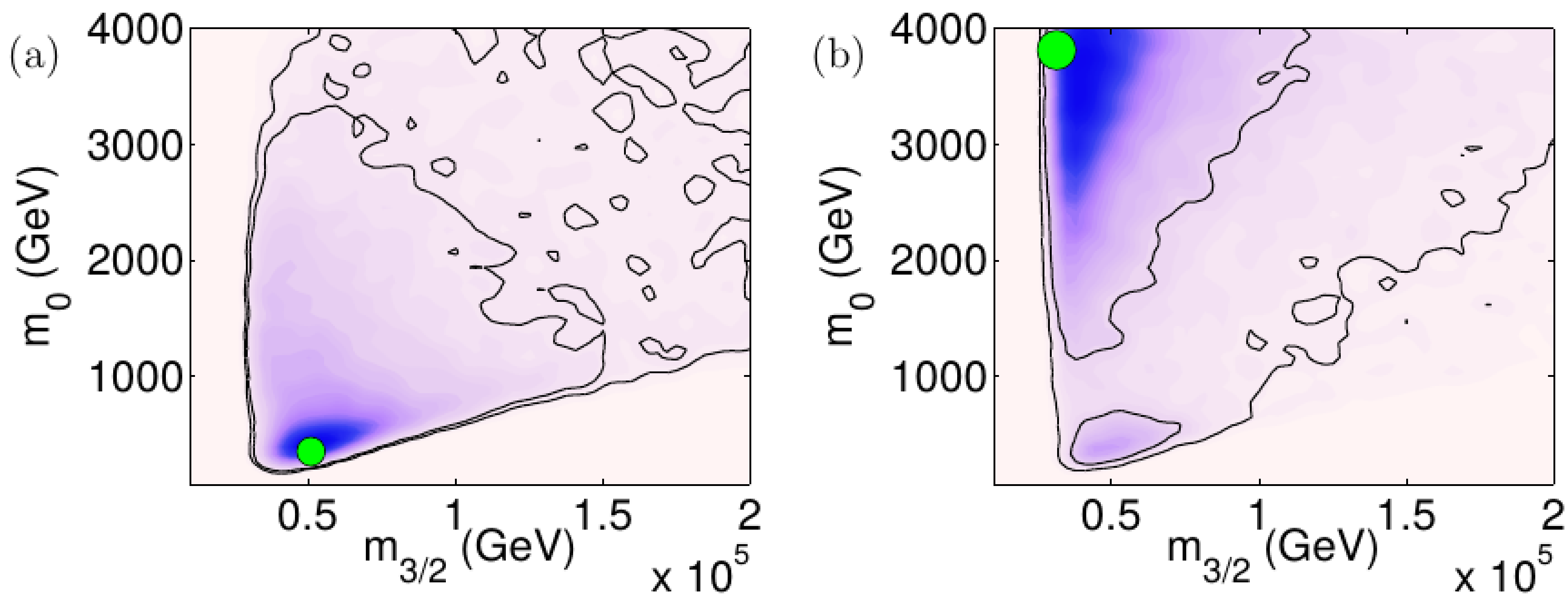
Global Fits,

or How I Learned to Stop Worrying and Love the Bayes



COEPP Summer School 2017

Matthew Dolan

University of Melbourne

CoEPP
ARC Centre of Excellence for
Particle Physics at the Terascale

THE UNIVERSITY OF
MELBOURNE

What’s a global fit?

- In particle physics we are often interested in detailed questions of data analysis such as

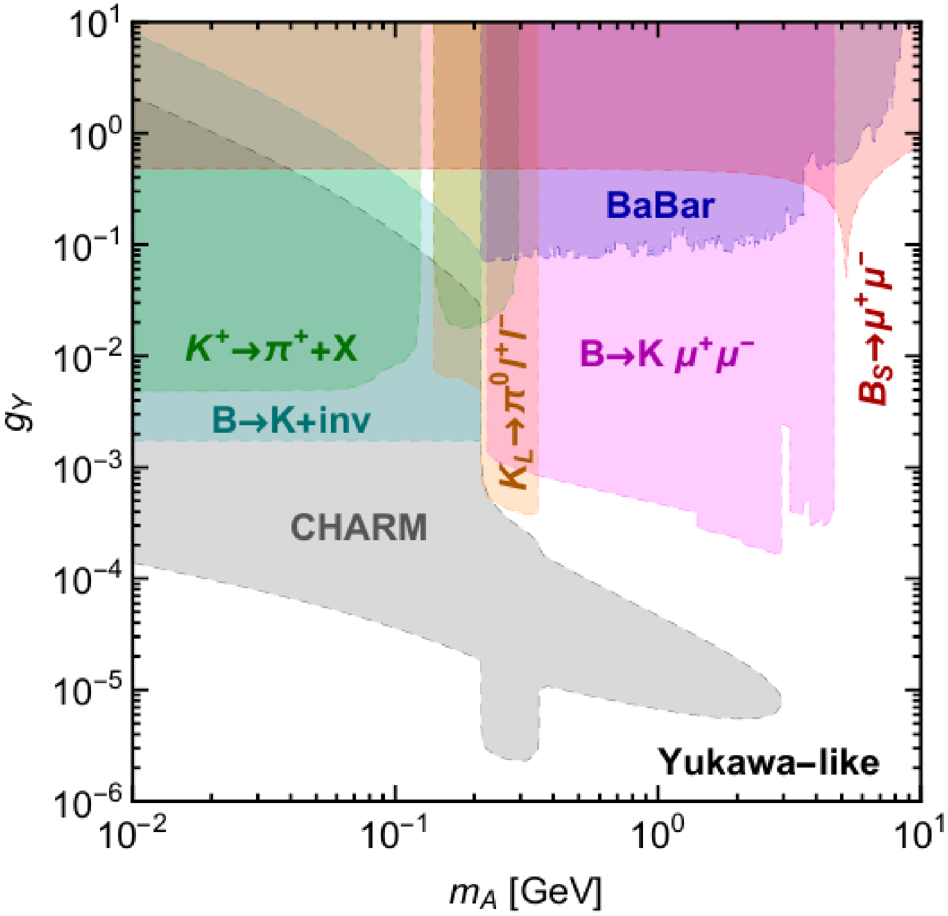
Model Comparison: What model fits the data best?

Parameter Estimation: Given a model, what values do its parameters take?

Goodness-of-fit: Given a model, is it consistent with the data?

Do we really need to fit?

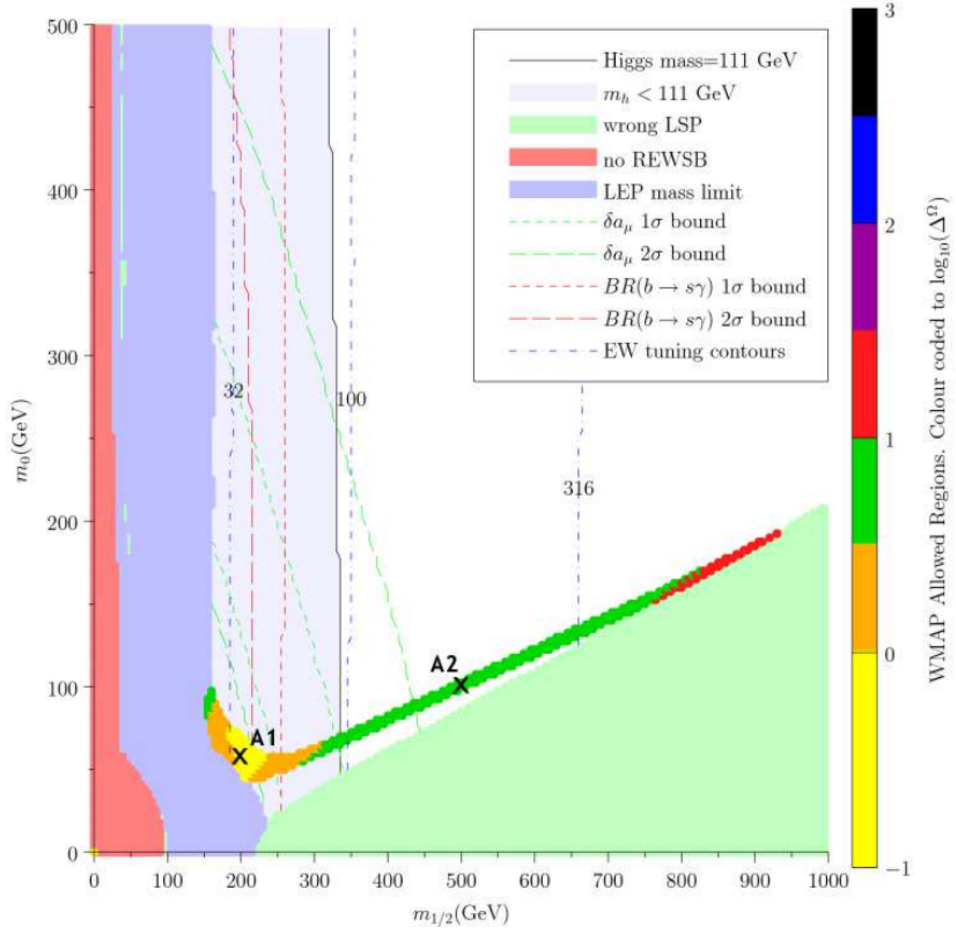
- With only 1-2 parameters could just plot constraints individually
- Example shows flavour constraint on a light pseudoscalar
- Only parameters are mass and coupling
- But what if we had 4, or 12 parameters?



Dolan et al, 2014

Do we really need to fit?

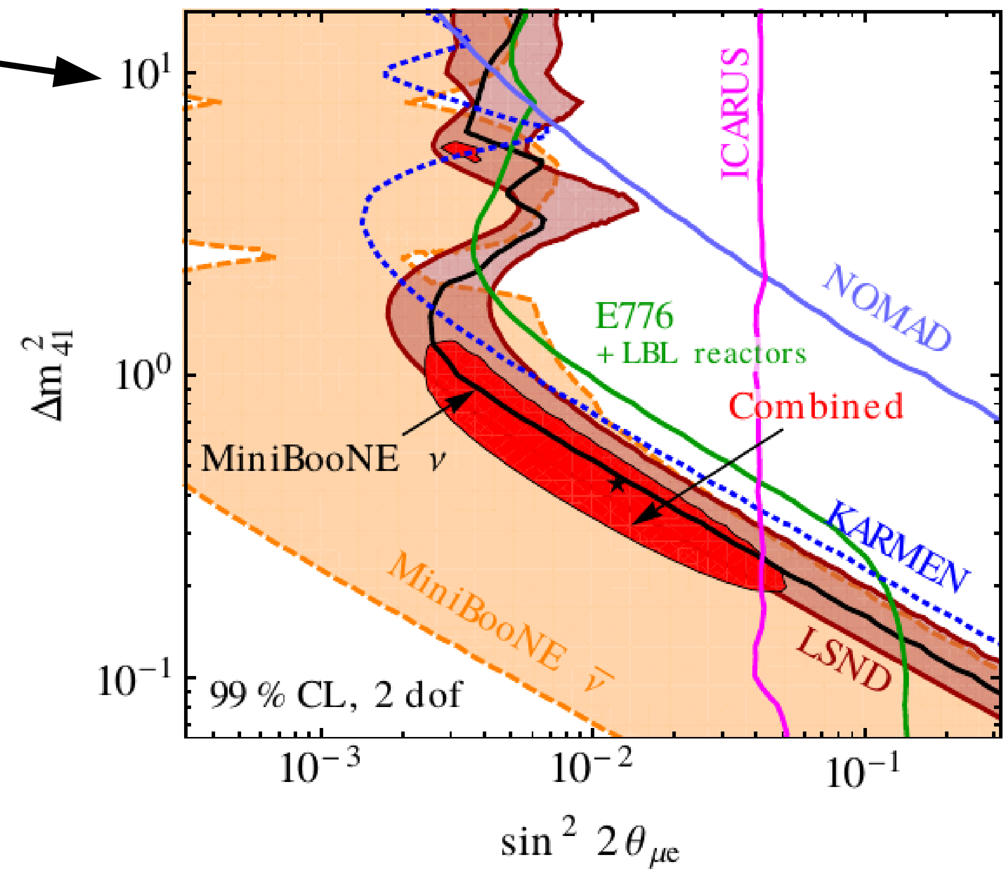
- Previous practice: show 2D planes and fix the other parameters to some (arbitrary) values.
- This was quite popular in BSM studies (example is for a SUSY model with 4 params).
- But it is clearly inadequate: we should combine constraints statistically.



Ellis et al, 2007

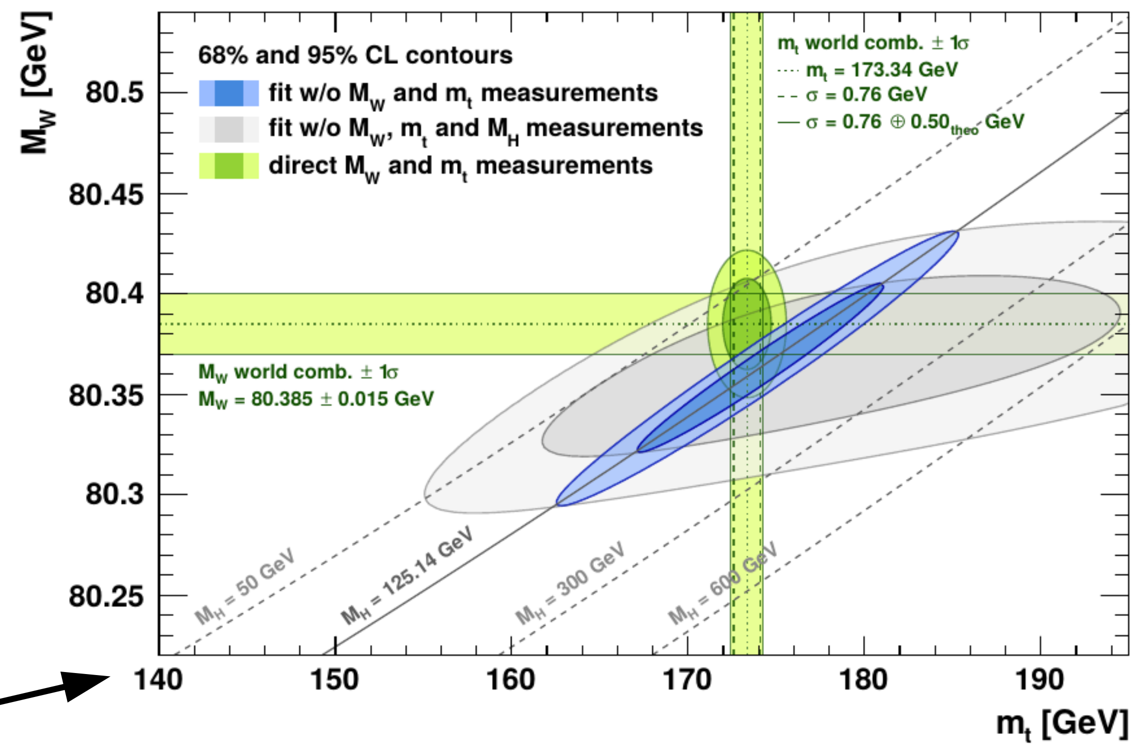
Examples of global fits.

- Sterile neutrinos
  Kopp et al, 2013
- DM Direct Detection
  Kopp, Schwetz, Zupan 2009
- DM Indirect Detection
  Cuoco et al, 2016
- Parton Distribution Functions
  MSTW, 2009   NNPDF, 2014
- SM Electroweak Fit
  GFitter Collaboration
- BSM: Many SUSY analyses MasterCode (Dolan), 2015
  Balasz, M. White et al, 2012



Examples of global fits.

- Sterile neutrinos
  Kopp et al, 2013
- DM Direct Detection
  Kopp, Schwetz, Zupan 2009
- DM Indirect Detection
  Cuoco et al, 2016
- Parton Distribution Function
  MSTW, 2009 NNPDF, 2014
- SM Electroweak Fit
  GFitter Collaboration



- BSM: Many SUSY analyses
  MasterCode (MJD et al), 2015
  Balasz, M. White et al, 2012

How to do a global fit?

- To combine constraints over a multidimensional space we need:

- An efficient method to explore structure of parameter space
- A statistical framework to interpret to the results
- Visualisation/prediction through marginalising/profiling down to lower dimensions

How to do a global fit?

- An efficient method to explore structure of parameter space
- A statistical framework to interpret to the results
- Visualisation/prediction through marginalising/profiling down to lower dimensions

We will use Bayesian statistics to provide an answer to these questions.

First off, how do we associate a probability to a point in a parameter space?

Bayesian Statistics

- Consider a hypothesis H, defined by model parameters m, which should describe some data D. Let's fix the hypothesis H for now.
- We want to know the probablility of some parameters, given the data.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) \neq 0,$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \text{ if } P(A) \neq 0,$$

$$\Rightarrow P(A \cap B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A),$$

Bayes' Theorem

$$\boxed{\Rightarrow P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}, \text{ if } P(B) \neq 0.}$$

Bayesian Statistics

- Consider a hypothesis H, defined by model parameters m, which should describe some data D. Let's fix the hypothesis H for now.
- We want to know the probablility of some parameters, given the data.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) \neq 0,$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \text{ if } P(A) \neq 0,$$

$$\Rightarrow P(A \cap B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A),$$

$$\boxed{\Rightarrow P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}, \text{ if } P(B) \neq 0.}$$

$$p(m|d) = p(d|m)\frac{p(m)}{p(d)},$$

Bayesian Statistics

$$p(m|d) = p(d|m)\frac{p(m)}{p(d)},$$

- The likelihood $\mathcal{L} \equiv p(d|m)$ is the probability of reproducing the data d, given a point m in the model parameter space.
- The posterior $p(m|d)$ is what we are interested in: the probability of the specific model point m, given the data d.
- The prior $p(m)$ is the probabilty of the model point m being correct.
- The Bayesian evidence is $p(d)$, the total probability of the data being reproduced, integrating over the model parameter space.
- Can now compare the relative probabilities of two points in parameter space (evidence cancels):

$$\frac{p(m_1|d)}{p(m_2|d)} = \frac{p(d|m_1)p(m_1)}{p(d|m_2)p(m_2)}.$$

Priors

$$\frac{p(m_1|d)}{p(m_2|d)} = \frac{p(d|m_1)p(m_1)}{p(d|m_2)p(m_2)}.$$

- The appearance of the prior distributions p(m) is an (in)famous feature of Bayesian statistics.
- For a Bayesian, probability is *degree of belief* or *state of knowledge* for the hypothesis to be true.
- How to detemine priors? Personal belief, theoretical considerations, results of other measurements.
- Bayesian stats provides a factorisation between the likelihood and prior, between objective and subjective.
- Often assume ratio of priors is 1: all regions of parameter space equally likely. Logarithmic priors also used.

Calculating Likelihoods

- What goes into the likelihood calculation?
- In principle, anything observable we can calculate within the model which has been measured or constrained.
- Cross-sections, branching ratios, event rates...

Observable	Constraint	Theory	Experiment
m_W[GeV]	80.399 ± 0.027	[43, 44]	[45]
$\sin^2\theta^l_{eff}$	0.23149 ± 0.000173	[43, 44]	[46]
$\delta a_\mu \times 10^{10}$	29.5 ± 8.8	[47, 48, 49, 50, 51, 52, 53, 54]	[55]
$\Omega_{\rm DM}h^2$	0.1143 ± 0.02	[47, 48, 49, 50]	[56]
m_h[GeV]	> 114.4GeV	[57]	[58]
Γ_Z^{tot}[GeV]	2.4952 ± 0.0023	[43]	[59, 60]
R_l^0	20.767 ± 0.025	[43]	[59, 60]
R_b^0	0.21629 ± 0.00066	[43]	[59, 60]
R_c^0	0.1721 ± 0.0030	[43]	[59, 60]
$A_{fb}^{0,b}$	0.0992 ± 0.0016	[43]	[59, 60]
$A_{fb}^{0,c}$	0.0707 ± 0.035	[43]	[59, 60]
$A_{LR}^0(SLD)$	0.1513 ± 0.0021	[43]	[59, 60]
$\mathcal{A}_b$	0.923 ± 0.020	[43]	[59, 60]
$\mathcal{A}_c$	0.670 ± 0.027	[43]	[59, 60]

AbdusSalam, MJD, et al 2009

Calculating Likelihoods

- Given a measurement $m_i \pm s_i$ and prediction p_i
- The likelihood is $\ln \mathcal{L}_i = -\frac{(m_i - p_i)^2}{2s_i^2} - \frac{1}{2}\ln(2\pi) - \ln s_i,$

Observable	Constraint	Theory	Experiment
m_W[GeV]	80.399 ± 0.027	[43, 44]	[45]
$\sin^2 \theta^l_{eff}$	0.23149 ± 0.000173	[43, 44]	[46]
$\delta a_\mu \times 10^{10}$	29.5 ± 8.8	[47, 48, 49, 50, 51, 52, 53, 54]	[55]
$\Omega_{\mathrm{DM}} h^2$	0.1143 ± 0.02	[47, 48, 49, 50]	[56]
m_h[GeV]	> 114.4GeV	[57]	[58]
Γ_Z^{tot}[GeV]	2.4952 ± 0.0023	[43]	[59, 60]
R_l^0	20.767 ± 0.025	[43]	[59, 60]
R_b^0	0.21629 ± 0.00066	[43]	[59, 60]
R_c^0	0.1721 ± 0.0030	[43]	[59, 60]
$A_{fb}^{0,b}$	0.0992 ± 0.0016	[43]	[59, 60]
$A_{fb}^{0,c}$	0.0707 ± 0.035	[43]	[59, 60]
$A_{LR}^0(SLD)$	0.1513 ± 0.0021	[43]	[59, 60]
$\mathcal{A}_b$	0.923 ± 0.020	[43]	[59, 60]
$\mathcal{A}_c$	0.670 ± 0.027	[43]	[59, 60]

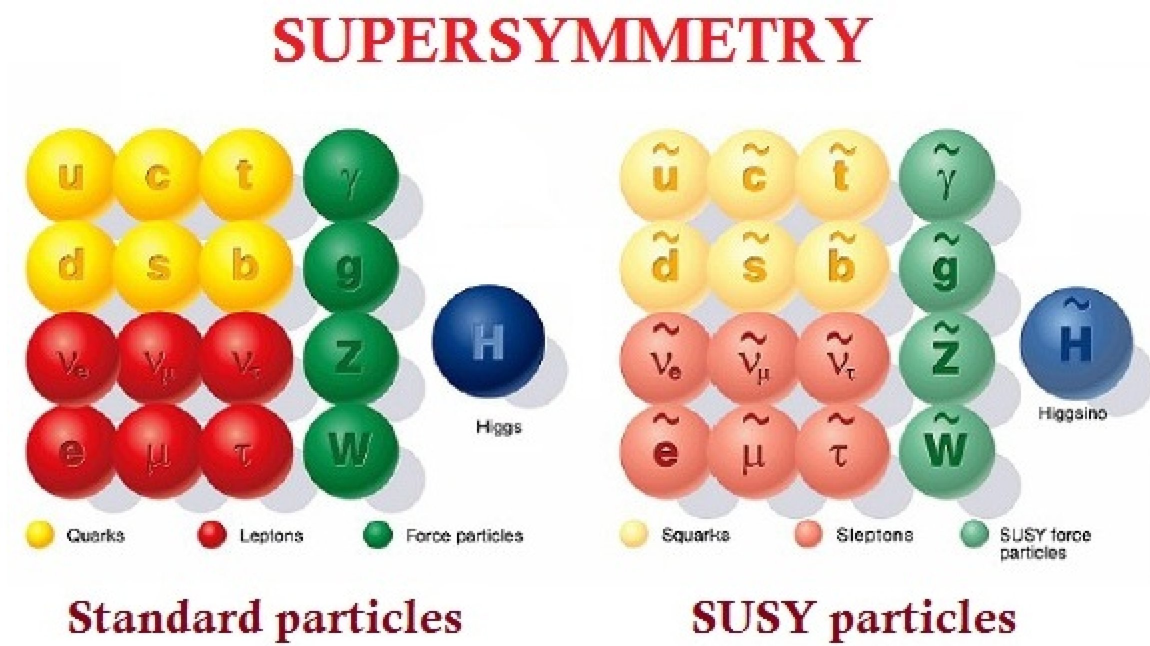
AbdusSalam, MJD, et al 2009

Example: The Constrained MSSM

Constrained Minimal Supersymmetric Standard Model:
Benchmark new physics model with 4 free parameters

$$m_0, \quad m_{1/2}, \quad A_0, \quad \tan\beta$$

These control the masses/couplings of new supersymmetric particles.

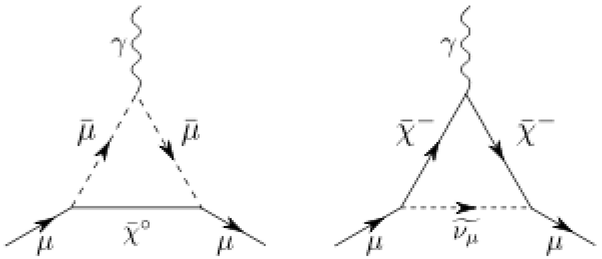


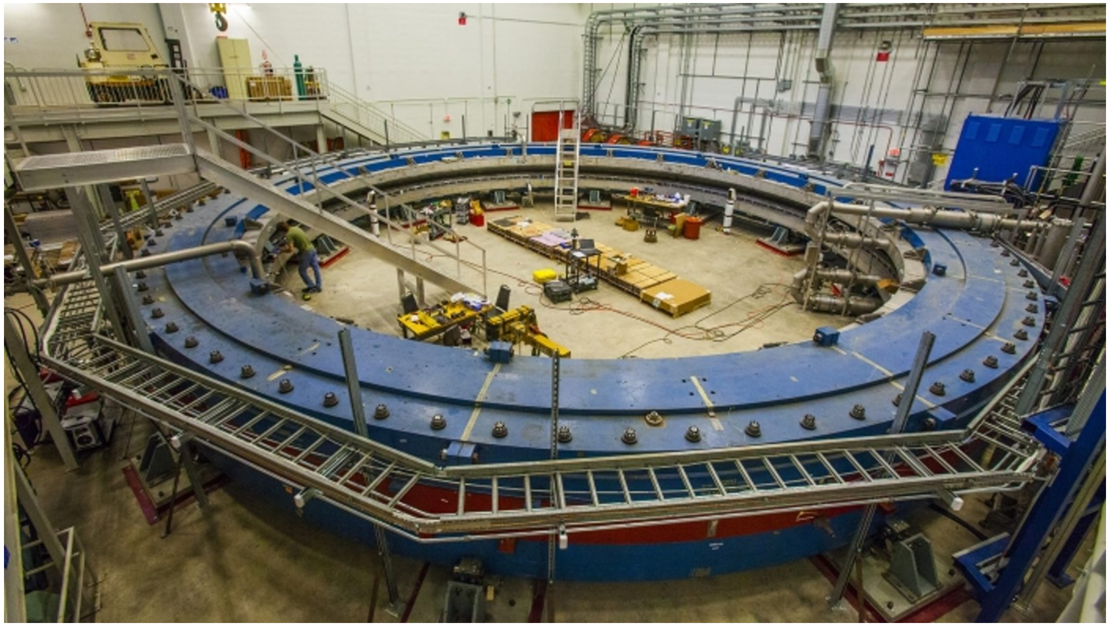
See Akula’s talk.

Example: The Constrained MSSM

Supersymmetric particles give new contributions to measured quantities:

The anomalous magnetic moment of the muon

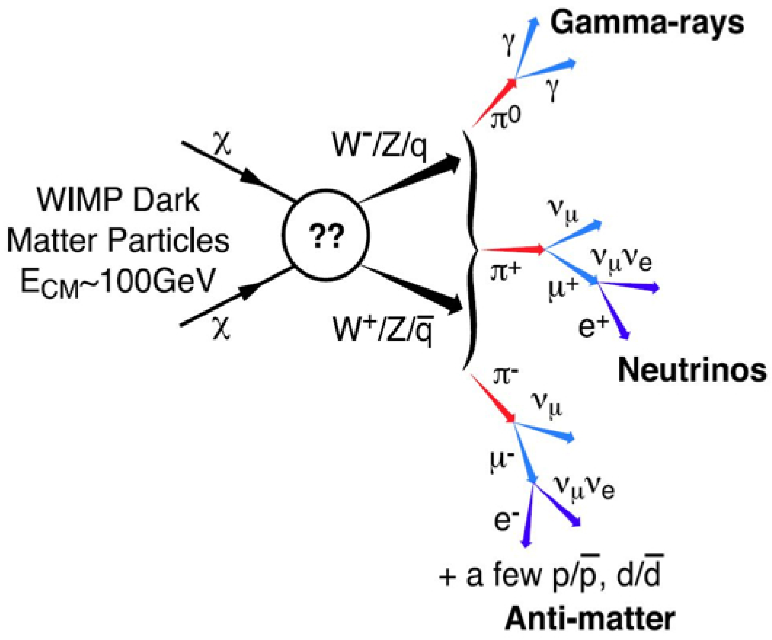




Example: The Constrained MSSM

Supersymmetric particles give new contributions to measured quantities:

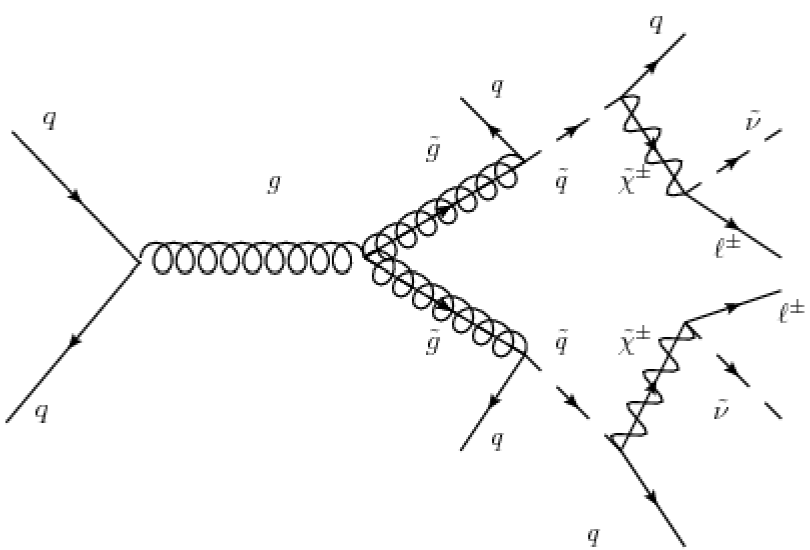
SUSY Dark matter can annihilate into SM particles in the Milky Way

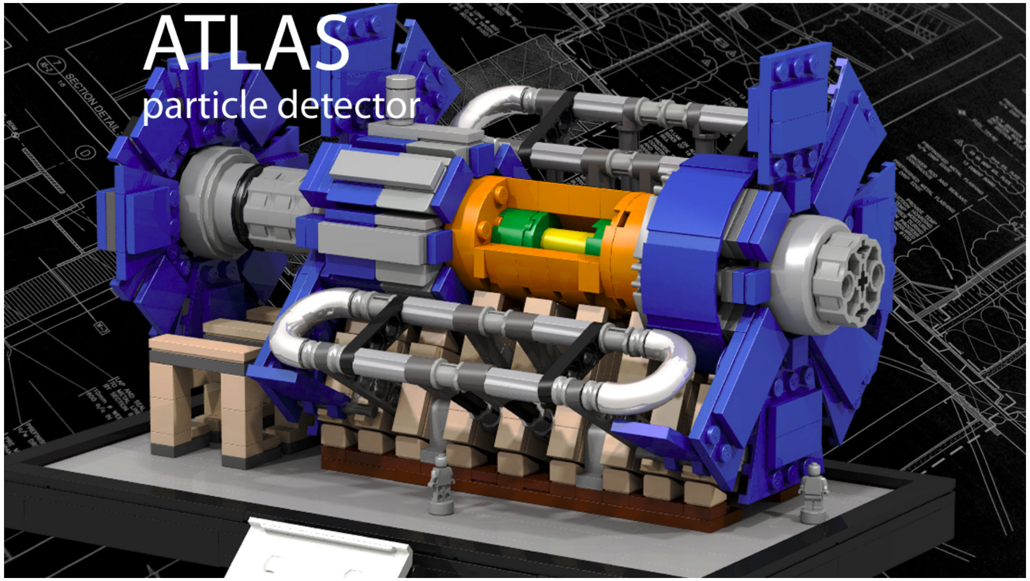


Example: The Constrained MSSM

Supersymmetric particles give new contributions to measured quantities:

SUSY particles can be created at the LHC





Exploring the Parameter Space

- So, we can construct the likelihood and hence posterior for a given point in parameter space.
- How do we efficiently sample the whole parameter space for a global picture of the posterior?
- **Linear Scanning?** Only in very few dimensions. Might miss small features
- **Random Scanning?** Same problems, and lacking in statistical meaning.
- We want to sample directly from the posterior distribution

Metropolis-Hastings Algorithm: Words

- Produces a sequence of samples which approximate the true posterior distribution.
- Samples are produced iteratively with each one depending only on the one before it – the sequence thus forms a Markov chain.

Metropolis-Hastings Algorithm: Words

- Produces a sequence of samples which approximate the true posterior distribution.
- Samples are produced iteratively with each one depending only on the one before it – the sequence thus forms a Markov chain.

- At each iteration, pick a candidate for the next point based on current one.
- With some probability, the candidate is accepted (and added to the sequence)
- Or else rejected, and we sample again
- The acceptance probability is based on the ratio of posteriors of the two points

Metropolis-Hastings Algorithm: Maths

- A **Markov chain** is a list of parameter points $(\mathbf{x}^{(t)})$ and likelihoods $(\mathcal{L}^{(t)} \equiv \mathcal{L}(\mathbf{x}^{(t)}))$, where t is an index for the link in the chain.
- Consider a point at the end of the chain $(\mathbf{x}^{(t)})$. We pick another point $(\mathbf{x}^{(t+1)})$ nearby using a proposal PDF $Q(\mathbf{x}; \mathbf{x}^{(\mathbf{t})})$.
- This is usually symmetric $Q(\mathbf{x_a}; \mathbf{x_b}) = Q(\mathbf{x_b}, \mathbf{x_a})$, and often taken to be a Gaussian.
- Calculate the likelihood for the new point.
- If $\mathcal{L}^{(t+1)} > \mathcal{L}^{(t)}$ add the point to the chain, and continue.
- Otherwise, accept the point with probability $\mathcal{L}^{(t+1)} / \mathcal{L}^{(t)}$ and (andif rejected add $(\mathbf{x}^{(t)})$ to the end).

Metropolis-Hastings: Comments

- In higher dimensions efficiency scales linearly with D (as opposed to power-law for a linear scan).
- Independent of proposal function Q for $t \to \infty$
- Still need to choose Q wisely.
- Too narrow: small step size, long convergence time.
- Too broad: hops all over parameter space too quickly.
- The proposal function should satisfy the detailed balance condition, guaranteeing convergence to the posterior

$$T(\mathbf{x}_a; \mathbf{x}_b)\mathcal{L}(\mathbf{x}_b) = T(\mathbf{x}_b; \mathbf{x}_a)\mathcal{L}(\mathbf{x}_a),$$

$$T(\mathbf{x}_b; \mathbf{x}_a) \equiv Q(\mathbf{x}_b; \mathbf{x}_a) \times \min(1, \mathcal{L}(\mathbf{x}_b)/\mathcal{L}(\mathbf{x}_a))$$

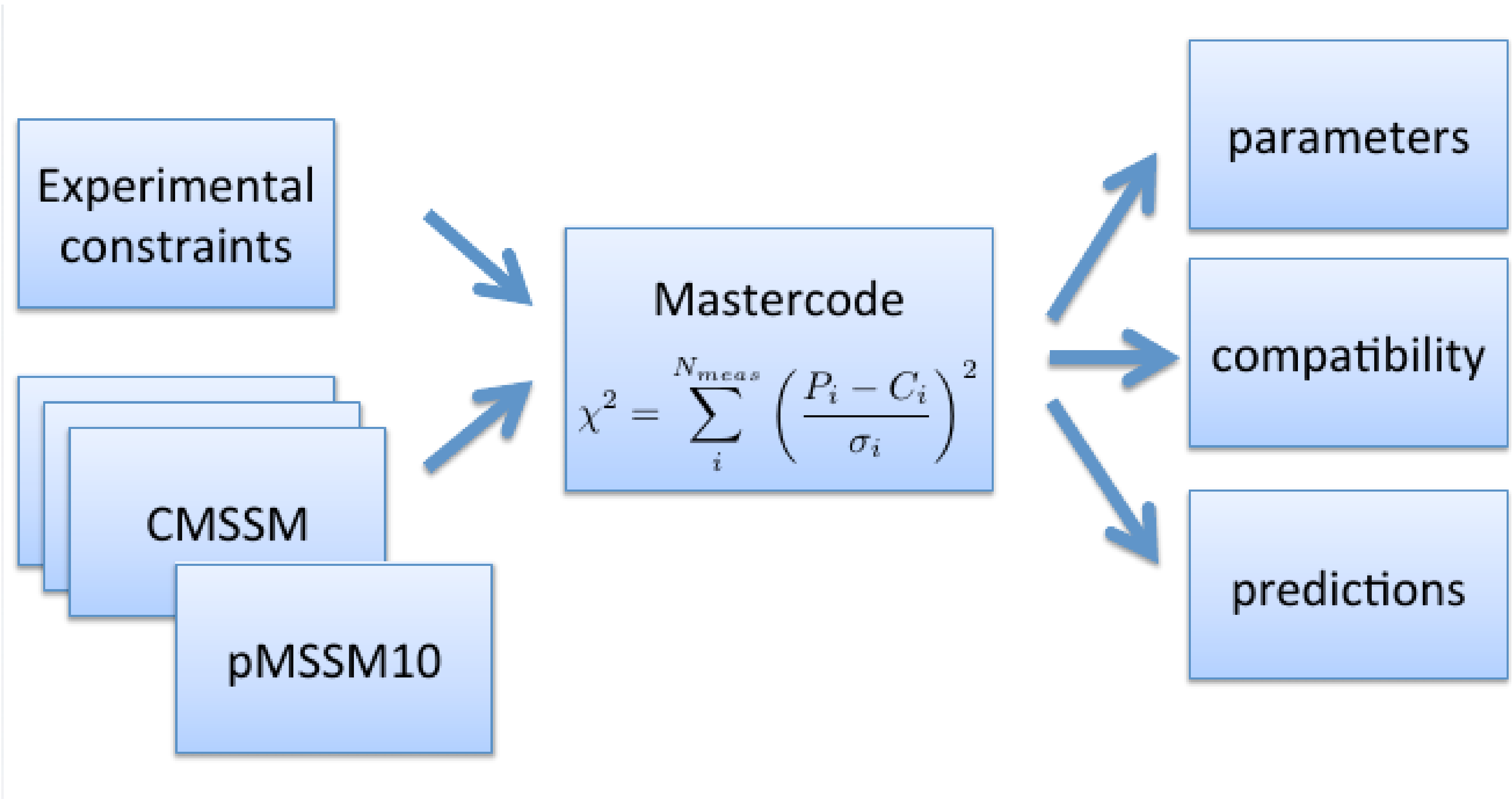
Marginalising and nuisance parameters

- How do we visualise these results from a high-dimensional parameter space?
- Integrate over the extra dimension – this is known as **marginalisation.**
- So, to get the posterior for m_0 only we would calculate

$$p(m_0|\text{data}) = \int dM_{1/2}\, dA_0\, d\tan\beta\, ds\; p(m_0, M_{1/2}, A_0, \tan\beta, s|\text{data}).$$

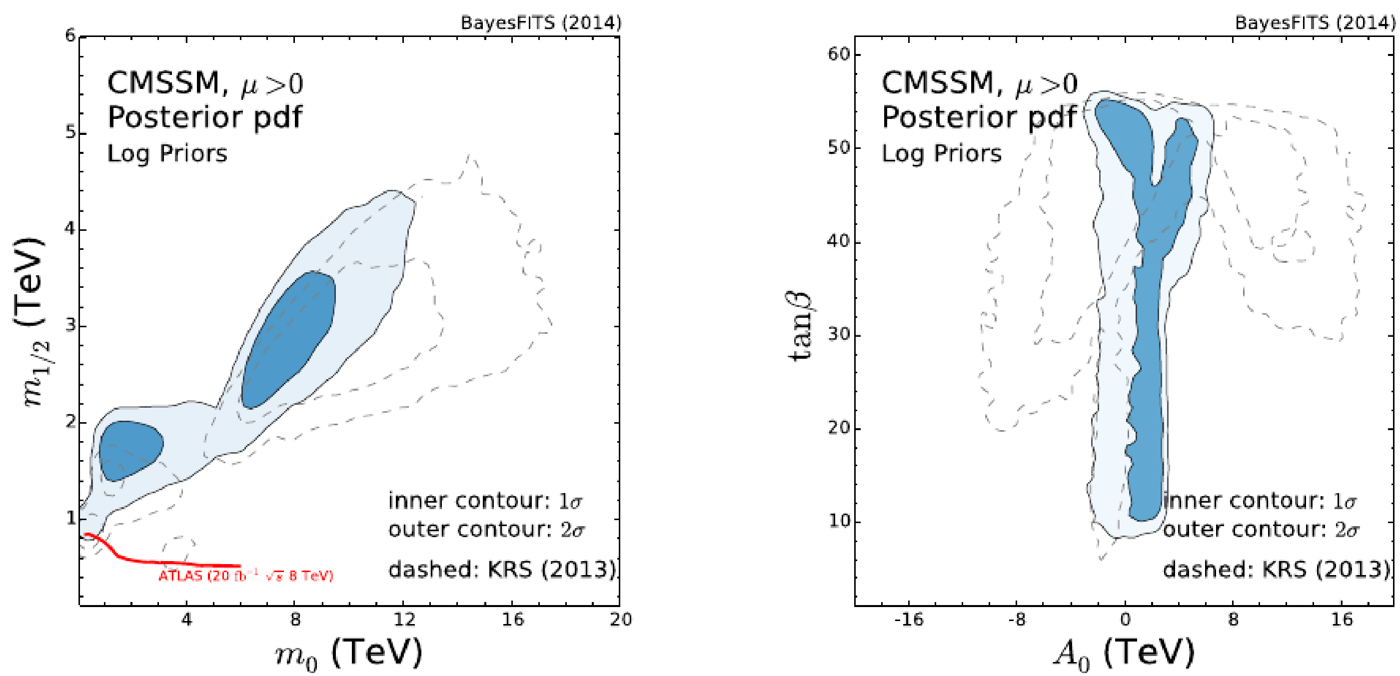
- Some SM parameters are not of immediate interest, but should be varied and accounted for in the analysis too – these are called **nuisance parameters**.
- Examples: $\alpha_S, m_t, m_b, \alpha_{EM}$
- These are always marginalised over

The Global Fit Game



Example: The Constrained MSSM

- Bayesian confidence/credible intervals are not given by $\Delta\chi^2$ contours.
- Instead, 68% of the posterior is contained within the 1σ contour.

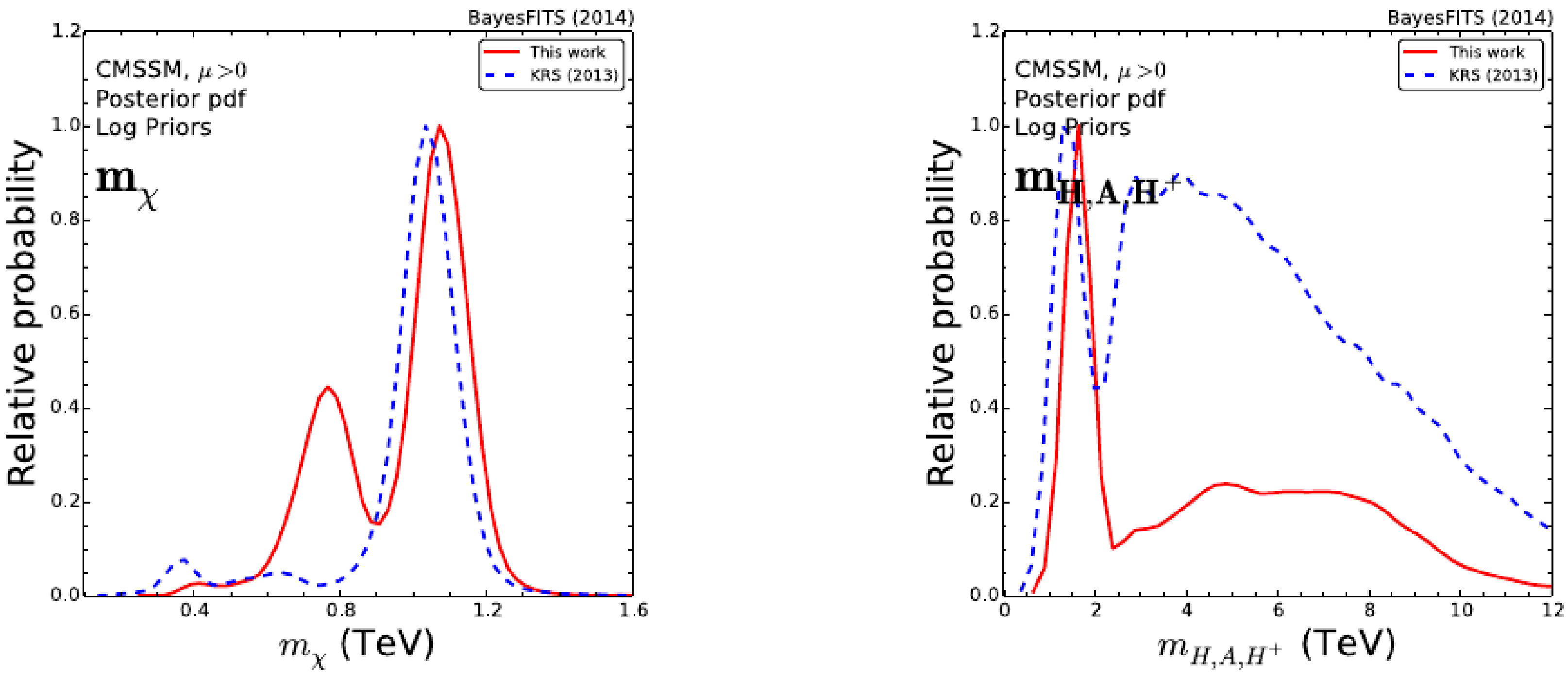


(a) (b)

Roszkowski et al, 2014

Example: The Constrained MSSM

- Bayesian confidence/credible intervals are not given by $\Delta\chi^2$ contours.
- Instead, 68% of the posterior is contained within the 1σ contour.

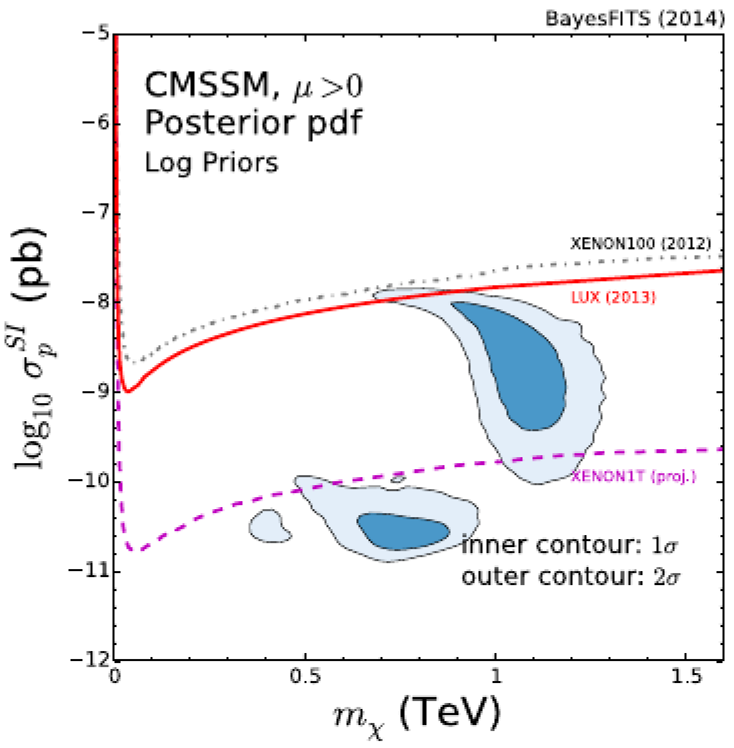


(a) (b)

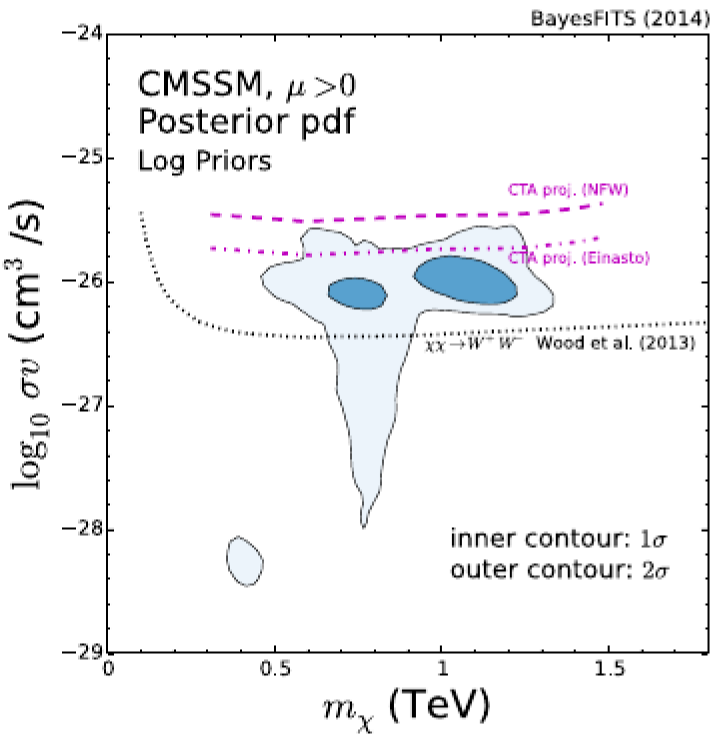
Roszkowski et al, 2014

Example: The Constrained MSSM

- Bayesian confidence/credible intervals are not given by $\Delta\chi^2$ contours.
- Instead, 68% of the posterior is contained within the 1σ contour.



(a)



(b)

Roszkowski et al, 2014

Bayes vs Frequentist

- I have not said much about frequentist approaches.
- Along with differences in philosophy, there are practical differences
- Instead of marginalising, frequentists construct **profile likelihoods**: in each bin take the point with the maximum likelihood (i.e. minimum χ^2)
- This can lead to differences with Bayesian marginalising: integrating out extra parameters means regions of parameter space with low posterior but large volume can give large contributions to the marginalised posterior. These are **volume effects.**

$$p(m_0|\mathrm{data}) = \int dM_{1/2}\ dA_0\ d\tan\beta\ ds\ p(m_0, M_{1/2}, A_0, \tan\beta, s|\mathrm{data}).$$

Bayes vs Frequentist

- Frequentist confidence intervals are also different.
- In practice, fitters usually calculate frequentist confidence intervals by plotting contours of $\Delta\chi^2$ around the best-fit point (point of maximum likelihood).
- In practice this is only right for Gaussian distributions, and one should use pseudo-measurements (“throwing toys”) to correctly estimate confidence intervals.
- I will not say any more about frequentist statistics.

Bridges et al, 2010 Feroz et al, 2011 Bechtle et al, 2015

Bayesian Model Selection

- So, we now know how to construct posterior pdfs for a single model.
- But what if we want to compare or rank models/hypotheses?

We calculate the **Bayesian evidence!**

- This is the factor which normalises the posterior in Bayes theorem (sometimes called $\mathcal{Z}$)

$$p(m|d) = p(d|m)\frac{p(m)}{p(d)},$$

Bayesian Model Selection

- So, we now know how to construct posterior pdfs for a single model.
- But what if we want to compare or rank models/hypotheses?
- In Bayesian statistics, we can only compare between models.

To do this we calculate the **Bayesian evidences!**

- This is the factor which normalises the posterior in Bayes theorem (sometimes called $\mathcal{Z}$)

$$\boxed{p(d|H) \equiv \int p(d|m,H)p(m|H)dm}$$

Bayesian Model Selection

- The Bayesian evidence is the integral of the likelihood times prior over the model parameter space.

$$\boxed{p(d|H) \equiv \int p(d|m,H)\,p(m|H)\,dm}$$

We ignore a factor of p(d) here.

Another application of Bayes theorem tells us:

$$p(H|d) \propto p(H)\,p(d|H)$$

So that we can compare two different hypotheses:

$$\boxed{\frac{p(H_0|d)}{p(H_1|d)} = \frac{p(d|H_0)}{p(d|H_1)}\frac{p(H_0)}{p(H_1)}}$$

Bayesian Model Selection

$$\frac{p(H_0|d)}{p(H_1|d)} = \frac{p(d|H_0)}{p(d|H_1)} \frac{p(H_0)}{p(H_1)}$$

Bayes factor B

Prior ratio for different hypotheses. For N hypotheses. Usually take $p(H_i) = 1/N$

$\|\ln B_{01}\|$	Odds	Probability	Strength of evidence
< 1.0	$\lesssim 3:1$	< 0.750	Inconclusive
1.0	$\sim 3:1$	0.750	Weak evidence
2.5	$\sim 12:1$	0.923	Moderate evidence
5.0	$\sim 150:1$	0.993	Strong evidence

Jeffrey’s scale for interpreting Bayes factor. This is empirical!

Bayesian Model Selection and Occam's Razor

- The Bayesian evidence automatically implements Occam's razor in a neat way.

$$p(d|H) \equiv \int p(d|m, H)p(m|H)dm$$

Say we add another parameter to a model H.
This will decrease the prior normalisation.
Unless the likelihood also increases, this will lead to a decrease in the evidence.

So,extra parameters must lead to a better fit in order to be favoured by the Bayesian evidence.

Trotta, 2008

Bayesian Model Selection and Occam's Razor

- The Bayesian evidence also penalises fine-tuning.
- Consider two model hypotheses, with parameter spaces of the same volume.
- In Model 1, the likelihood is sharply peaked in small region of parameter space.
- In Model 2 the likelihood is high over most of parameter space (plateau-style)
- The Bayesian evidence will then favour model 2.

$$p(d|H) \equiv \int p(d|m,H)p(m|H)dm$$

Public Tools

There are many of public tools available for fitting.

- MultiNest: General purpose sampler for Frequentist+Bayesian applications. Calculates evidence, wrappers for C/C++/Python/R/Matlab: MultiNest PyMultiNest
- MontePython (more for cosmo): MontePython
- PyMC: Python implementation of MCMC: PyMC
- PolyChord (next-gen Nested Sampling): PolyChord

- SuperPlot (python, A. Fowlie): SuperPlot
- Pippi (python): Pippi
- CosmoMC (GetDist): GetDist

To appear: GAMBIT (M White talk)...

Summary

- Global fits are a powerful and useful tool in the phenomenologists' arsenal
- Allow to extract maximum information about the effects of data on a model, in a statistically coherent way.
- May prove critical in discovery of New Physics.
- Do require some knowledge of statistics…
- But an excellent array of public tools are available!
- Get sampling!

References

- Frequentism vs Bayesianism: Pythonic Perambulations
- Bayes and Frequentism: A Particle Physicist's Perspective, Louis Lyons, 1301.1273
- Probability and Statistics for Particle Physicists, J. Ocariz, 1405.3402
- Practical Statistics for the LHC, Kyle Cranmer, 1503.07622
- Bayesian Methods in Cosmology, Roberto Trotta, 0803.4089 and 1701.01467
- Why isn't every physicist a Bayesian?, R Cousins, PDF Link