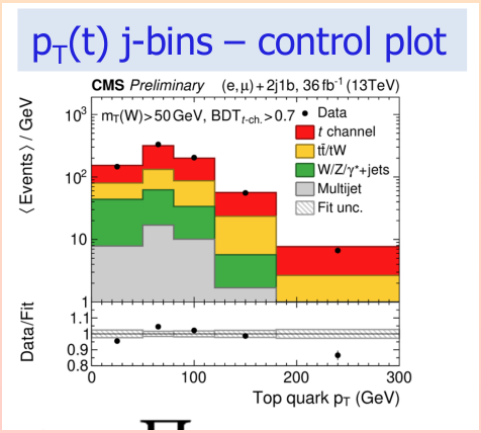
# Combining data taking eras for background-dominated analyses

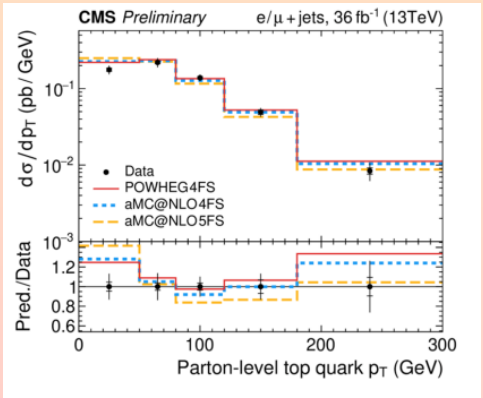
- Estimate bin-wise signal yields $y_j$ in template fit to discriminator distribution
- Usual maximum likelihood fit for signal fractions

$$\mathcal{L} = \prod_{j\ bins} \prod_{i\ bins} exp[-(y_j s_{ij} + b_{ij})] \cdot (y_j s_{ij} + b_{ij})^{n_{ij}} \cdot \prod Constraints$$

- All $y_i$ fitted simultaneously, then TUnfold (general strategy for background-dominated analyses)
  - $b_{ij}$ depends on all background normalizations, fitted for each bin $j$
  - $s_{ij}$ and $b_{ij}$ depend also on nuisance parameters, that are also fitted
- Combine eras of data taking: write combined likelihood as $\mathcal{L} = \mathcal{L}_{2016} \times \mathcal{L}_{2017} \times \mathcal{L}_{2018}$, fit the three yields simultaleously, and feed them to TUnfold
- Fitting in reco space assumes the Standard Model!!!
  - Loose capability of comparing with BSM models in generator level space



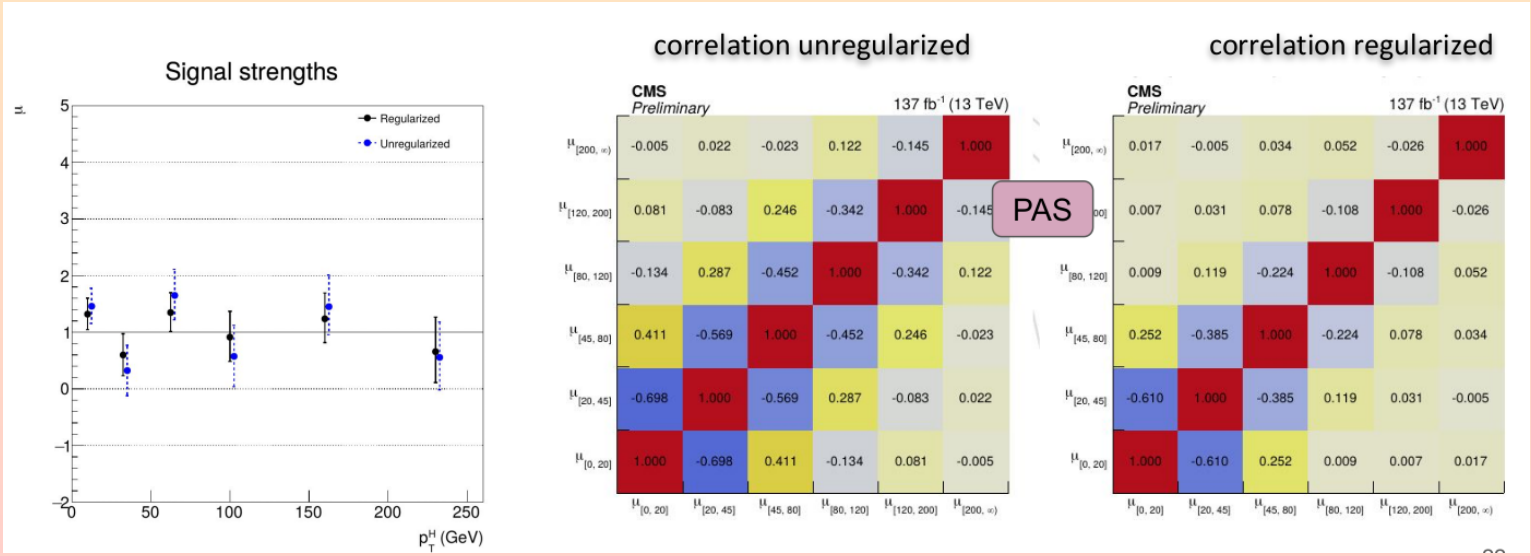




Plots from Olaf and TOP-17-023 (t-channel single top production)

## Combining data taking eras when number of unfolded distributions is small

- Maximum likelihood fit based on Higgs combination tool https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/part3/regularisation
  - Now also L2 (“Tikhonov”) regularization is included
  - Based on Higgs combination tool datacards, with some minor tweaks
  - Computationally slower due to numerical minimization (TUnfold much faster)
  - So far tried simultaneous unfolding of up to about 13 differential cross sections
- Response matrix elements and background predictions can depend on nuisance parameters, and all related tools of `combine` can be used out of the box
- Also applicable in case of pre-unfolding template fit vs discriminator bins $i$
- Combination across the years done by using combined likelihood $\mathcal{L} = \mathcal{L}_{2016} \times \mathcal{L}_{2017} \times \mathcal{L}_{2018}$

$$\mathcal{L} = \prod_{j\ bins} exp\Big[ -\Big( \sum_{m\ bins} K_{jm}\lambda_m + b_j \Big)\Big] \cdot \Big( \sum_{m\ bins} K_{jm}\lambda_m + b_j \Big)^{n_j} \cdot \prod Constraints$$



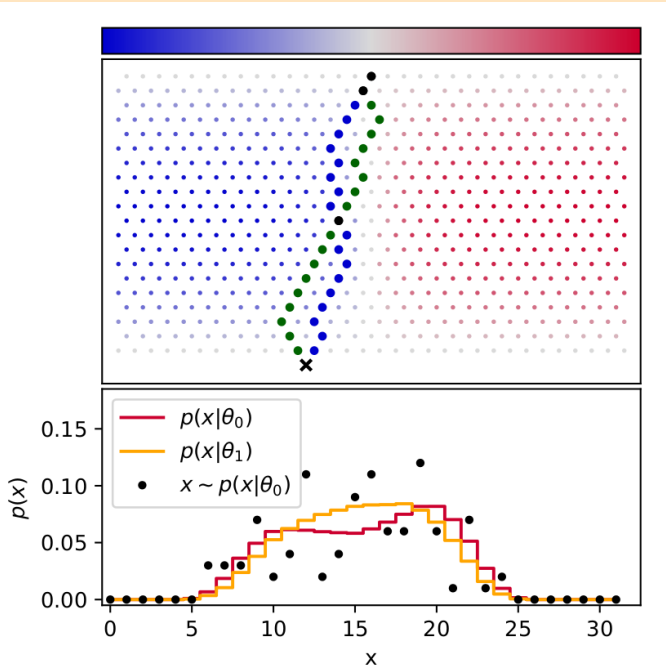

Plots from HIG-19-002

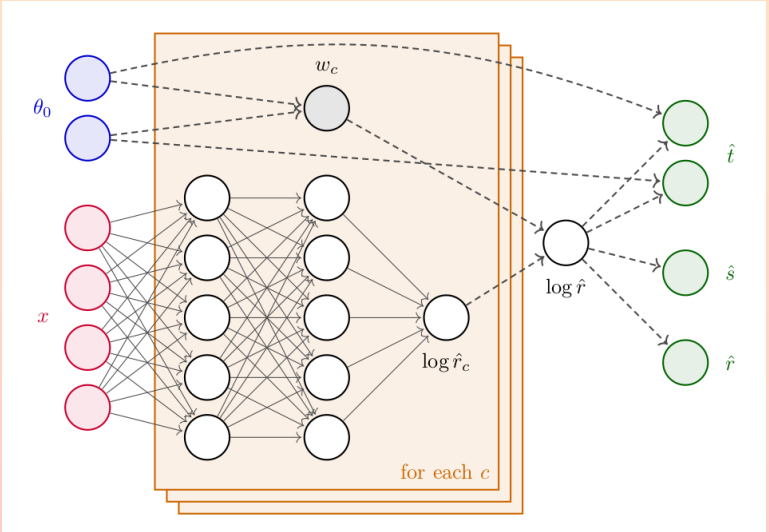
- If possible, combine the data
  - Only if data quality (response matrices) similar across the years
- Combinations based on building the total $\chi^2$ or the total likelihood
- Different strategies for different scenarios, all extensible to combination of data taking eras
  - **High stat, low background:** TUnfold after background subtraction
  - **High background:** TUnfold after template fit of detector-level signal yields $y_j$
  - **Small number of unfolded distributions:** maximum likelihood fit for $\hat{\lambda}$
- If you have to combine unfolded result, make sure they are not individually regularized
  - Otherwise equivalent to applying regularization several times
- Regardless of the scheme, be careful on how do you assess the correlations between systematic uncertainties across the years
  - A few POGs implemented mixed correlated/decorrelated components

- Combining results in the reconstructed-level space: inclusive and differential
  - Maximum likelihood fits are best if likelihoods and data are available
  - $\chi^2$ between correlated results when likelihoods not available
  - Simplified combination with BLUE when uncertainties are gaussian and correlation not too large
- Combining results in the generator-level space: differential
  - If data quality is constant, sum the data and unfold the sum
  - If using regularization, make sure you don't combine regularized results (triple counting!): eventually regularize the combined likelihood
  - Low background: simultaneous $\chi^2$ unfolding of the three years' yields
  - High background: template fit, followed by simultaneous $\chi^2$ unfolding
  - Low number of differential distributions to combine: maximum likelihood fit
- Take home messages
  - If possible, perform a combined measurement (either sum the data or build a combined likelihood)
  - $\chi^2$ valid substitute of likelihood when individual likelihoods not available or in case of computational constraints
  - Simplified methods like BLUE require precise measurement of correlation coefficient when it's close to 0 or 1
  - Correlation between the uncertainties across the years important for unfolding

## What if we don't have a likelihood?

- Likelihood $p(x|\theta) = \int dz p(x,z|\theta) = \int dz p_x(x|\theta,z) \prod_i p_i(z_i|\theta, z_{<i})$
  - Latent states sampled from $z_i \sim p_i(z_i|\theta, z_{<i})$
  - Final output sampled from $x \sim p_x(x|\theta, z)$
  - Observables $x$ from particle generator; dependency on latent $z$s (matrix element, parton shower, detector...)
- Want to do inference in $\theta$ given a $p(x|\theta)$ which is intractable; likelihood trick;
  - Train a classifier (NN) to separate samples from $p(x|\theta_0)$ and $p(x|\theta_1)$
  - Likelihood ratio between $\theta_0$ and $\theta_1$ by inverting the minimization of the binary cross-entropy loss
- Joint score $t(x,z|\theta_0)$ and likelihood ratio $r(x,z|\theta_0,\theta_1)$ computable from simulated samples
  - Train parameterized estimators, then likelihood ratio is the minimum of loss function
  - Or local approximation, then the score is a sufficient statistic for inference
- Rewrite the EFT likelihood in a basis in which it is a mixture model
- Calculate the full true parton-level likelihood starting from $N$ simulated events
  - Obtain a sufficient statistic for inference; exploit all available information!
  - Inference not limited anymore by the size of the generated samples




Images from arXiv:1805.12244 and arXiv:1805.00020

# Backup