Predictiveness as Discovery

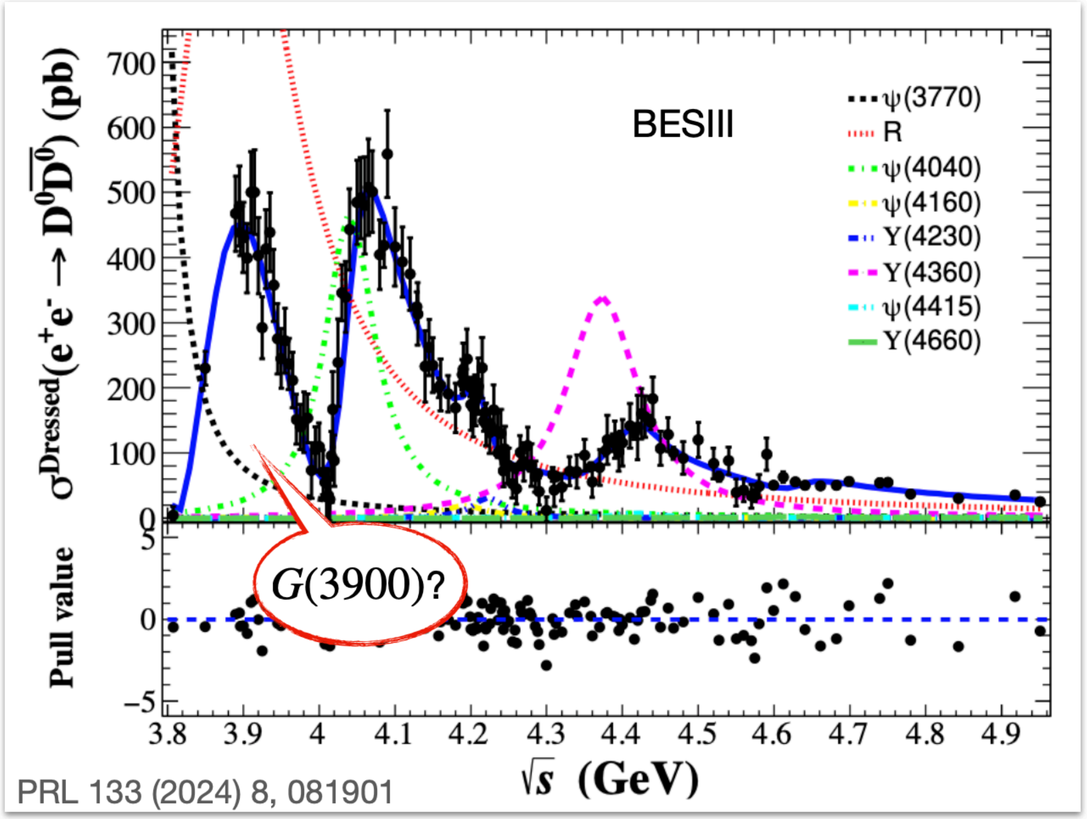
ExoHad
Exotic Hadrons Topical Collaboration

the traditional approach to "discovery"

G(3900)

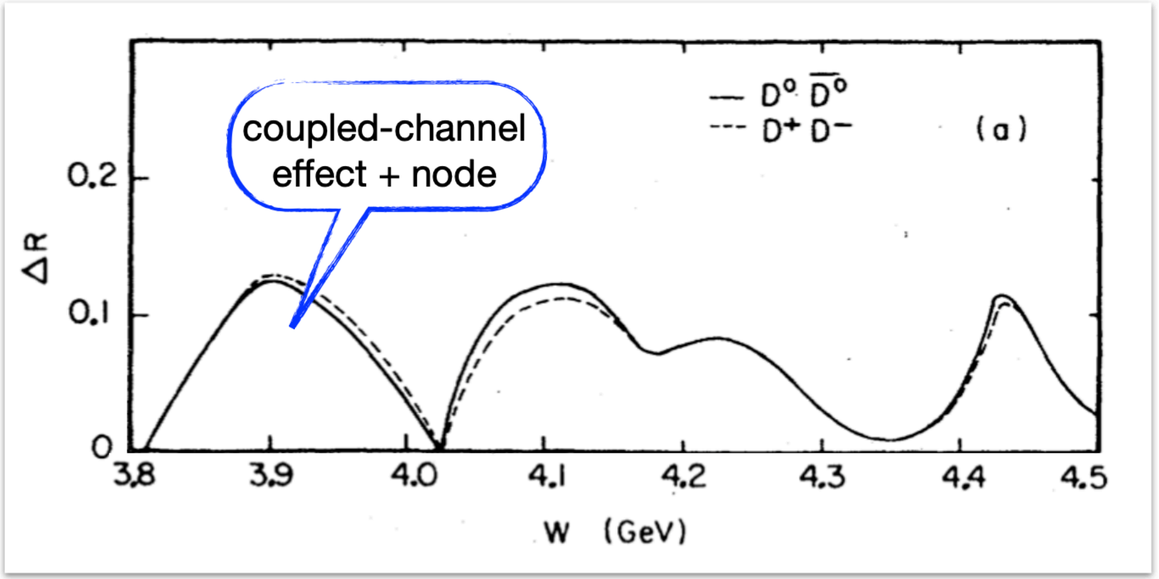
N. Hüsken, R.F. Lebed, R.E. Mitchell, E.S. Swanson, Y-Q Wang, 2404.03896

$e^+e^- \to D\bar{D}$



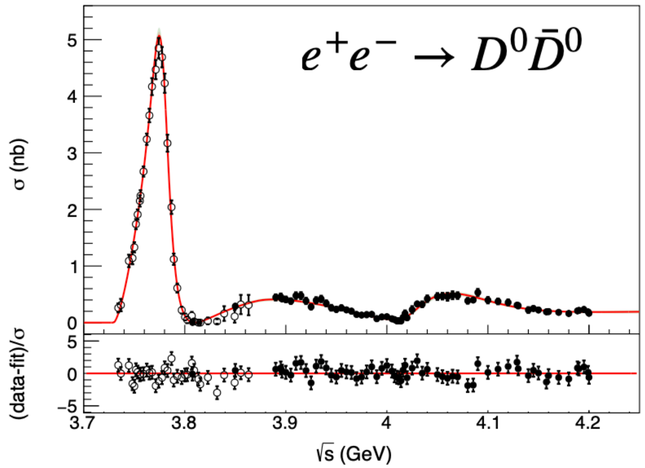
PRL 133 (2024) 8, 081901

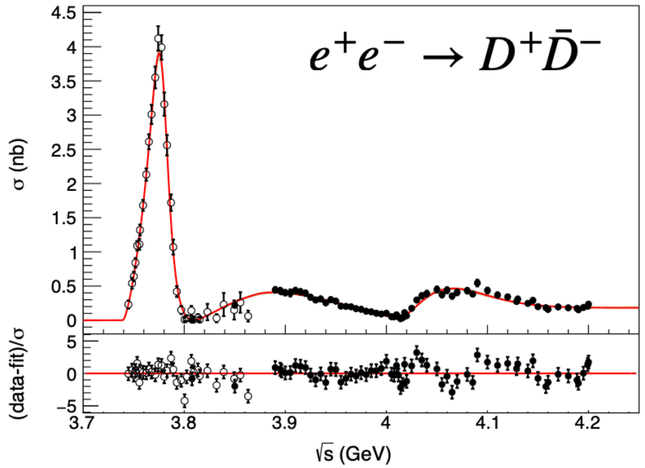
Eichten et al., Phys. Rev. D 21 (1980) 203

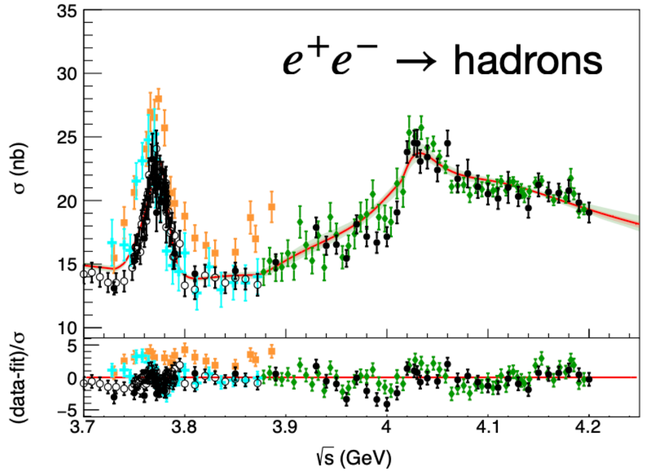


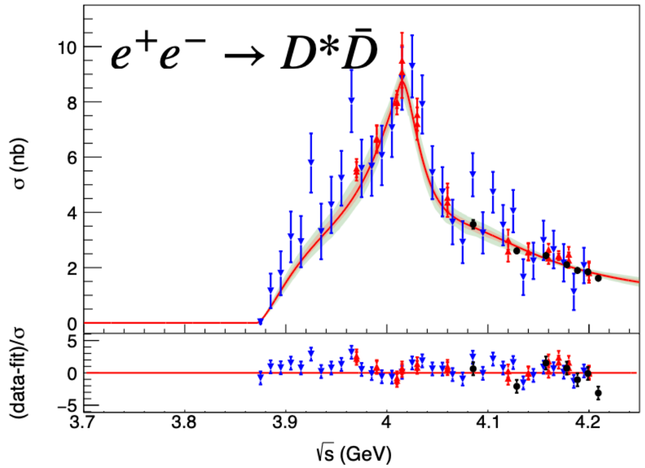
In our calculation there is some weak structure in the 3.9–4.0 GeV region. It does not arise from a $c\bar{c}$ resonance, but from the opening of the $D\bar{D}^* + D^*\bar{D}$ channel and a decrease in the $D\bar{D}$ channel due to a nearby zero in the $3S$ decay amplitude.

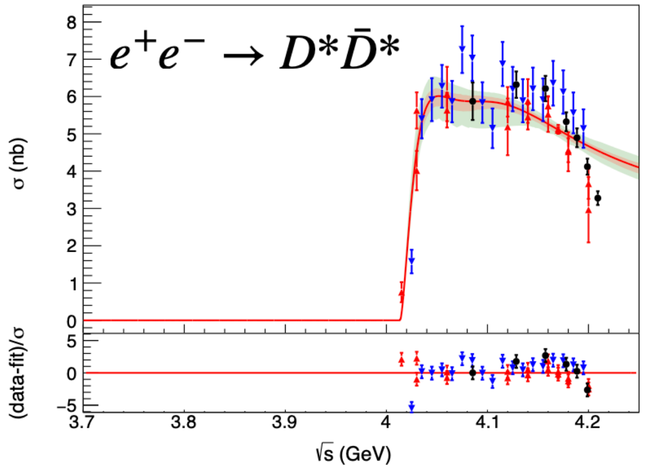
Refit with additional information and coupled channels







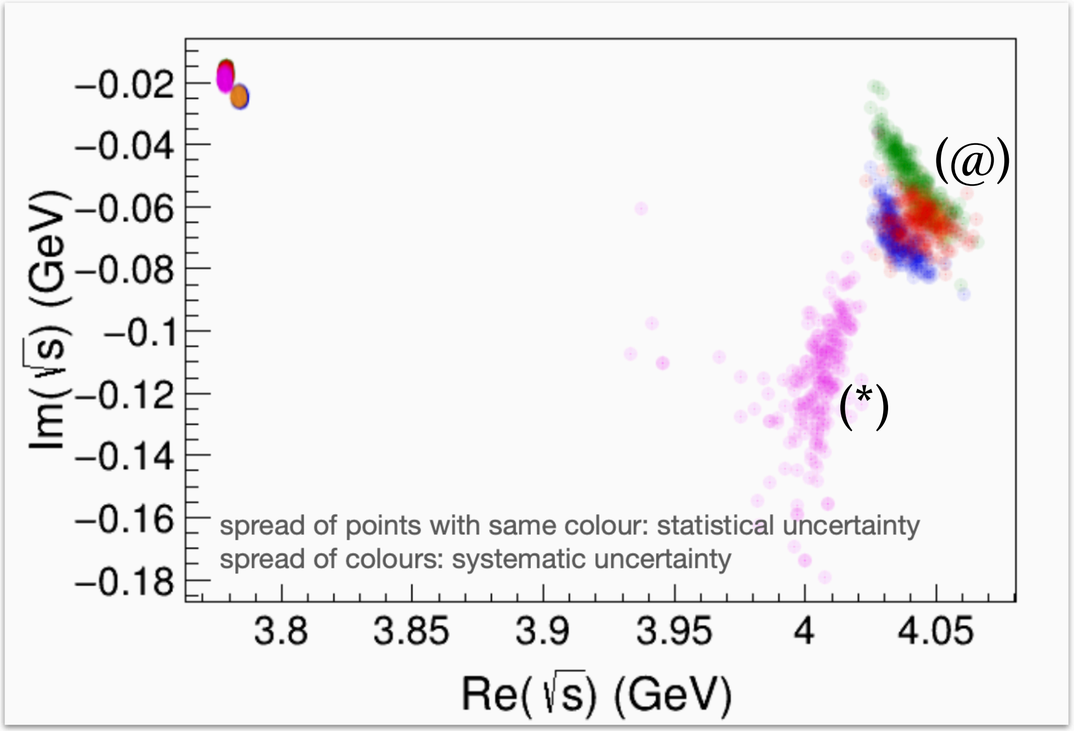




fit result
68% CL
90% CL

Data from:
BESIII (unoff.) Andy Julin, University of Minnesota
BESIII: PRL 133 (2024) 8, 081901
BESIII: JHEP 05 (2022) 155
Belle: Phys.Rev.D 97 (2018) 1, 012002
CLEO: Phys.Rev.D 80 (2009) 072001
BES: PRL 88, 101802 (2002)
BESII: PRL 97, 262001 (2006)
SPEAR: PRL 39, 526 (1977);
A. Osterheld et al. 86; Schindler 79

(@) 4 models w/ 24-30 parameters



(*) model with node

the devil's in the details:

© Chat

- given our choices of K_{ij}, $n(s)$, Σ, we find the data can be described without a $G(3900)$ pole
- details we varied:
  - isospin-constraints between D^+D^- and $D^0\bar{D}^0$
  - threshold opening of channel that absorbs *missing* intensity
  - node in $\hat{n}(s) = n(s) \cdot (1 - k^2/k_0^2)$
- details we did not vary:
  - type of barrier factor $b_L(s)$
  - choice of Σ
- very valuable that other groups use different assumptions entirely different approaches

PRD 109 (2024) 11, 114015
PRL 133 (2024) 24, 241903
PRD 112 (2025) 5, 054027
PRD 112 (2025), 016015
arXiv:2509.17679

some find a $G(3900)$, some do not…

+ more in other energy regions

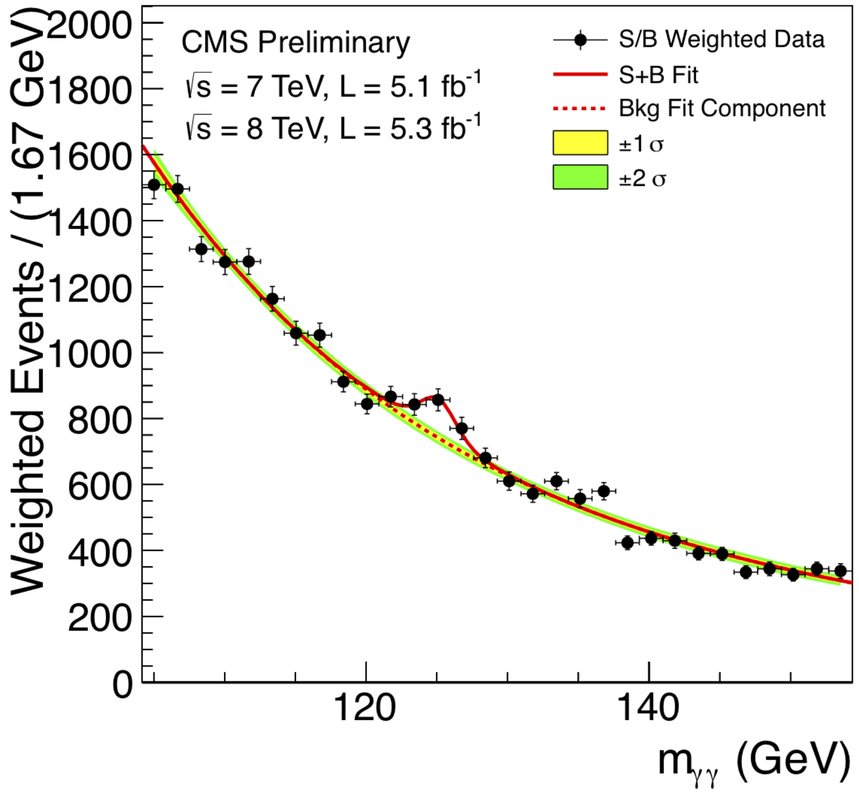
the traditional approach to discovery

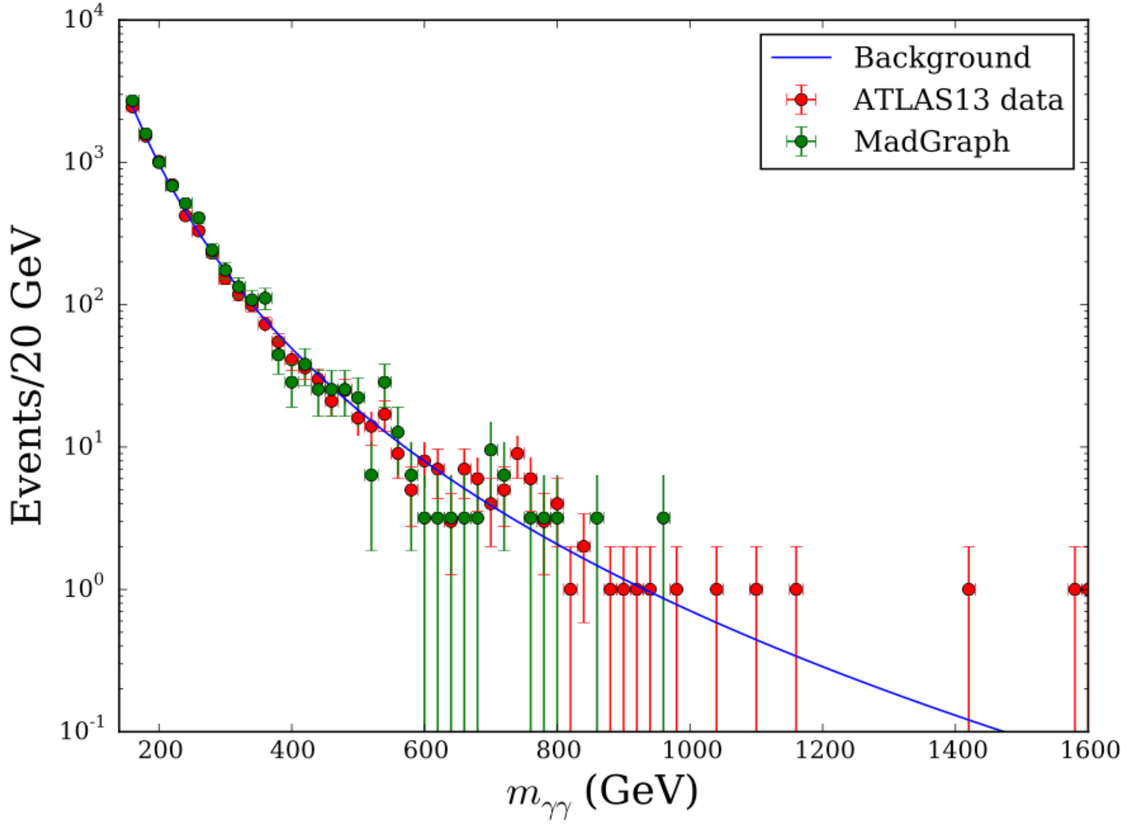
i. fit two model amplitudes to data, $M_0(\ldots)$ and $M_1(\ldots; m_R, \Gamma_R)$

ii. determine the P-value ~ the probability of obtaining the observed effect (or greater) given that the null hypothesis is true.

$$2\log\frac{L_1}{L_0} \to \chi^2_{d_1-d_0}, \quad L_0 \subset L_1.$$ (Wilk's theorem)

iii. declare a discovery if $P < 3\cdot 10^{-7} \to 5\sigma$ $[P < 0.0027 \to 3\sigma]$. The new physics is described by the fit parameters, m_R, Γ_R.





problems with the traditional approach

i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it

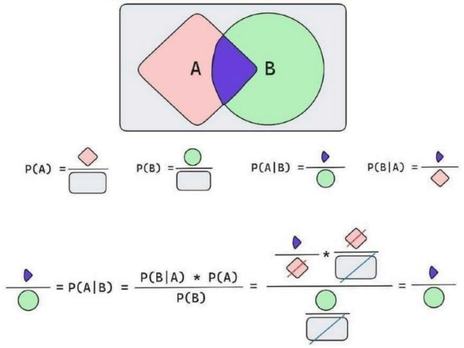
   a. LASSO

   b. AIC

   c. BIC

   d. Bayesian Model Averaging :

$$p(M \mid \mathscr{D}) = \int d\theta \, p(\mathscr{D} \mid \theta; M) \, p(\theta \mid M) \, p(M)$$



Priors are problematic! What is a uniform prior? Lack of reparametrization invariance. What is the totality of causative agents?

additional concerns

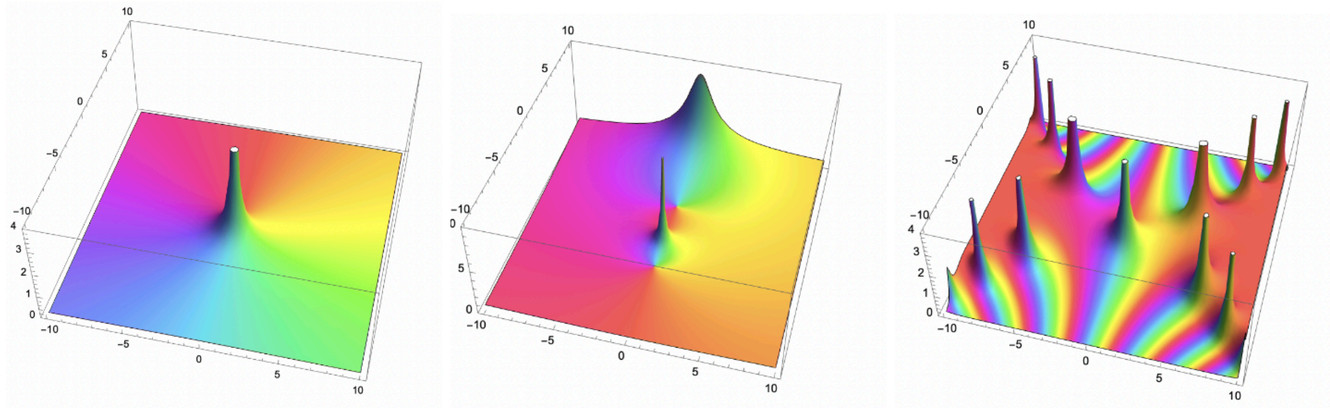
i. data sets are not providential

ii. no model is providential

iii. model parameters are (approximately?) meaningless

iv. we are not interested in the model, but rather the analytic structure of the model

(or, at least, should be!)

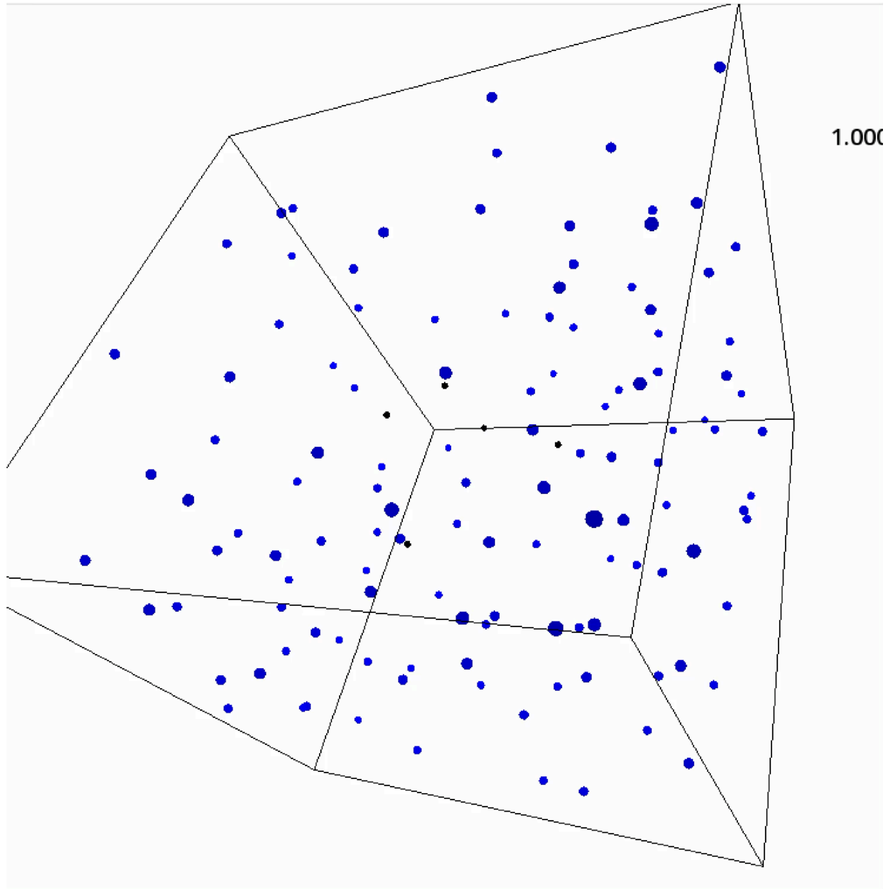


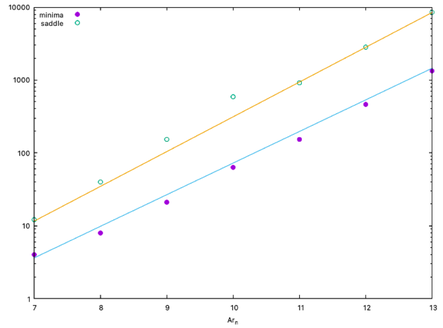
bigger problems

i. what does it mean to model?

ii. what does it mean to minimize?

A 38 element Lennard-Jones system has ~ 10^{14} local minima (!!)





C.J. Tsai and K.D. Jordan, J. Phys Chem, **97** 227 (1993).

a way forward:

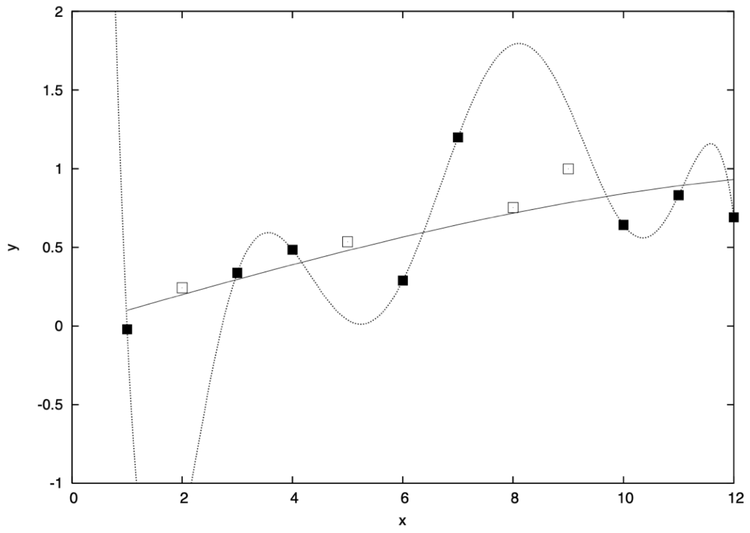
- One way in which one can claim the discovery of a state is if that knowledge permits better, or *predictive*, statements to be made about future experiments.

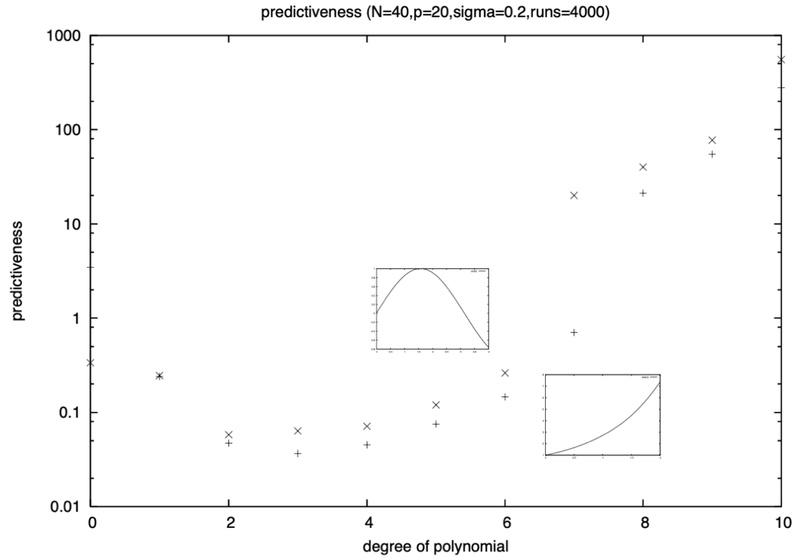
"How well does my model fit the data?" ➔ "How well can I predict the outcomes of future experiments?"

- We can use cross-validation to avoid overfitting.

- Combine these ideas by splitting the data set into training (ante) and validation (post) sets.

A simple example





Bayesian Predictivity

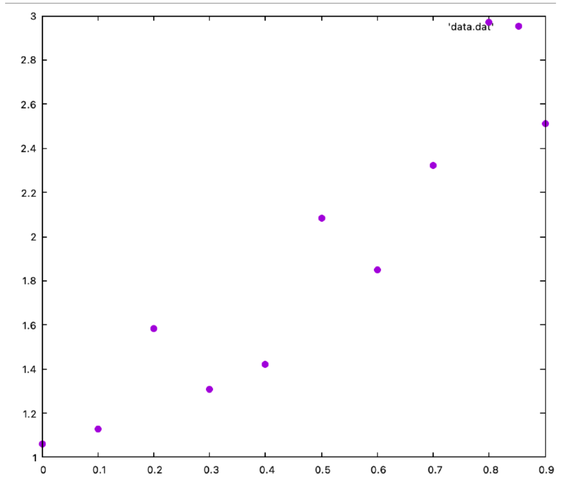
- abandon the idea that we "know" the model -- work in "super-model space" (which, of course, is a model; but now we seek a degree of agnosticism).
- stochastically explore model space. [One could model average by averaging over the trajectory. This is not our primary goal here.]
- explore the continuous portion of model space with Markov chain Monte Carlo. Metropolis-Hastings update according to

$$p(\theta \to \theta') = \min\left(1, \frac{p(\theta' \mid \mathscr{D})}{p(\theta \mid \mathscr{D})} \frac{f_{\theta,\theta'}}{f_{\theta',\theta}}\right) \qquad p(\theta \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \theta)\, p(\theta)}{p(\mathscr{D})}$$

1000 MCMC steps starting at (0.5,0.5); uniform step of size $\in [-0.1, 0.1]$

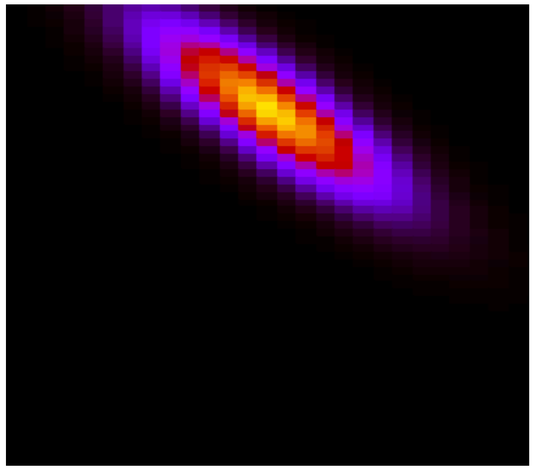
Every point in this space is a model, some are just better than others.

$\mathscr{D}(1 + 2x + \hat{\eta}(0.2))$



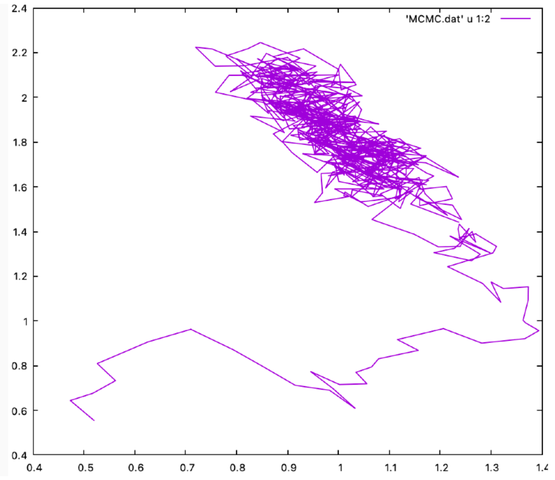
scan of the posterior

b



a

b



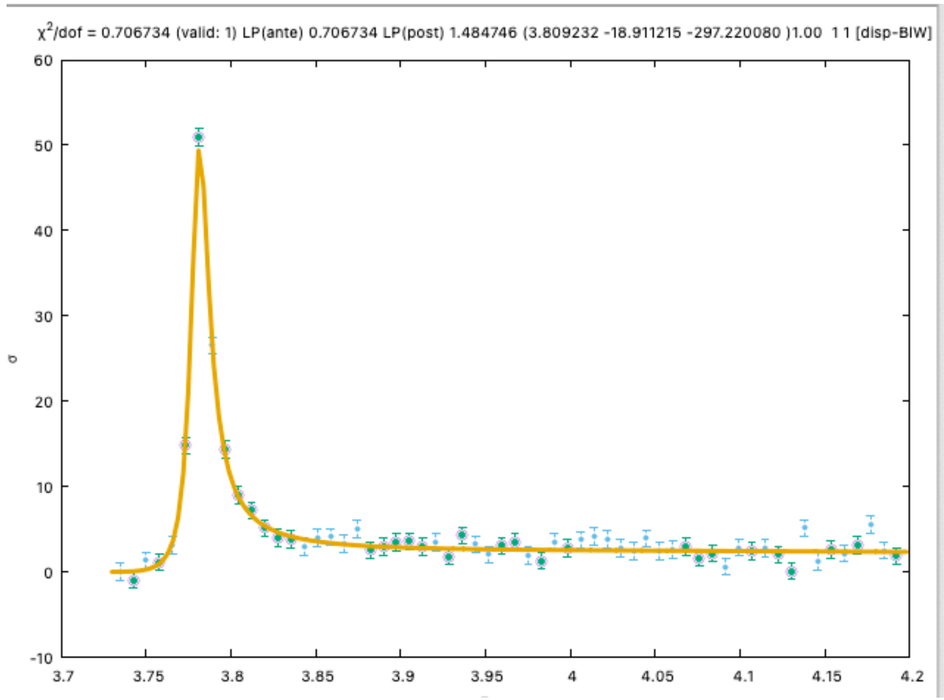
a

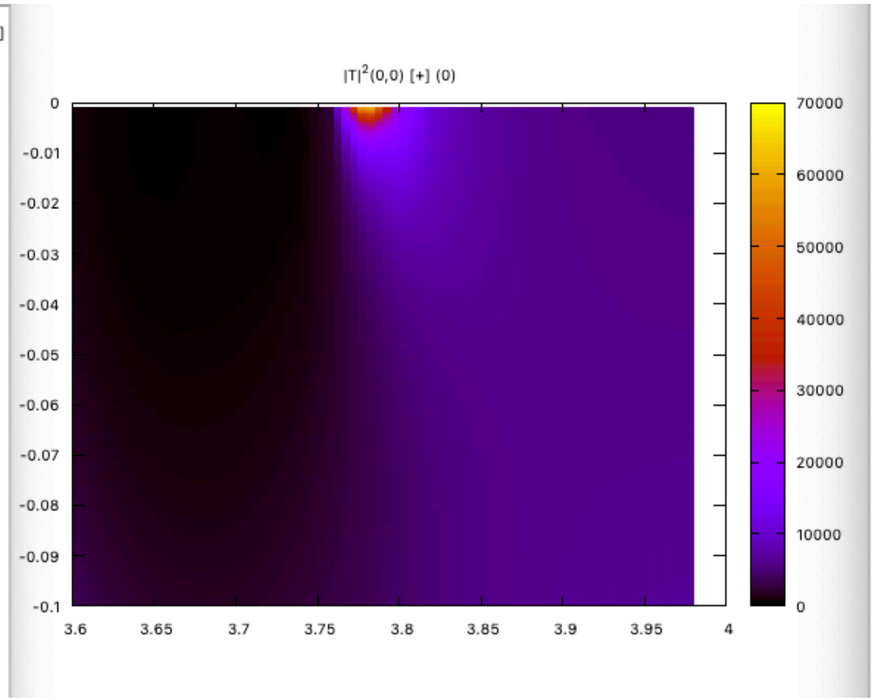
Single Channel K-matrix

Generate synthetic cross section via $\mathscr{M}^{-1} = K^{-1} - i\rho \rightarrow K^{-1} + C$ with $C = -\int_{s_0} \frac{ds'}{\pi} \frac{\rho(s')g^2(s')}{s' - s - i\epsilon}$.

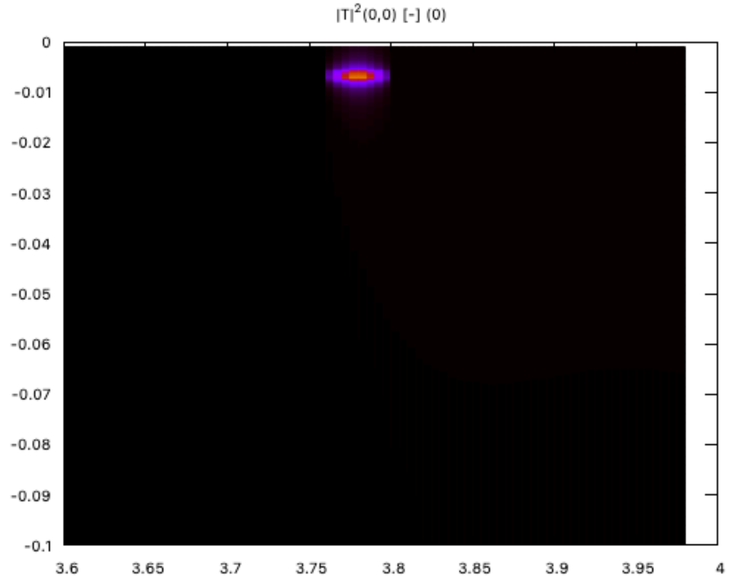
K-matrix parameterization comprises model space

$$M(\vec{\theta}; R, N, Q, C) \rightarrow K_{\mu\nu} = \sum_{r=1}^{R} \frac{g^{(Q)}_{R:\mu} g^{(Q)}_{R:\nu}}{m_R^2 - s} + \sum_{i=0}^{N} c^{(i)}_{\mu\nu} s^{i/2}$$





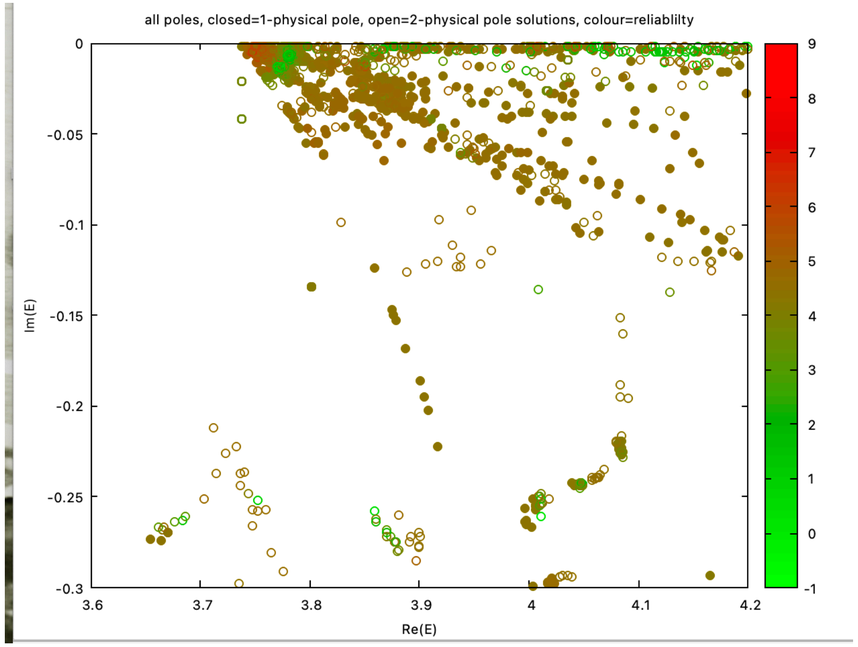
pole location on sheet II

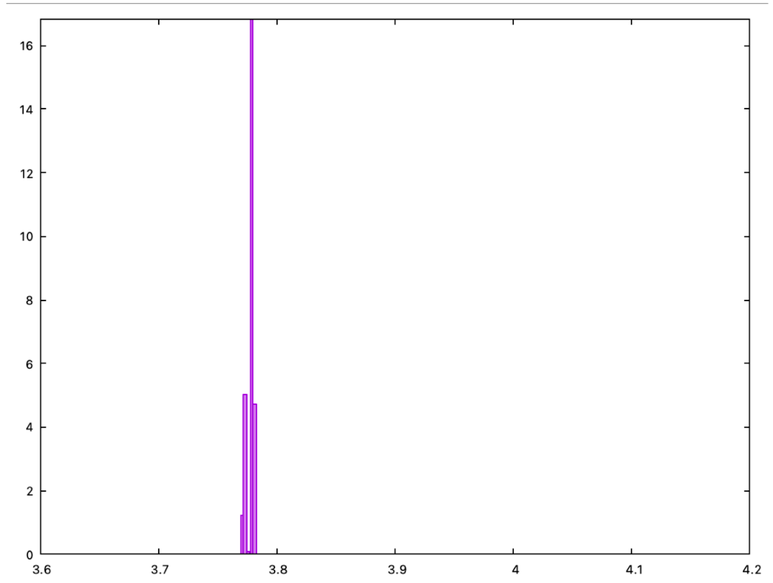


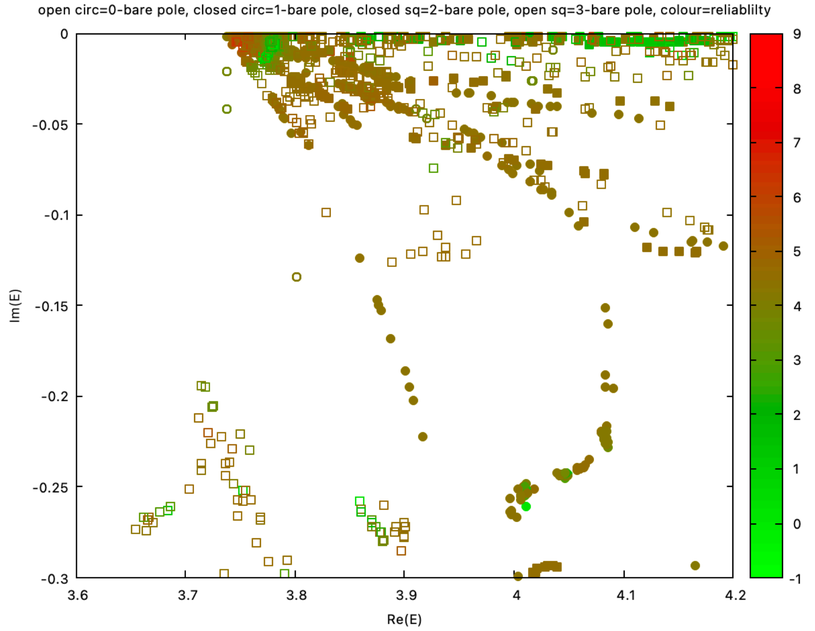
Pole Positions

We seek to say something about one-pole vs 2-pole fits. Here are the optimal reliabilties for each option

```
./MAK4
 number of available threads = 10
 enter verbose [0/1], make plots [0/1], exclude gen [0/1], nthreads, seed, max model size, number of data points, sig for data generation
0 0 1 8 891723 4 60 1.
 enter priorWidth, number of models to throw
1. 12
 enter NMC, nMeas (total MCMC=nMeas*NMC)
4 20
 enter the MCMC step size for  masses, couplings, bg's
.05 .1 .1


global average single pole  3.77841 +/- 0.00292129 (8.44008e-05)  +i -0.00823452 +/- 0.0031752 (9.17367e-05)

one pole optimum reliability: 0.483037 @ (3.78,-0.007) residue: (-13.9922,5.97456)  model:  beta = 2 nPoles = 3 nBack = 0 Copt = 1 FFopt = 1

two pole optimum reliability: 0.373839 @ (3.665,-0.268) residue: (9.54467,-4.69041) model:  beta = 1 nPoles = 3 nBack = 4 Copt = 2 FFopt = 2
two pole optimum reliability: 0.373839 @ (3.773,-0.014) residue: (-5.97938,8.18145) model:  beta = 1 nPoles = 3 nBack = 4 Copt = 2 FFopt = 2
```







big run Mmax=4 *predictiveness/ModAvg/MAK4.cpp*

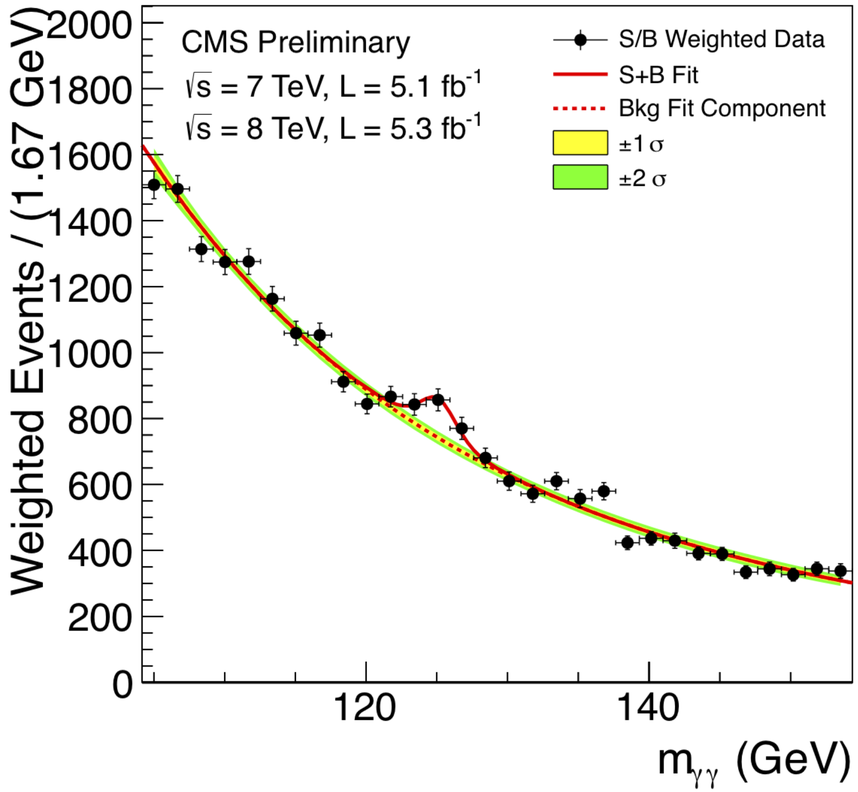
the traditional approach to discovery

i. fit two model amplitudes to data, $M_0(\ldots)$ and $M_1(\ldots; m_R, \Gamma_R)$

ii. determine the P-value ~ the probability of obtaining the observed effect (or greater) given that the null hypothesis is true.

$$2\log\frac{L_1}{L_0} \to \chi^2_{d_1-d_0}, \quad L_0 \subset L_1.$$ (Wilk's theorem)

iii. declare a discovery if $P < 3\cdot 10^{-7} \to 5\sigma$ $[P < 0.0027 \to 3\sigma]$. The new physics is described by the fit parameters, m_R, Γ_R.



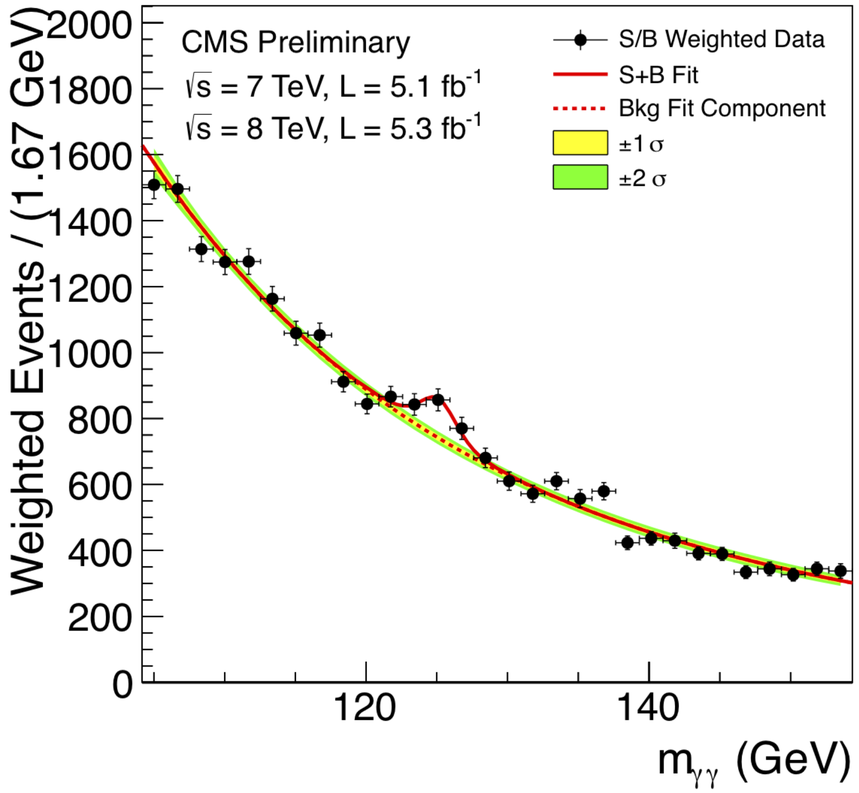
the traditional approach to discovery

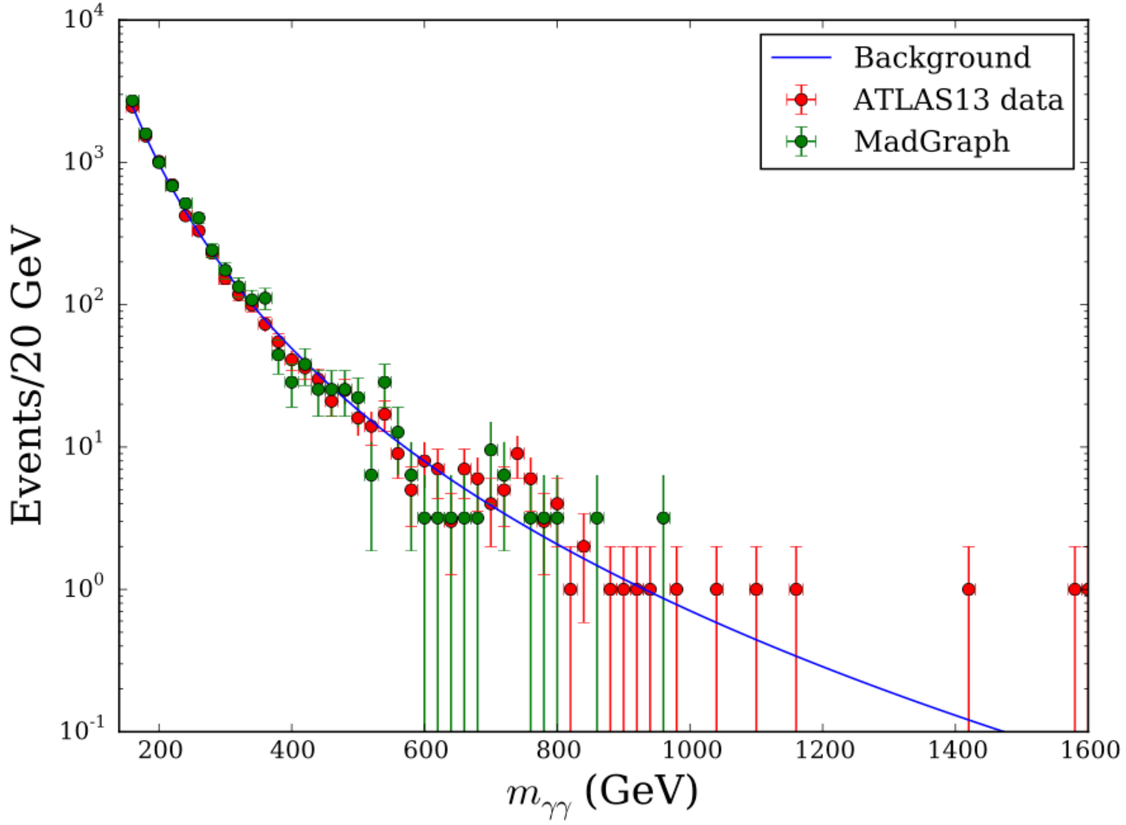
i. fit two model amplitudes to data, $M_0(\ldots)$ and $M_1(\ldots; m_R, \Gamma_R)$

ii. determine the P-value ~ the probability of obtaining the observed effect (or greater) given that the null hypothesis is true.

$$2\log\frac{L_1}{L_0} \to \chi^2_{d_1-d_0}, \quad L_0 \subset L_1.$$ (Wilk's theorem)

iii. declare a discovery if $P < 3\cdot 10^{-7} \to 5\sigma$ $[P < 0.0027 \to 3\sigma]$. The new physics is described by the fit parameters, m_R, Γ_R.





problems with the traditional approach

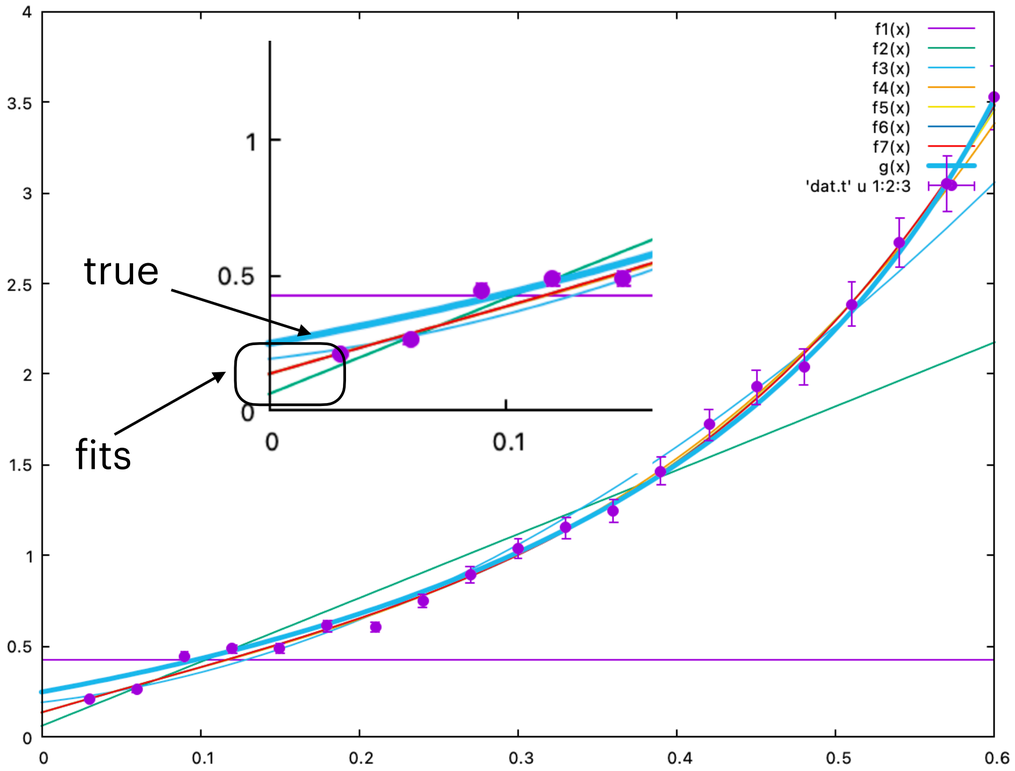
i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it



problems with the traditional approach

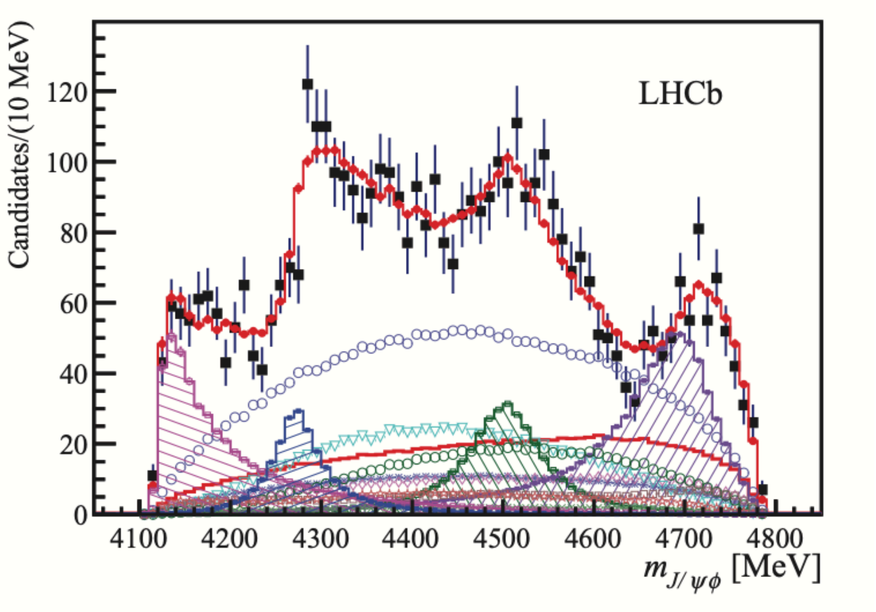
i. fluctuations in the data set may be important

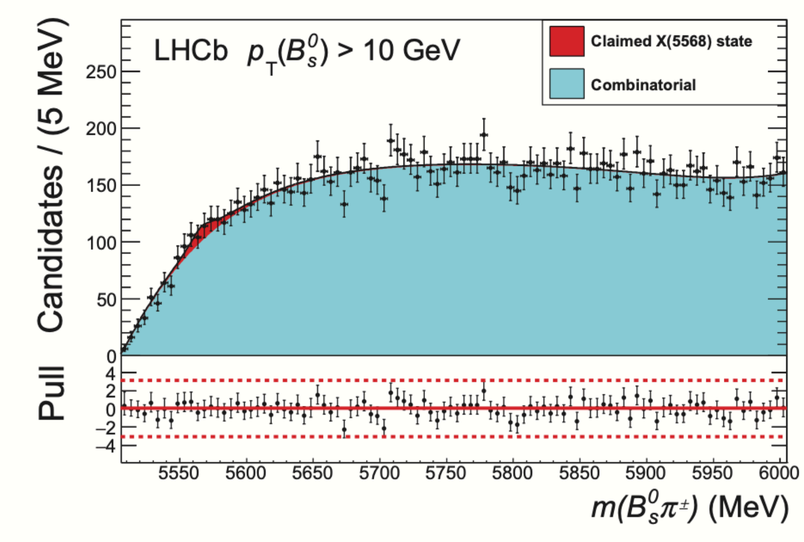
ii. models M0 and M1 are wrong!

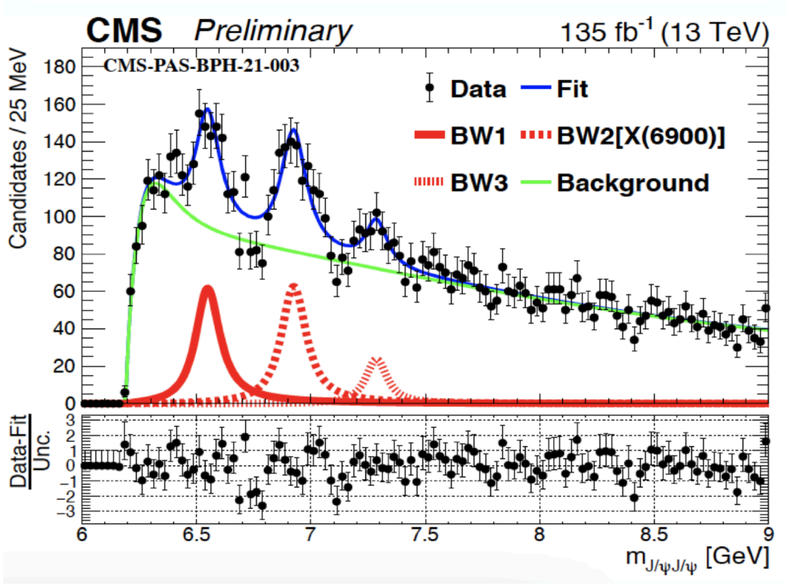
iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it







problems with the traditional approach

i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

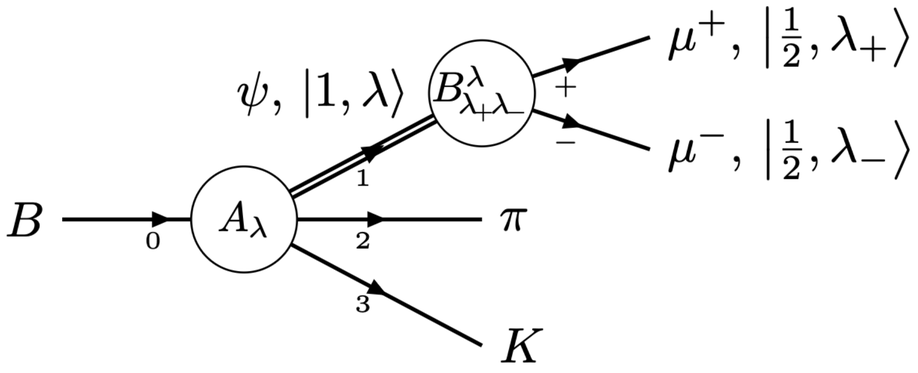
v. we wish to extract model structure, not assume

$$\mathrm{BW}(m|M_0,\Gamma_0) = \frac{1}{M_0^2 - m^2 - iM_0\Gamma(m)}\,,$$

$$\Gamma(m) = \Gamma_0 \left(\frac{q}{q_0}\right)^{2L_{\Lambda^*}+1} \frac{M_0}{m} B'_{L_{\Lambda^*}}(q,q_0,d)^2\,.$$

$$R_{\Lambda_n^*}(m_{Kp}) = B'_{L^{\Lambda_n^*}_{\Lambda_b^0}}(p,p_0,d) \left(\frac{p}{M_{\Lambda_b^0}}\right)^{L^{\Lambda_n^*}_{\Lambda_b^0}} \mathrm{BW}(m_{Kp}|M_0^{\Lambda_n^*},\Gamma_0^{\Lambda_n^*})\, B'_{L_{\Lambda_n^*}}(q,q_0,d) \left(\frac{q}{M_0^{\Lambda_n^*}}\right)^{L_{\Lambda_n^*}}. \quad (21$$

$$\mathcal{M}^{\Lambda^*}_{\lambda_{\Lambda_b^0},\lambda_p,\Delta\lambda_\mu} = \sum_n R_{\Lambda_n^*}(m_{Kp})\, \mathcal{H}^{\Lambda_n^*\to Kp}_{\lambda_p} \sum_{\lambda_\psi} e^{i\lambda_\psi\phi_\mu}\; d^1_{\lambda_\psi,\Delta\lambda_\mu}(\theta_\psi)$$
$$\times \sum_{\lambda_{\Lambda^*}} \mathcal{H}^{\Lambda_b^0\to\Lambda_n^*\psi}_{\lambda_{\Lambda^*},\lambda_\psi}\; e^{i\lambda_{\Lambda^*}\phi_K}\; d^{\frac{1}{2}}_{\lambda_{\Lambda_b^0},\lambda_{\Lambda^*}-\lambda_\psi}(\theta_{\Lambda_b^0})\; d^{J_{\Lambda_n^*}}_{\lambda_{\Lambda^*},\lambda_p}(\theta_{\Lambda^*}).$$



unitary?

correct analytic structure?

crossing symmetric?

non-perturbative?

problems with the traditional approach

i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

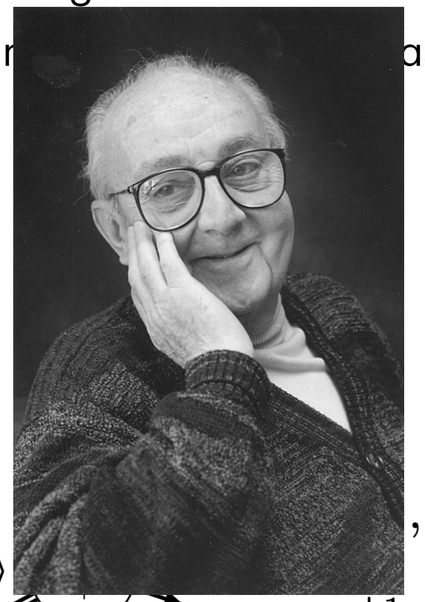
v. we wish to extract r… assume

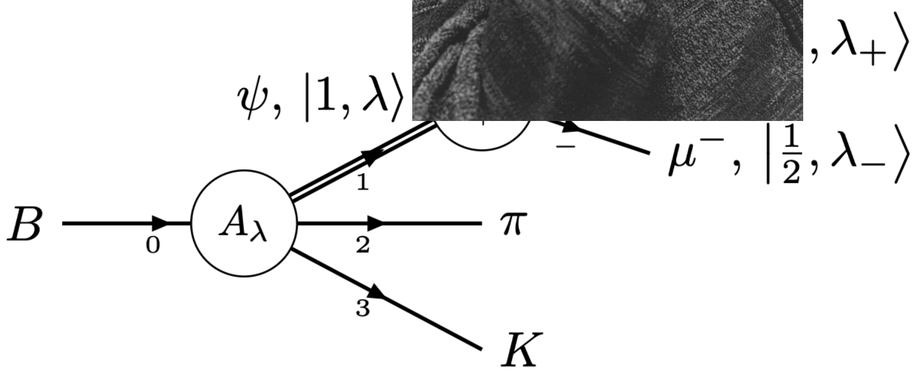
$$\mathrm{BW}(m|M_0,\Gamma_0) = \frac{1}{M_0^2 - m^2 - iM_0\Gamma(m)},$$

$$\Gamma(m) = \Gamma_0 \left(\frac{q}{q_0}\right)^{2L_{\Lambda^*}+1} \frac{M_0}{m} B'_{L_{\Lambda^*}}(q,q_0,d)^2 .$$

$$R_{\Lambda_n^*}(m_{Kp}) = B'_{L^{\Lambda_n^*}_{\Lambda_b^0}}(p,p_0,d) \left(\frac{p}{M_{\Lambda_b^0}}\right)^{L^{\Lambda_n^*}_{\Lambda_b^0}} \mathrm{BW}(m_{Kp}|M_0^{\Lambda_n^*},\Gamma_0^{\Lambda_n^*})\, B'_{L_{\Lambda_n^*}}(q,q_0,d) \left(\frac{q}{M_0^{\Lambda_n^*}}\right)^{L_{\Lambda_n^*}} . \quad (21$$

$$\mathcal{M}^{\Lambda^*}_{\lambda_{\Lambda_b^0},\lambda_p,\Delta\lambda_\mu} = \sum_n R_{\Lambda_n^*}(m_{Kp})\, \mathcal{H}^{\Lambda_n^*\to Kp}_{\lambda_p} \sum_{\lambda_\psi} e^{i\lambda_\psi\phi_\mu}\; d^1_{\lambda_\psi,\Delta\lambda_\mu}(\theta_\psi)$$
$$\times \sum_{\lambda_{\Lambda^*}} \mathcal{H}^{\Lambda_b^0\to\Lambda_n^*\psi}_{\lambda_{\Lambda^*},\lambda_\psi}\; e^{i\lambda_{\Lambda^*}\phi_K}\; d^{\frac{1}{2}}_{\lambda_{\Lambda_b^0},\lambda_{\Lambda^*}-\lambda_\psi}(\theta_{\Lambda_b^0})\; d^{J_{\Lambda_n^*}}_{\lambda_{\Lambda^*},\lambda_p}(\theta_{\Lambda^*}).$$





unitary?

correct analytic structure?

crossing symmetric?

non-perturbative?

problems with the traditional approach

i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

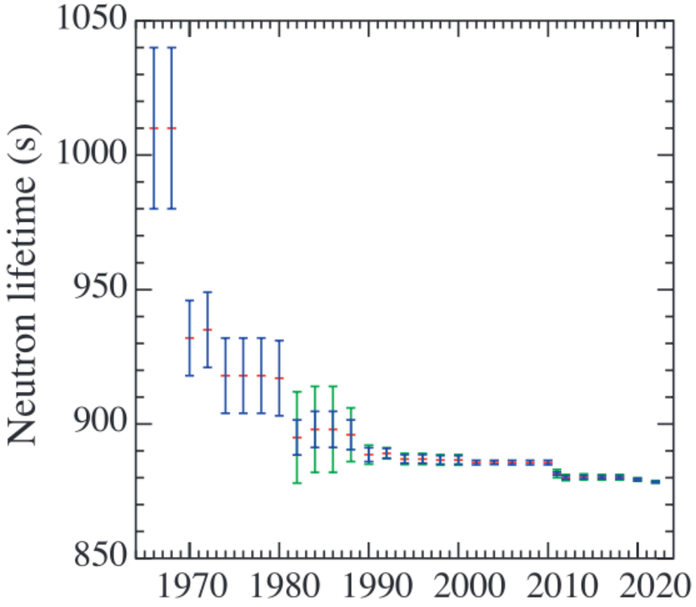
iii. systematic errors are often underestimated

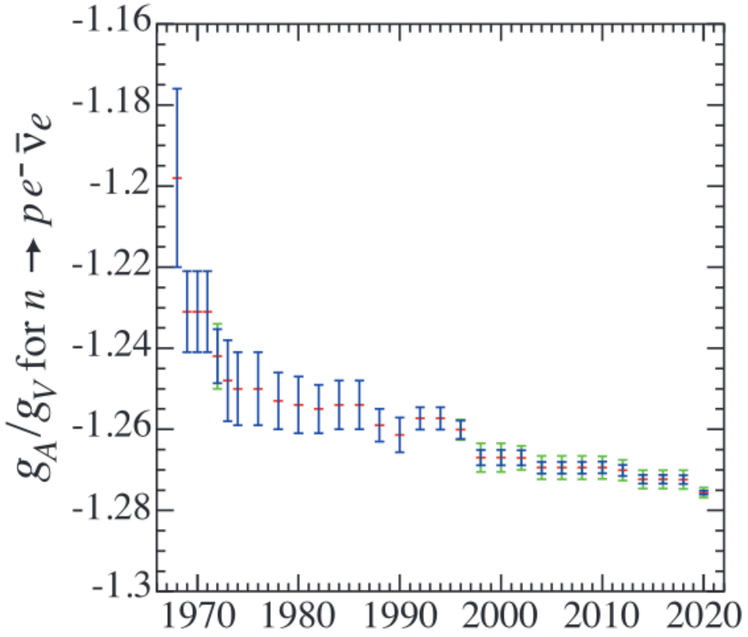
iv. problematic overfitting

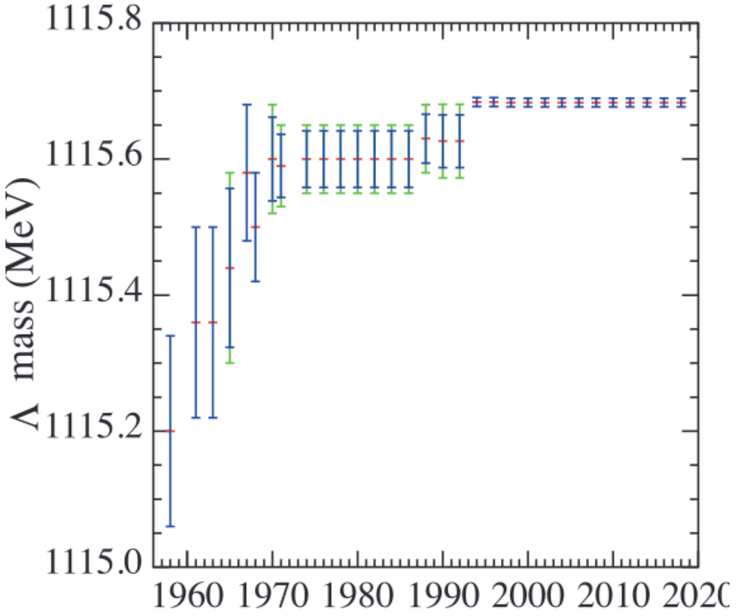
v. we wish to extract model structure, not assume it

i.e. over-confidence in one's model.

i.e. one acts as if the model generated the data.







problems with the traditional approach

i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

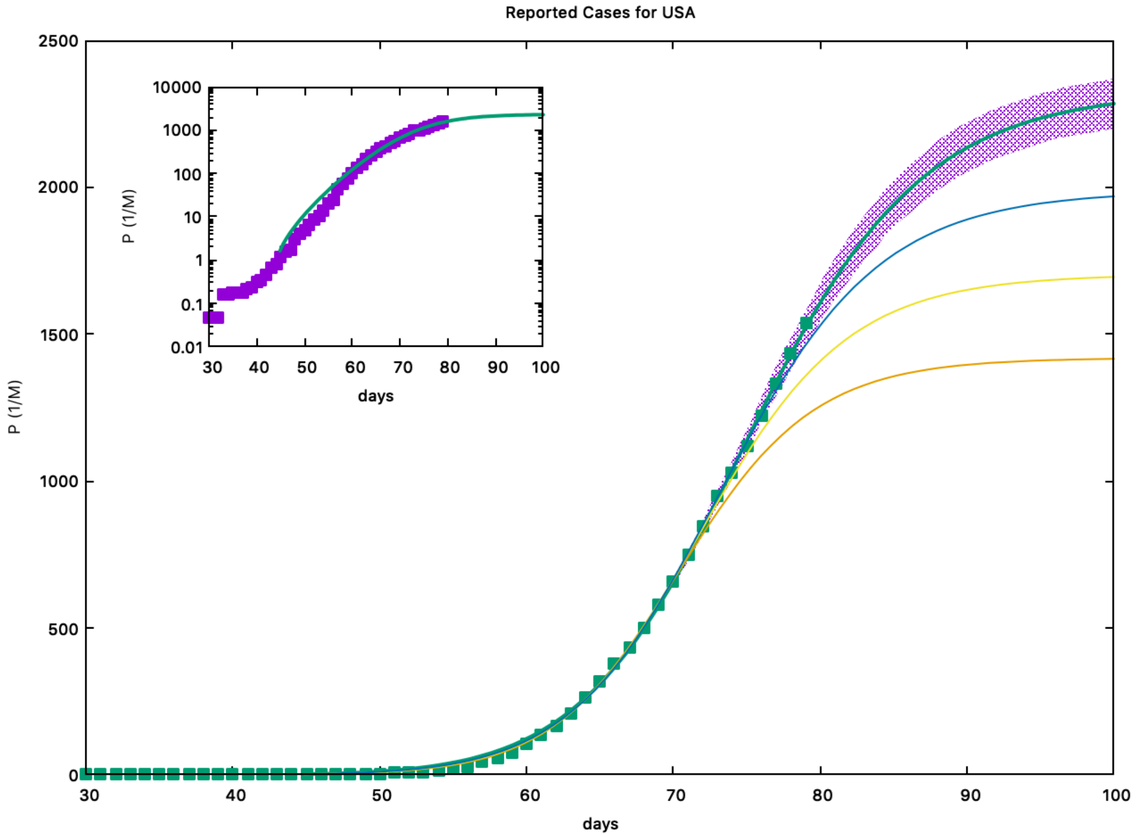
iv. problematic overfitting

v. we wish to extract model structure, not assume it

i.e. over-confidence in one's model.

i.e. one acts as if the model generated the data.

fits of a logistic function to coronovirus cases in the US, over time



USA Apr 10, 2020

problems with the traditional approach

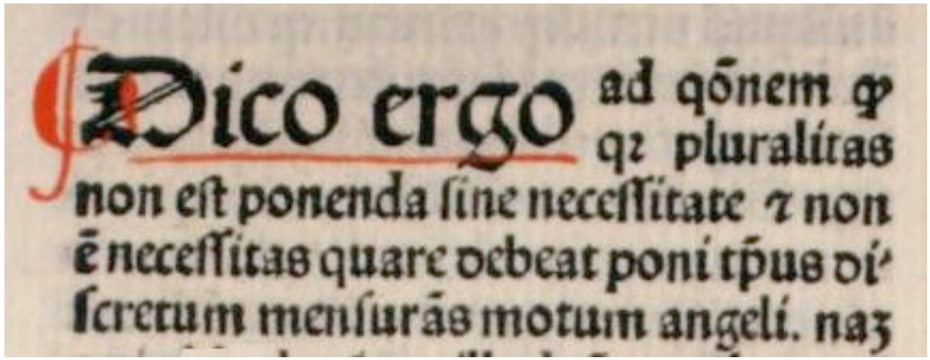
i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it



problems with the traditional approach

i. fluctuations in the data set may be important

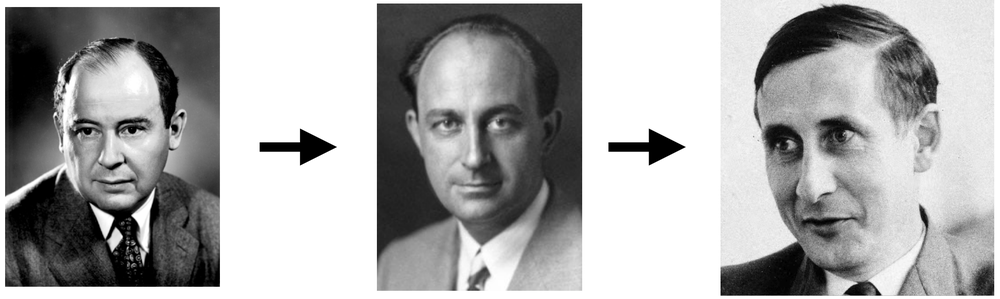
ii. models M0 and M1 are wrong!

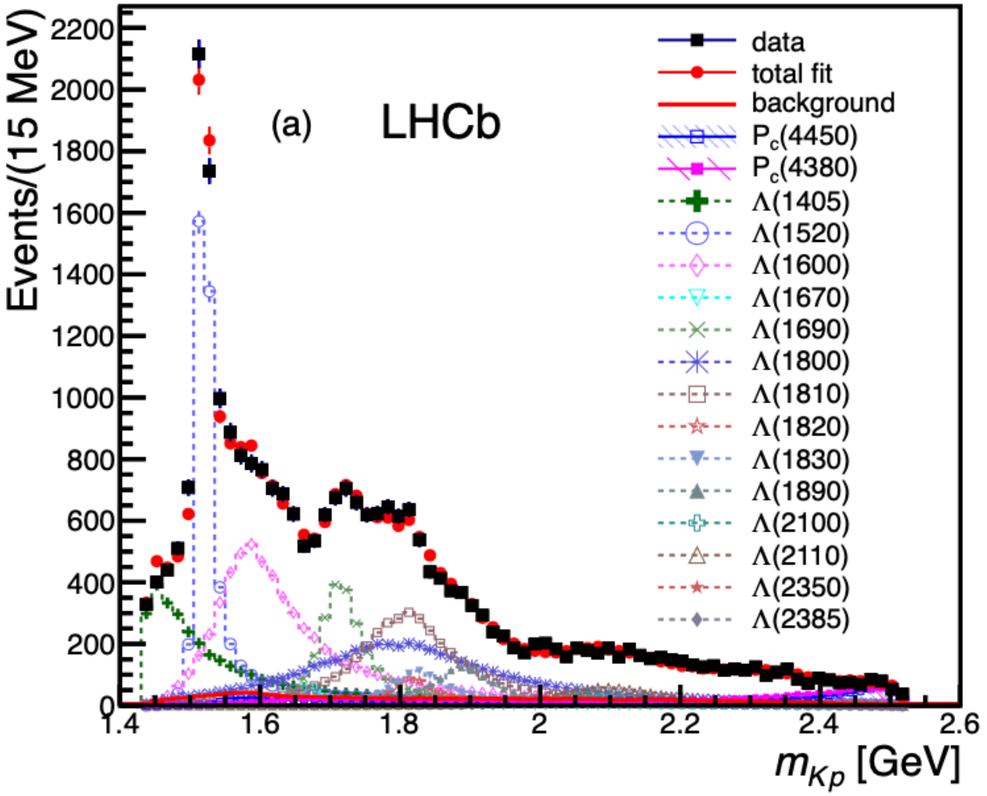
iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."





problems with the traditional approach

i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it

The optimal values of parameters can give a misleading representation of the full posterior distribution. Eg, they are not in the "typical region".

problems with the traditional approach

i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it

a. LASSO **--** add an L1 penalty to the objective function in an attempt to prune parameters

problems with the traditional approach

i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

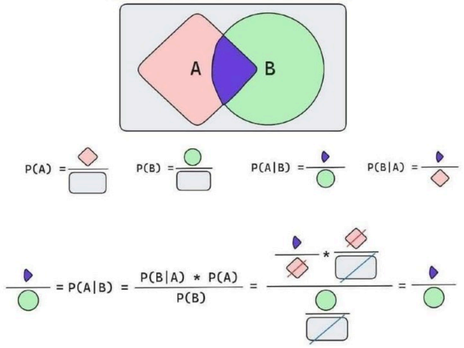
iv. problematic overfitting

v. we wish to extract model structure, not assume it

a. LASSO

b. AIC -- $E_g(2k - 2\log f(\mathscr{D} \mid \hat{\theta})) \rightarrow -2E_g(\log(f(\mathscr{D} \mid \theta))$, ***thus large model are penalized. (g is the generating model)***

problems with the traditional approach



i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it

a. LASSO

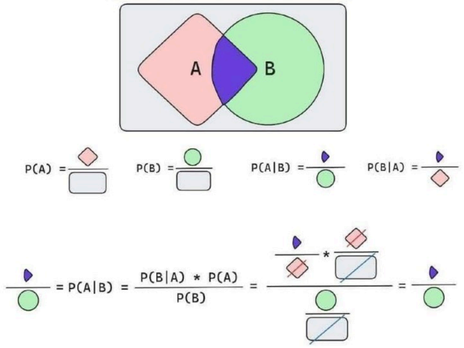
b. AIC

c. *BIC* -- $p(M \mid \mathscr{D}) = \int d\theta\, p(\mathscr{D} \mid \theta; M)\, p(\theta \mid M)\, p(M) \approx \int d\theta \exp[\log L(\hat{\theta}) - \frac{n}{2}(\theta - \hat{\theta}) \cdot I(\hat{\theta}) \cdot (\theta - \hat{\theta}) + \ldots]\, p(\hat{\theta} \mid M)\, p(M)$

$$p(M \mid \mathscr{D}) \approx \exp\left[-\frac{1}{2}(k \ln n - 2 \log L(\hat{\theta})) + O(1)\right] p(M)$$

Large models are penalized more with large data sets.

problems with the traditional approach



i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it

a. LASSO

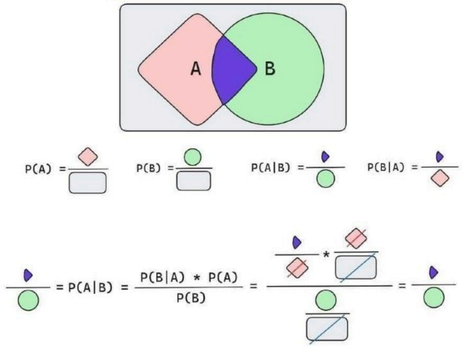
b. AIC

c. BIC

d. Bayesian Model Averaging :

$$p(M \mid \mathscr{D}) = \int d\theta \, p(\mathscr{D} \mid \theta; M) \, p(\theta \mid M) \, p(M)$$

problems with the traditional approach



i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it

a. LASSO

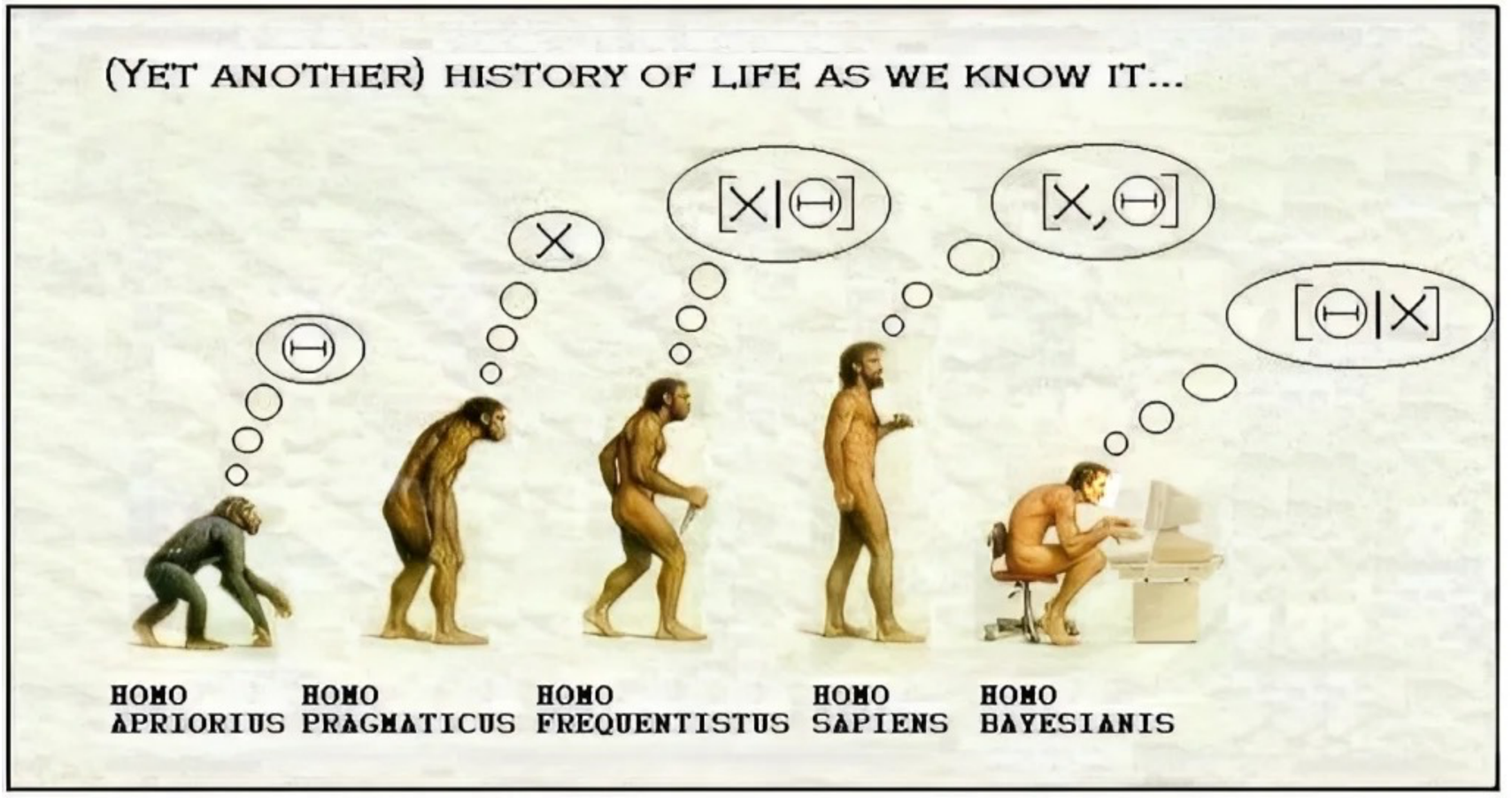
b. AIC

c. BIC

d. Bayesian Model Averaging :

$$p(M \mid \mathscr{D}) = \int d\theta \, p(\mathscr{D} \mid \theta; M) \, p(\theta \mid M) \, p(M)$$

Priors are problematic! What is a uniform prior? Lack of reparametrization invariance. What is the totality of causative agents?


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Θ
X
[X|Θ]
[X,Θ]
[Θ|X]
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

problems with the traditional approach

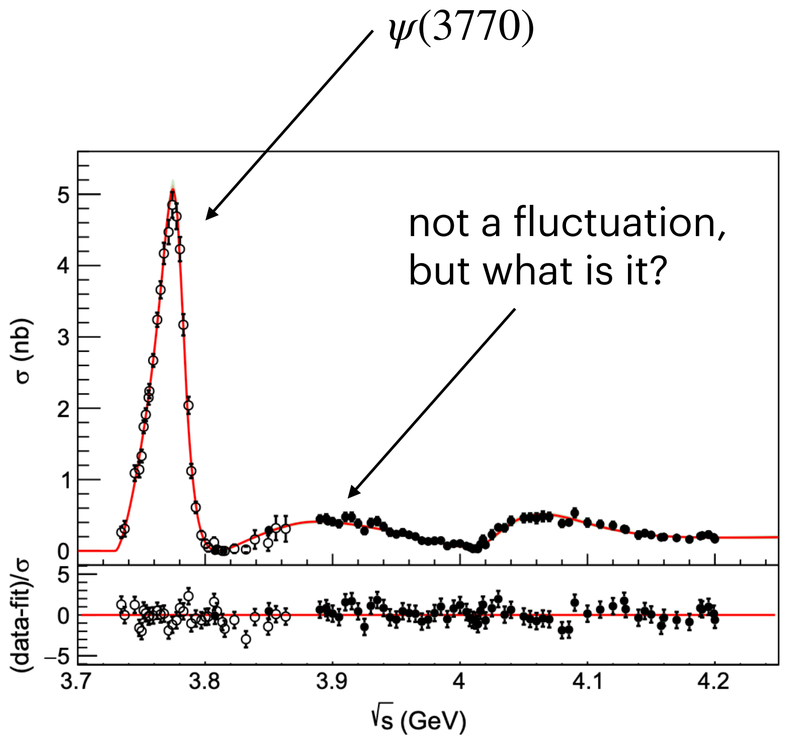
i. fluctuations in the data set may be important

ii. models M0 and M1 are wrong!

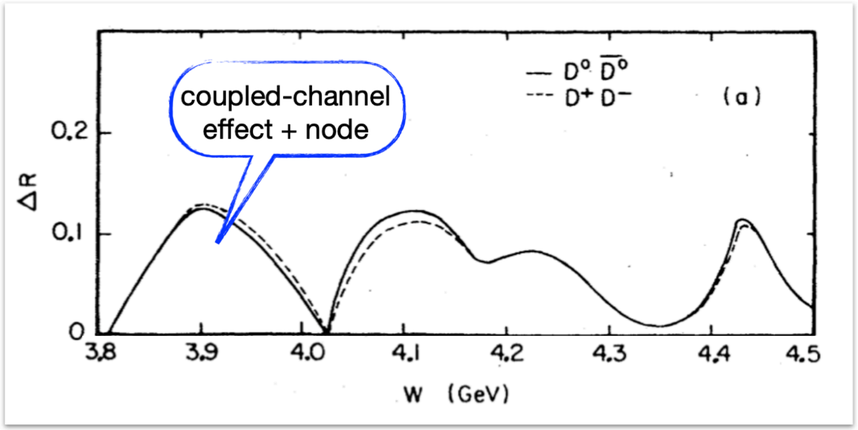
iii. systematic errors are often underestimated

iv. problematic overfitting

v. we wish to extract model structure, not assume it



Eichten et al., Phys. Rev. D 21 (1980) 203



In our calculation there is some weak structure in the 3.9–4.0 GeV region. It does not arise from a $c\bar{c}$ resonance, but from the opening of the $D\bar{D}^* + D^*\bar{D}$ channel and a decrease in the $D\bar{D}$ channel due to a nearby zero in the $3S$ decay amplitude.

additional concerns

i. data sets are not providential
ii. no model is providential
} summary so far

iii. model parameters are (approximately?) meaningless

iv. we are not interested in the model, but rather the analytic structure of the model

additional concerns

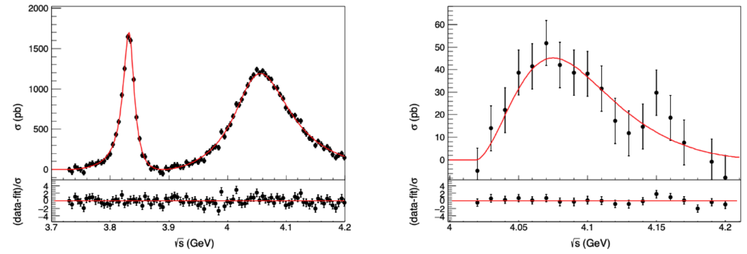
i. data sets are not providential

ii. no model is providential

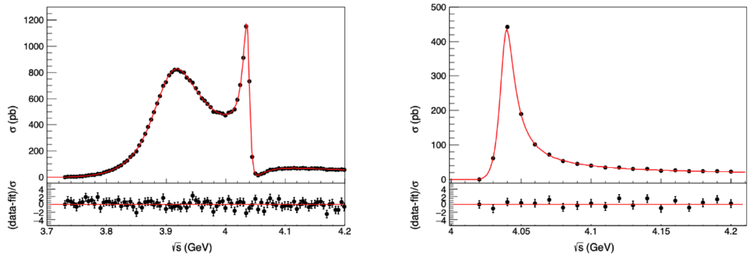
iii. model parameters are (approximately?) meaningless

iv. we are not interested in the model, but rather the analytic structure

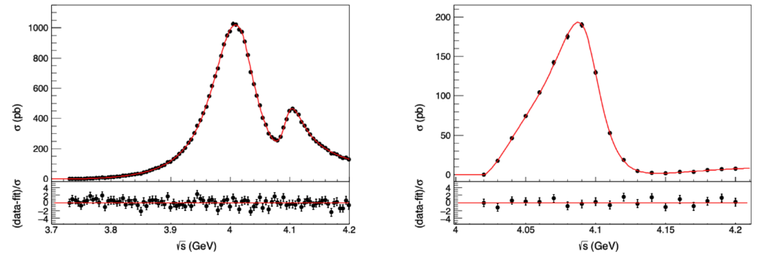
A



B



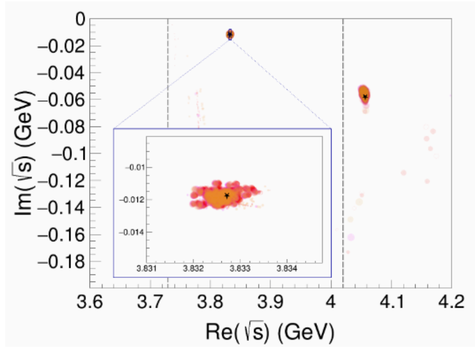
C

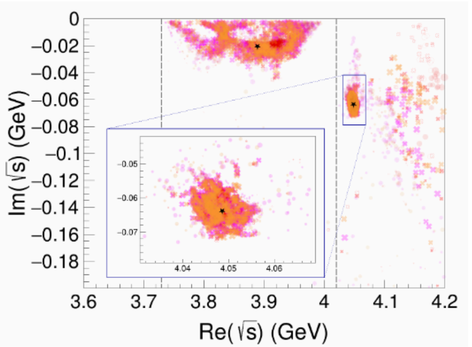


sheet II

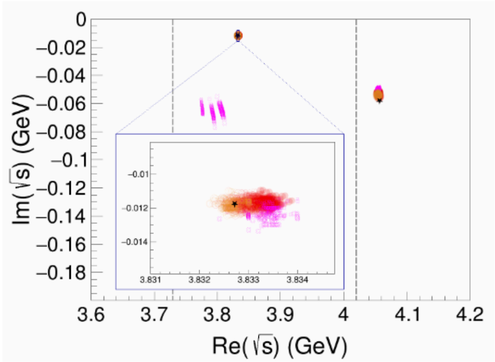
sheet III

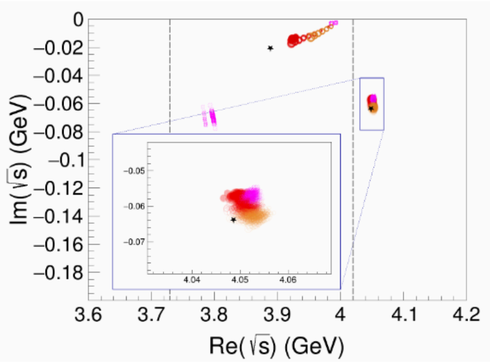
2 bare poles



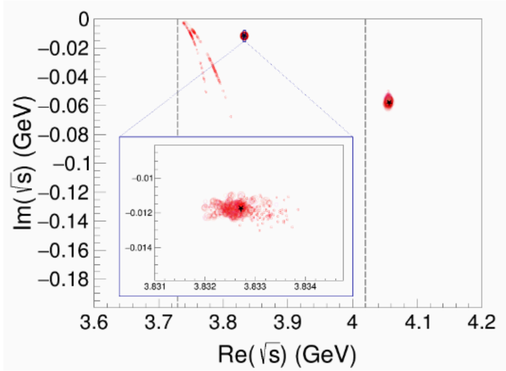


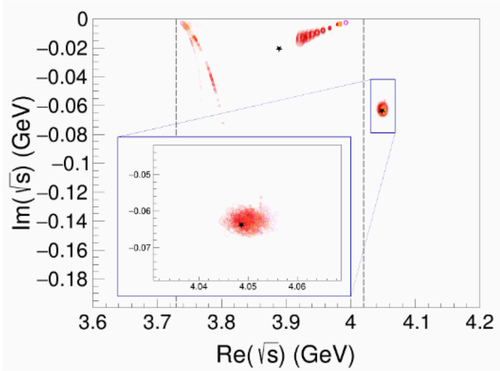
1 bare pole





0 bare poles





additional concerns

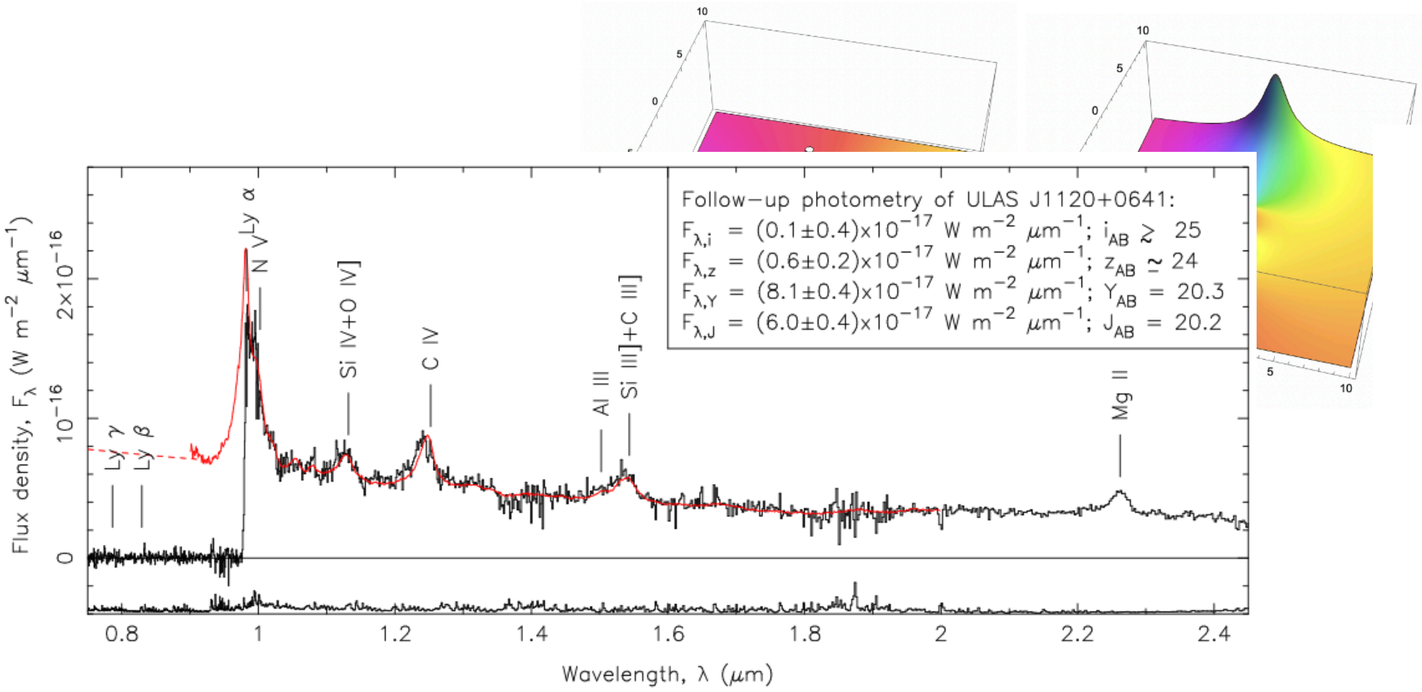
i. data sets are not providential

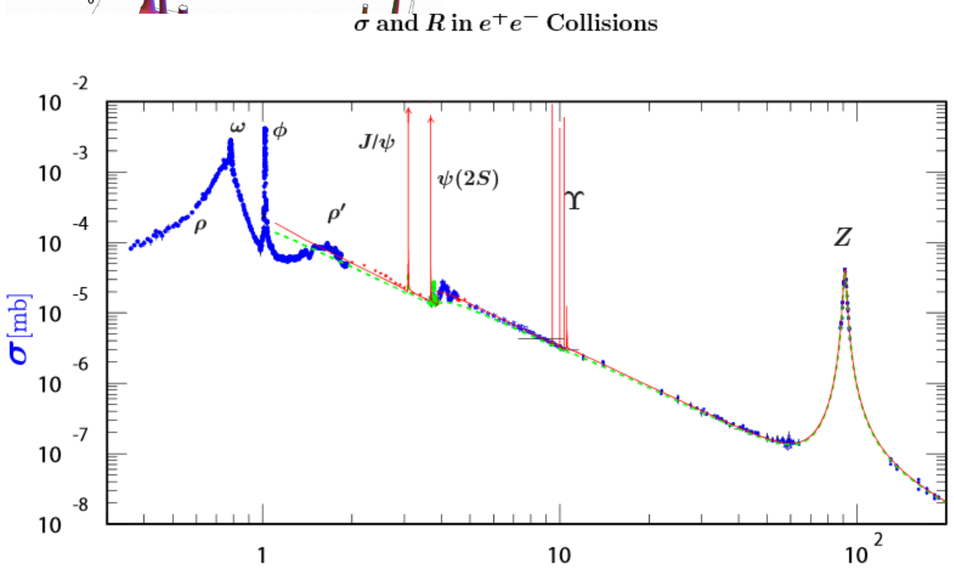
ii. no model is providential

iii. model parameters are (approximately?) meaningless

iv. we are not interested in the model, but rather the analytic structure of the model

this applies to other fields/spectroscopy as well!





additional concerns

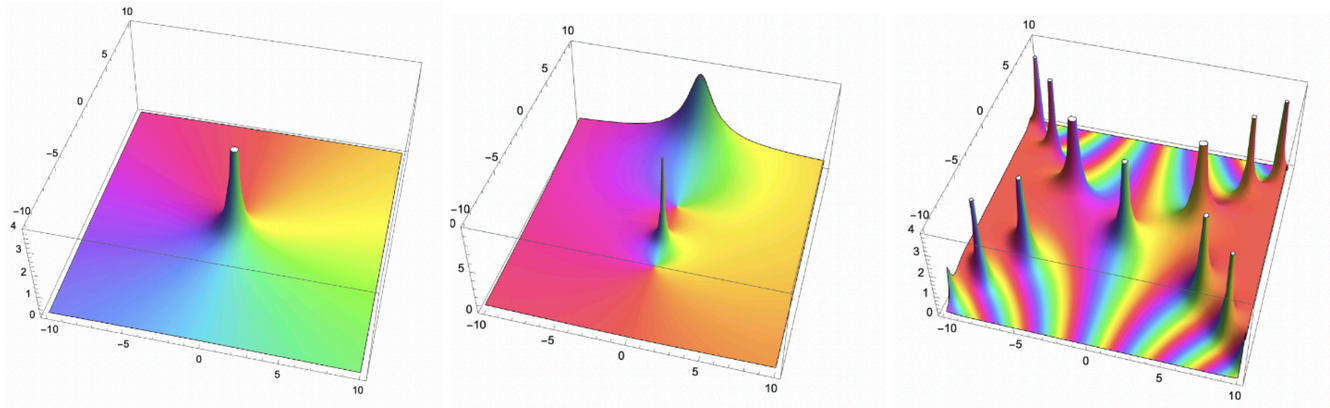
i. data sets are not providential

ii. no model is providential

iii. model parameters are (approximately?) meaningless

iv. we are not interested in the model, but rather the analytic structure of the model

(or, at least, should be!)



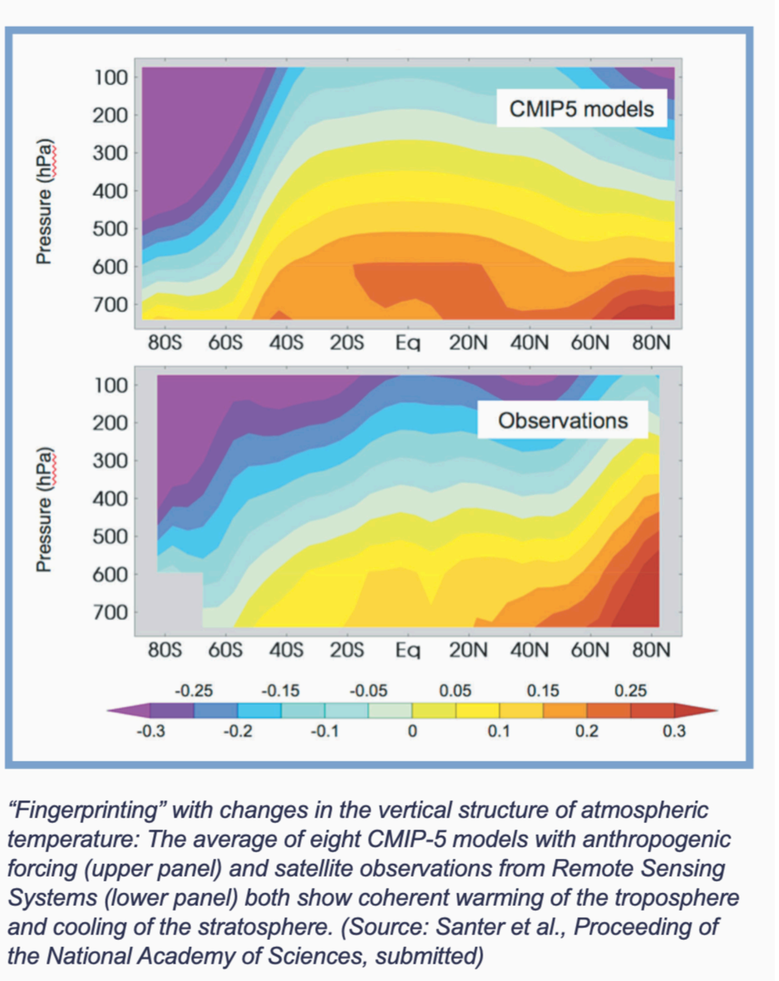
h/t Randall Munroe


I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT YEARS AGO.

bigger problems

i. what does it mean to model?

ii. what does it mean to minimize?



"Fingerprinting" with changes in the vertical structure of atmospheric temperature: The average of eight CMIP-5 models with anthropogenic forcing (upper panel) and satellite observations from Remote Sensing Systems (lower panel) both show coherent warming of the troposphere and cooling of the stratosphere. (Source: Santer et al., Proceeding of the National Academy of Sciences, submitted)

Models should not be thought of as right or wrong, rather skillful or less skillful.

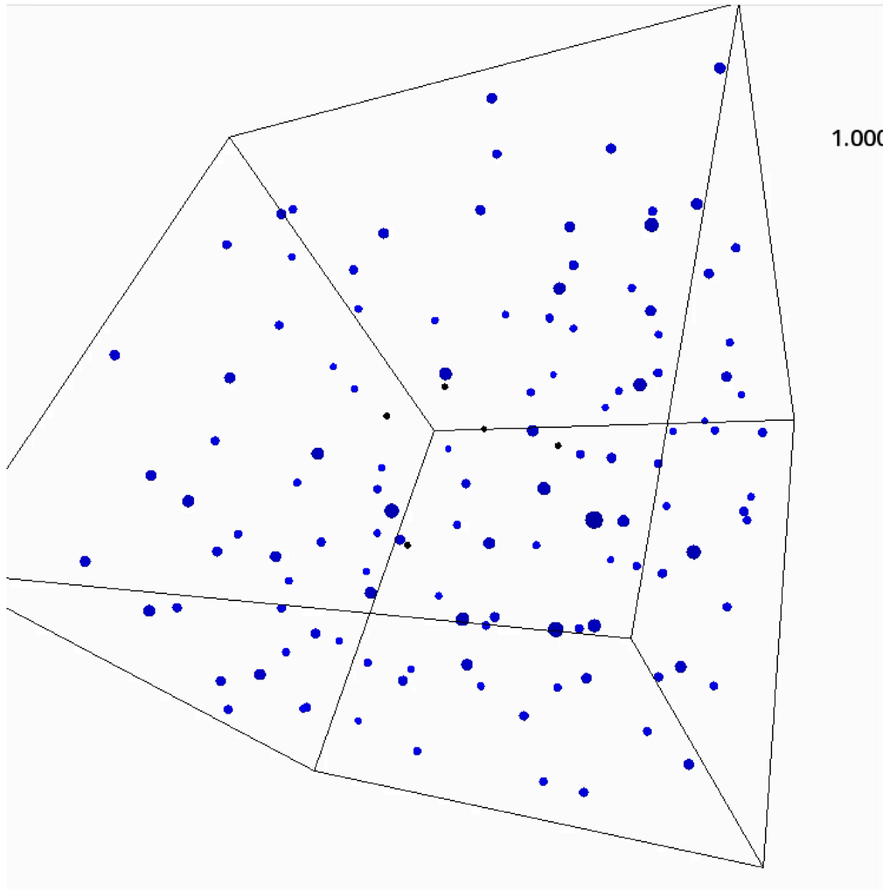
https://web.archive.org/web/20141225022241/http://climatemodeling.science.energy.gov/f/Fact_Sheets/PCMDI_Fact_Sheet.pdf

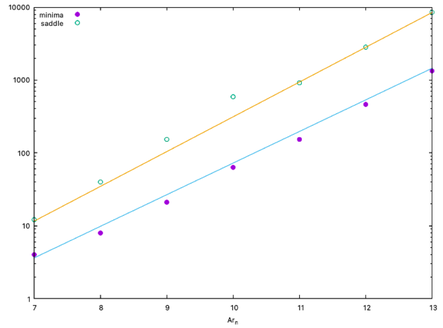
bigger problems

i. what does it mean to model?

ii. what does it mean to minimize?

A 38 element Lennard-Jones system has ~ 10^{14} local minima (!!)



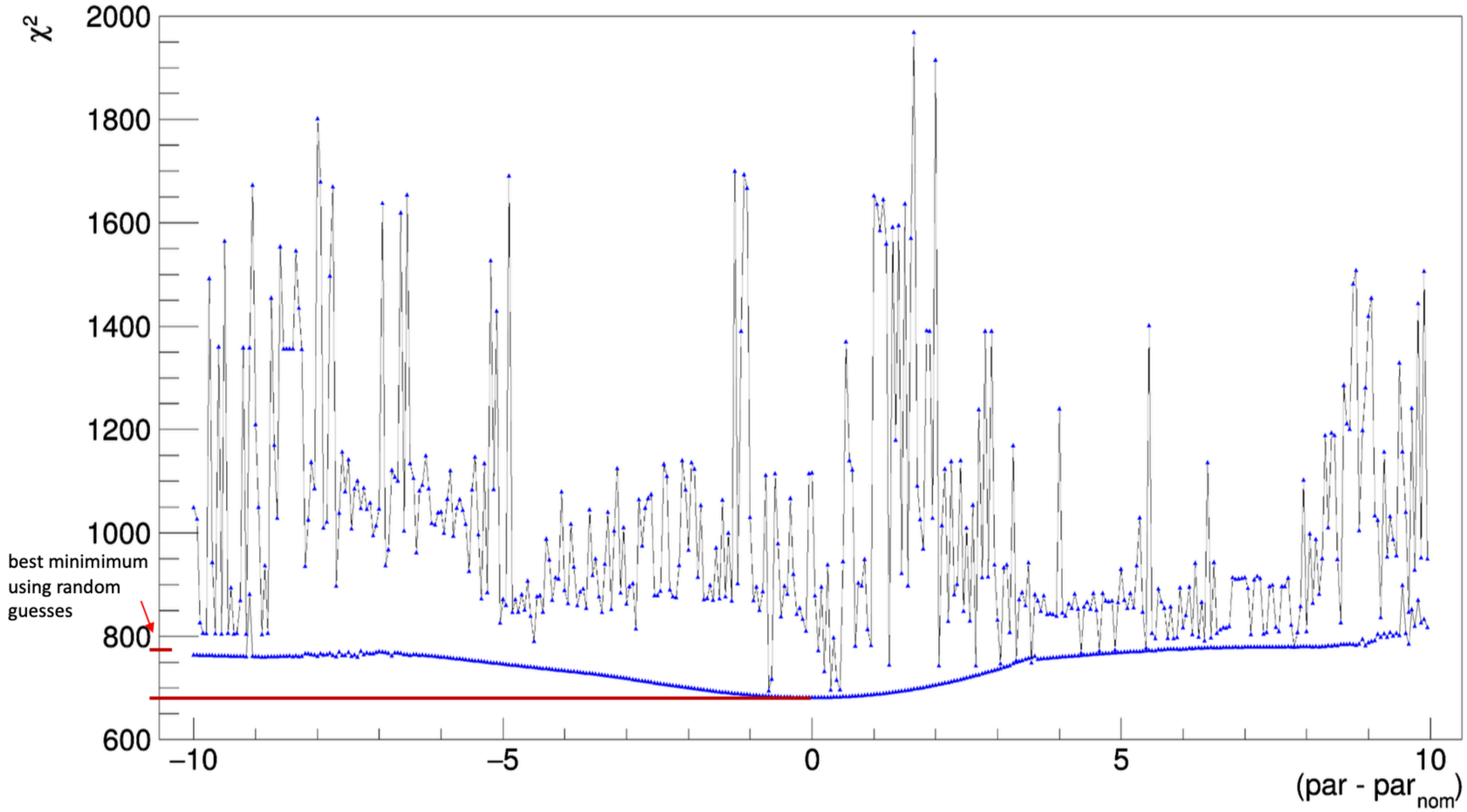


C.J. Tsai and K.D. Jordan, J. Phys Chem, **97** 227 (1993).

bigger problems

i. what does it mean to model?

ii. what does it mean to minimize?



a way forward

a way forward:

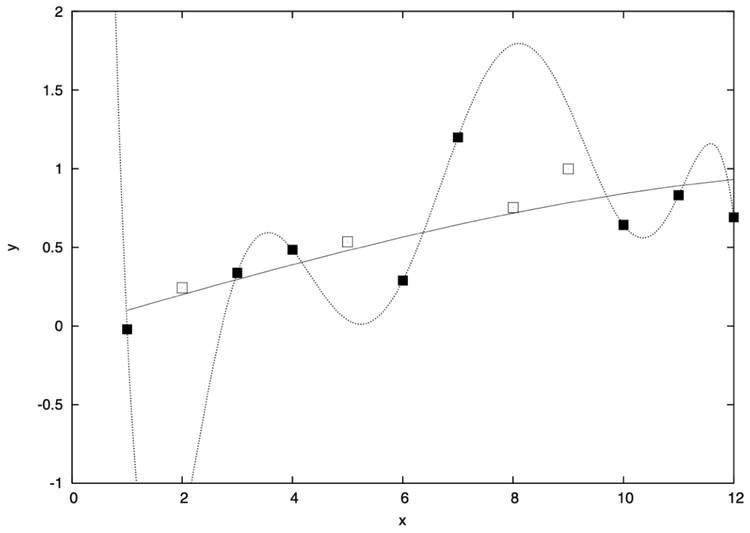
- One way in which one can claim the discovery of a state is if that knowledge permits better, or *predictive*, statements to be made about future experiments.

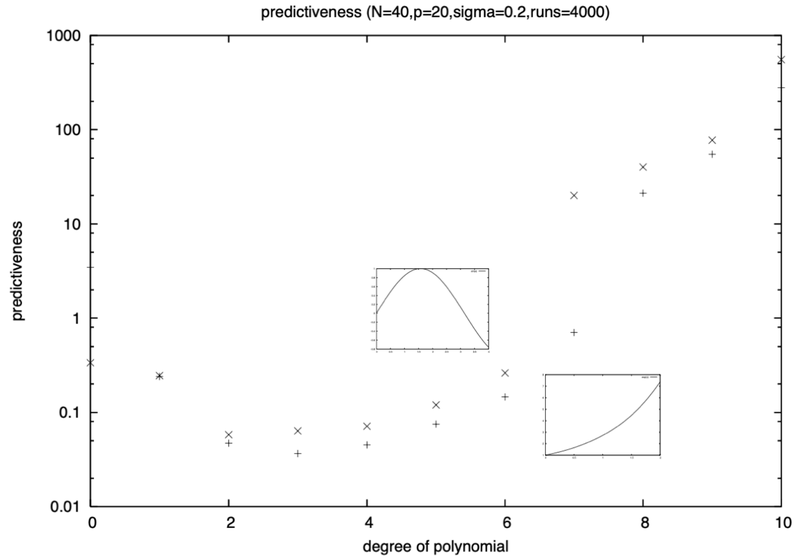
"How well does my model fit the data?" ➔ "How well can I predict the outcomes of future experiments?"

- We can use cross-validation to avoid overfitting.

- Combine these ideas by splitting the data set into training (ante) and validation (post) sets.

A simple example





Bayesian Predictivity

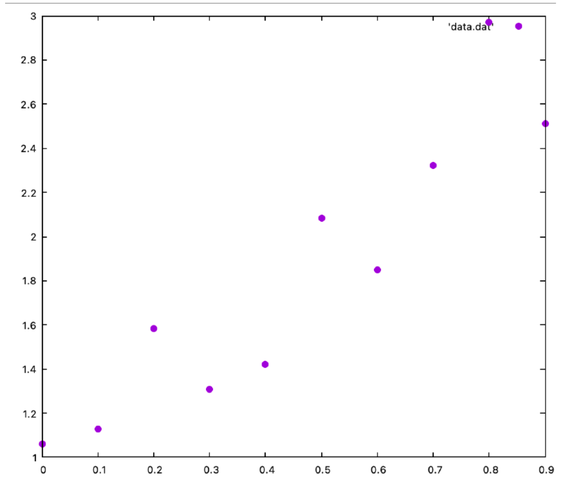
- abandon the idea that we "know" the model -- work in "super-model space" (which, of course, is a model; but now we seek a degree of agnosticism).
- stochastically explore model space. [One could model average by averaging over the trajectory. This is not our primary goal here.]
- explore the continuous portion of model space with Markov chain Monte Carlo. Metropolis-Hastings update according to

$$p(\theta \to \theta') = \min\left(1, \frac{p(\theta' \mid \mathscr{D})}{p(\theta \mid \mathscr{D})} \frac{f_{\theta,\theta'}}{f_{\theta',\theta}}\right) \qquad p(\theta \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \theta)\, p(\theta)}{p(\mathscr{D})}$$

1000 MCMC steps starting at (0.5,0.5); uniform step of size $\in [-0.1, 0.1]$

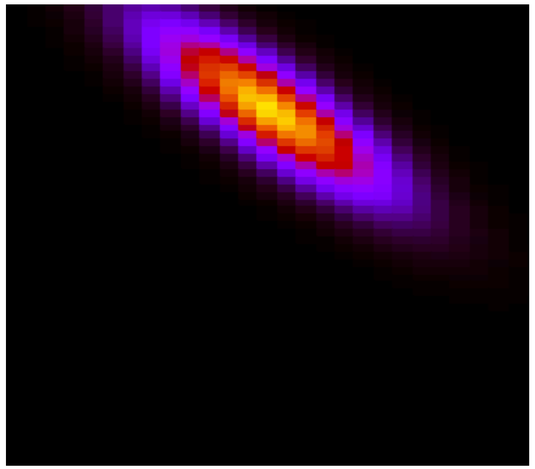
Every point in this space is a model, some are just better than others.

$\mathscr{D}(1 + 2x + \hat{\eta}(0.2))$



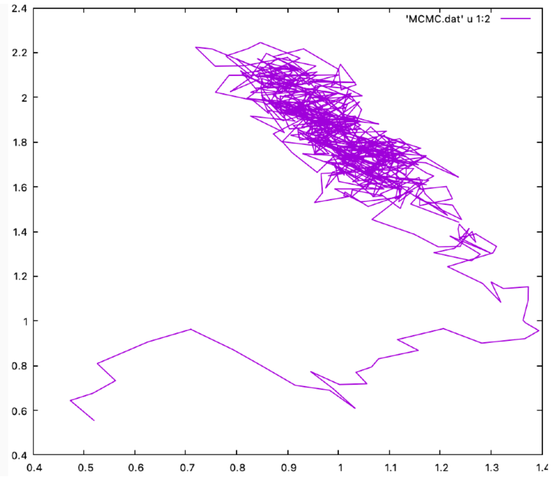
scan of the posterior

b



a

b



a

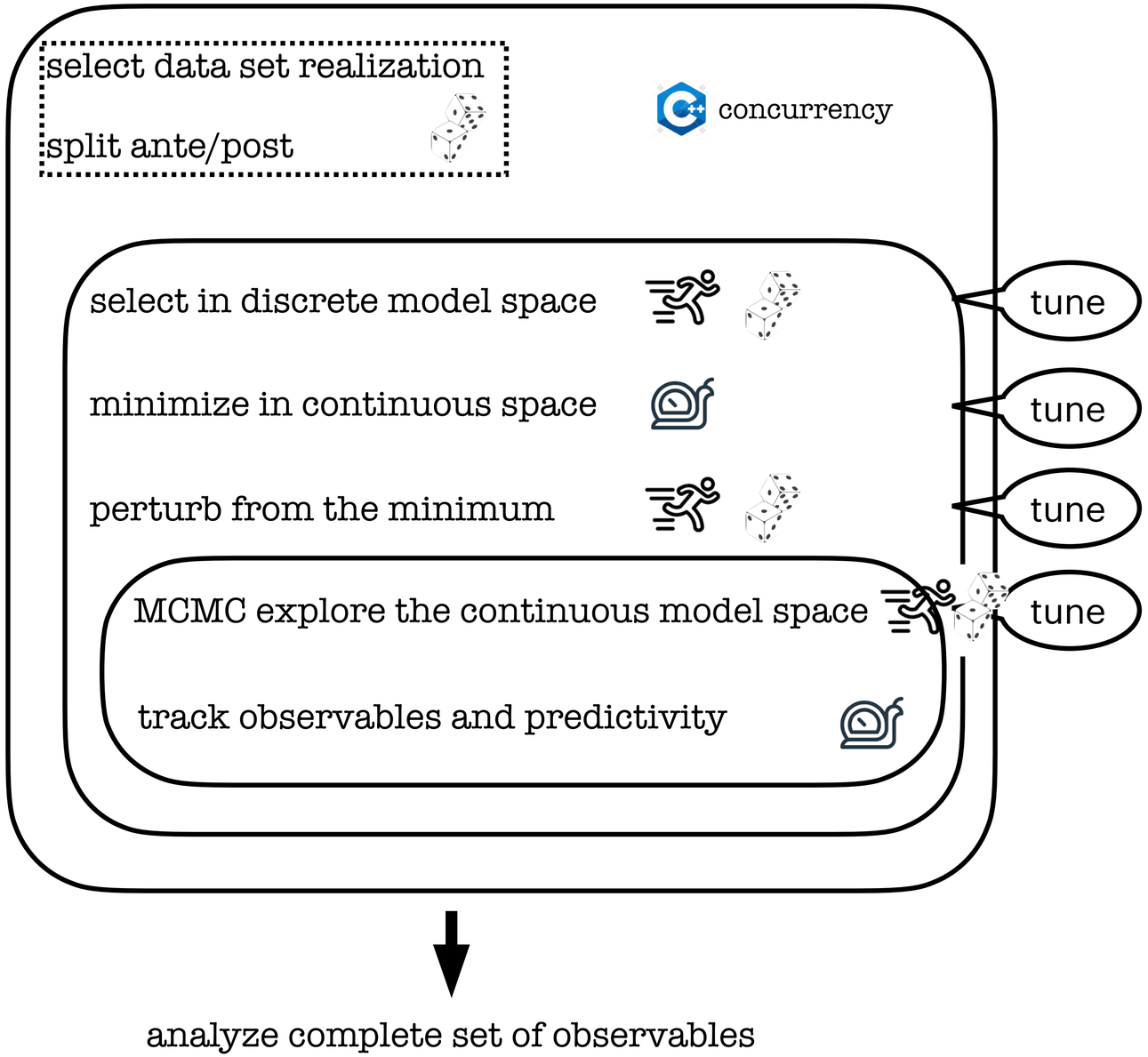
Bayesian Predictivity

- it is not necessary that "minimization" occur! Rather we seek to average over (or explore) model space

  In fact, model comparison can be done with the *evidence*: $p(\mathcal{D} \mid \mathcal{H}_k) = \int d\vec{\theta}\, p(\mathcal{D} \mid \vec{\theta}\, \mathcal{H}_k)\, p(\vec{\theta} \mid \mathcal{H}_k)$, which has no fitting!

- in practice it is useful to minimize a given model in its continuous parameters, perturb this, and use this as a starting point for the MCMC exploration
- exploration in the discrete space is done at random, with minimization at each step
- it's convenient to use logPost$(\vec{\theta} \mid \mathcal{D}) \equiv -2 \log p(\vec{\theta} \mid \mathcal{D})$ and logPost$(\vec{\theta} \mid \mathcal{D}') = -2 \log p(\vec{\theta} \mid \mathcal{D}')$. We want both of these to be small. Thus define $r = \log[\text{logPost}(\vec{\theta} \mid \mathcal{D}) + \text{logPost}(\vec{\theta} \mid \mathcal{D}')]$
- the model prior is chosen to be $\prod_i \exp(-(\theta_i - \hat{\theta}_i)^2 / W^2)$. There is no "complexity penalty" because we use predictivity to select the model, not an anthropomorphic guess
- many data realizations are used concurrently
- there is no need to form a perfect model, only useful ones
- systematic errors should be more accurately estimated

Bayesian Predictivity - Algorithm



focus on predictiveness

de-emphasize fitting and fit quality

explore a large model space to enhance the reliability of the conclusions

use data realizations to enhance the reliability of the conclusions

be as agnostic wrt priors and models as possible

model parameters are not physical

Bayesian Model Averaging

$$p(t \mid D)$$

what we are interested in

$$p(t \mid D) = \sum_{\Theta} p(t \mid \Theta, D)\, p(\Theta \mid D)$$

marginalize over unknown quantities

$$p(t \mid D) = \sum_{\Theta} p(t \mid \Theta, D)\, p(\Theta \mid D) = \sum_{M} \int d\theta_M\, p(t \mid \theta_M, M, D)\, p(\theta_M, M \mid D)$$

split these into M (hyperparameters) and θ_M, parameters.

$$p(t \mid D) = \sum_{\Theta} p(t \mid \Theta, D)\, p(\Theta \mid D) = \sum_{M} \int d\theta_M\, p(t \mid \theta_M, M, D)\, p(\theta_M, M \mid D) = \sum_{M} \int d\theta_M\, p(t \mid \theta_M, M, D)\, p(\theta_M \mid M, D)\, p(M \mid D)$$

apply the chain rule (in this order because $\theta = \theta_M \in \mathbb{R}^n$). This introduces an apparent asymmetry between parameters and hyperparameters.

$$p(t \mid D) = \sum_{\Theta} p(t \mid \Theta, D)\, p(\Theta \mid D) = \sum_{M} \int d\theta_M\, p(t \mid \theta_M, M, D)\, p(\theta_M, M \mid D) = \sum_{M} \overbrace{\int d\theta_M\, p(t \mid \theta_M, M, D)\, \underbrace{p(\theta_M \mid M, D)}_{\text{posterior}}}^{\text{posterior predictive}} \cdot \overbrace{p(M \mid D)}^{\text{model weight}}$$

the posterior represents our modified belief on the prior $p(\theta_M \mid M)$ given the data

$$p(t \mid D) = \ldots \sum_{M} \int d\theta_M\, p(t \mid \theta_M, M, D)\, p(\theta_M \mid M, D)\, p(M \mid D) = \sum_{M} \int d\theta_M\, p(t \mid \theta_M, M, D)\, p(\theta_M \mid M, D) \cdot \overbrace{p(D \mid M)}^{\text{model evidence}}\, p(M) / \overbrace{p(D)}^{\text{evidence}}$$

Bayes theorem

$$p(t \mid D) = \sum_{M} \int d\theta_M\, p(t \mid \theta_M, M, D)\, p(\theta_M \mid M, D) \cdot \int d\theta'\, \overbrace{p(D \mid \theta'_M, M)}^{\text{likelihood}}\, p(\theta'_M \mid M)\, p(M) / p(D)$$

marginalize, implements Occam's razor

Bayesian Model Averaging -- approximations

$$p(t \mid D) = \sum_M \int d\theta_M \, p(t \mid \theta_M, M, D)\, p(\theta_M \mid M, D) \cdot \frac{\int d\theta' \, p(D \mid \theta'_M, M)\, p(\theta'_M \mid M)\, p(M)/p(D)}{}$$

$$\to p(D \mid \hat{\theta}_M, M)\, p(\hat{\theta}_M \mid M)\, (2\pi)^{d/2} \sqrt{\det \Sigma}\, p(M)/p(D)$$

Laplace approximation (steepest descent)

$$\to \frac{1}{I} \sum_i p(D \mid \theta_M^{(i)}, M), \quad \theta_M \sim p(\theta_M \mid M)\, p(M)/p(D)$$

MCMC

approximate $p(\theta \mid M, D)$ as $q_\phi(\theta)$ by minimizing the ELBO ,

and set $p(M \mid D) \approx \exp(ELBO_k)\, p(M)/\text{norm}$

variational Bayes w/ ELBO

LOO, not strictly Bayesian, but practical, and often yields very similar results

$$\to \frac{1}{I} \sum_i p(t \mid \theta_{M_i}^{(i)}, M_i, D)$$

reversible jump Monte Carlo

explain ELBO?

Bayesian Model Averaging -- approximations

$$p(t \mid D) = \sum_M \int d\theta_M \, p(t \mid \theta_M, M, D) \, p(\theta_M \mid M, D) \cdot \int d\theta' \, p(D \mid \theta'_M, M) \, p(\theta'_M \mid M) \, p(M)/p(D)$$

$$p(t \mid M_i, D) = \frac{1}{I} \sum_i p(t \mid \theta^{(i)}_{M_i}, M_i, D) \quad \theta^{(i)}_M \sim p(\theta \mid M, D)$$

MCMC/HMB/NUTS

$$p(D \mid M_i) \, p(M_i)/p(D) \rightarrow$$

Laplace approx around MAP from NUTS

NB: the post/ante split is not necessary in BMA because the Bayes-Occam factor accounts for overfitting. However, one must remain practical!

example applications

a Simple Example

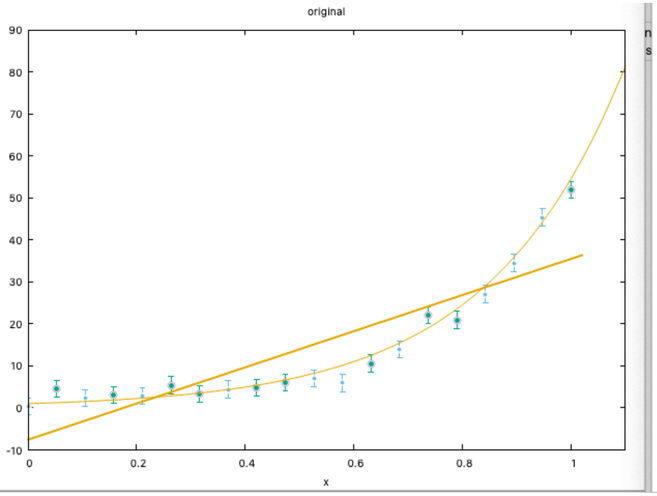
data generated with $\exp(4x) + \hat{\eta}$

fit with polynomials up to degree 11

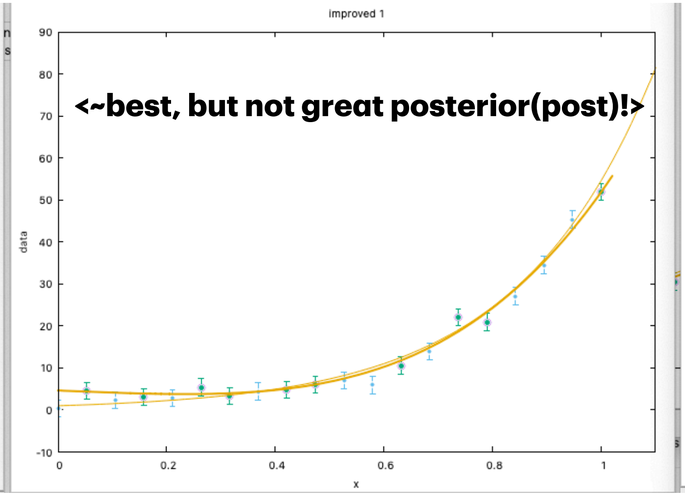
predictiveness(X2) = X2(post)/X2(ante)

predictiveness(posterior) = posterior(post)/posterior(ante)

logPost = - 2 log posterior/n , posterior = p(M|D)



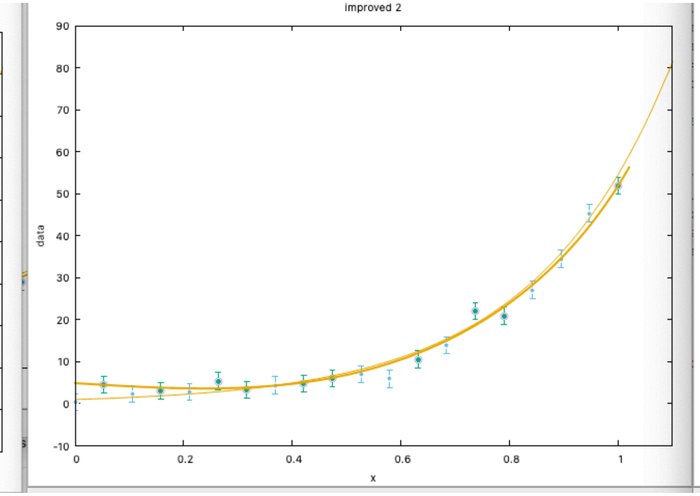

```
initial projector: 1 1 0 0 0 0 0 0 0 0 0 0
 x2(ante) = 18.3654
 predictiveness(X2) = 0.899731
 predictiveness(posterior) = 1581.33
```



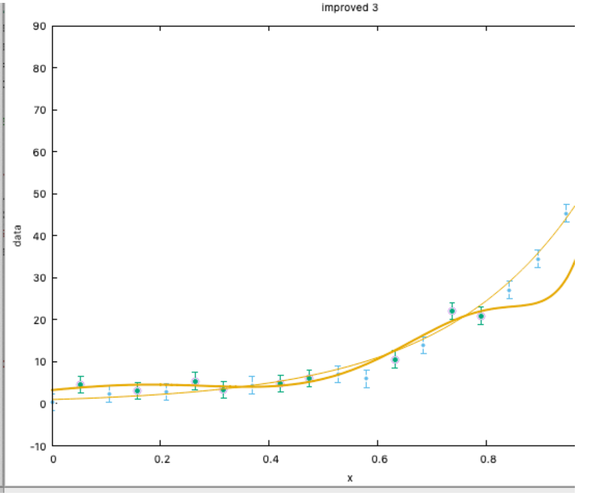

```
 best posterior(ante): 0.0621576
theta: 4.64047 -5.31853 23.2168
29.3319  projector: 1 1 0 1 1 0 0
0 0 0 0 0
 predictiveness(posterior):
0.059387
```

<~best>

```
 best posterior(ante): 0.0521653
theta: 3.7384 0.496066 47.9415
projector: 1 1 0 0 1 0 0 0 0 0 0
0
 predictiveness(posterior):
0.209822
```



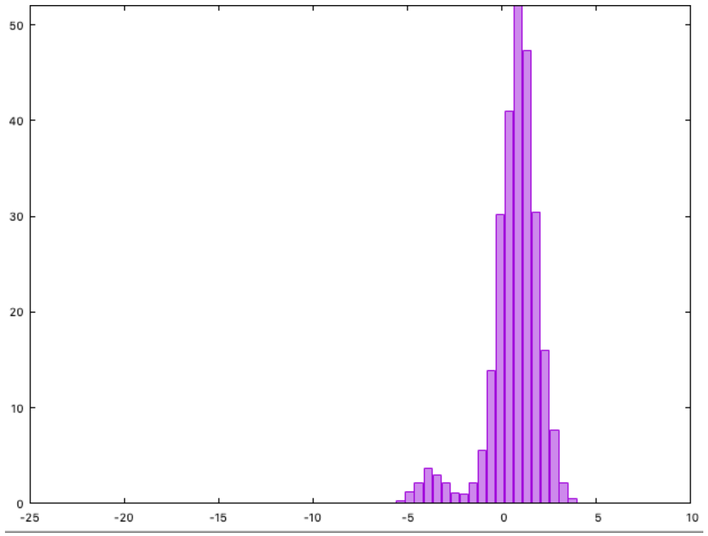

```
 best posterior(ante): 0.0636724
theta: 4.91426 -7.98628 45.6515
3.97777 5.36599  projector: 1 1
0 1 1 0 0 0 0 0 0 1
 predictiveness(posterior):
0.0273971
```



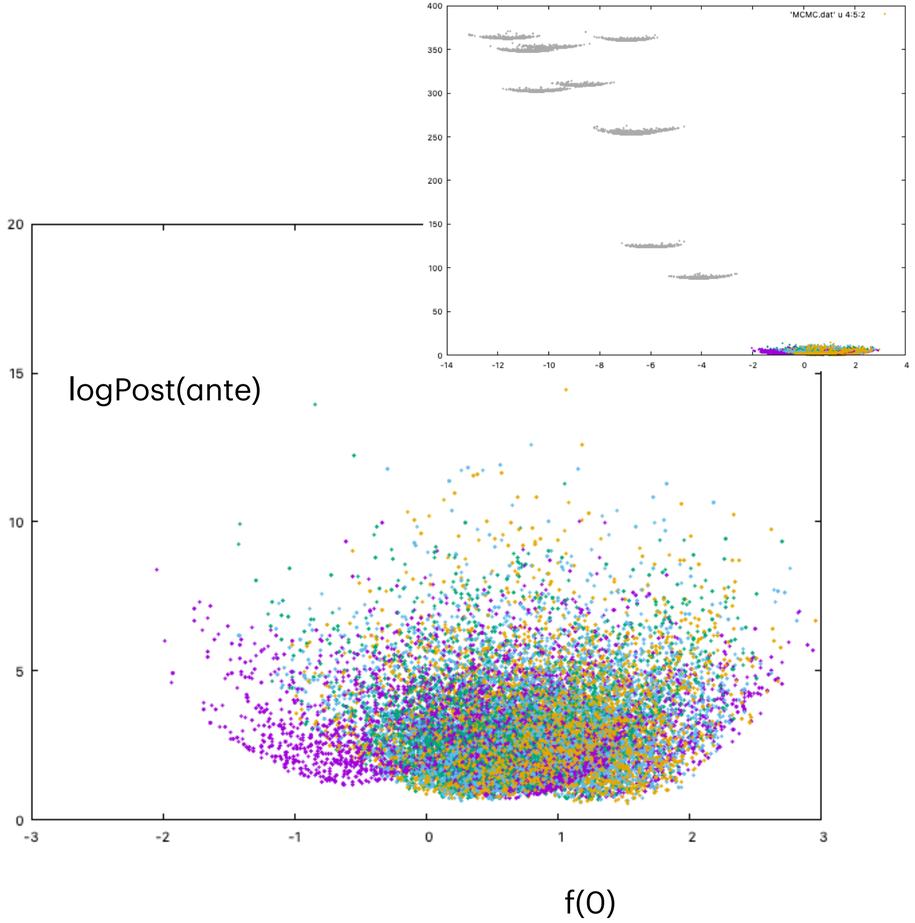

```
 best posterior(ante): 0.126803
theta: 3.26668 11.7815 -212.094
383.643 -1273.17 1138.53
projector: 1 1 0 1 1 0 0 0 0 0 1 1
 predictiveness(posterior):
1.53585e-21
```

a Simple Example

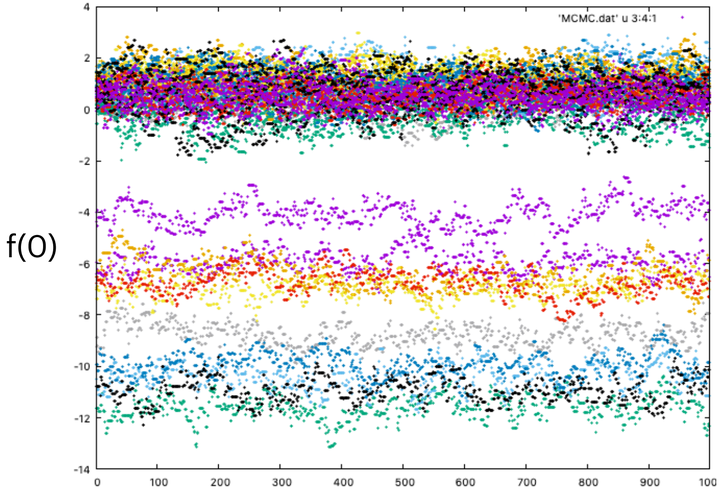
post(ante)*post(post)-weighted
distribution of f(0) over all models



distribution of f(0) vs logPost(ante)

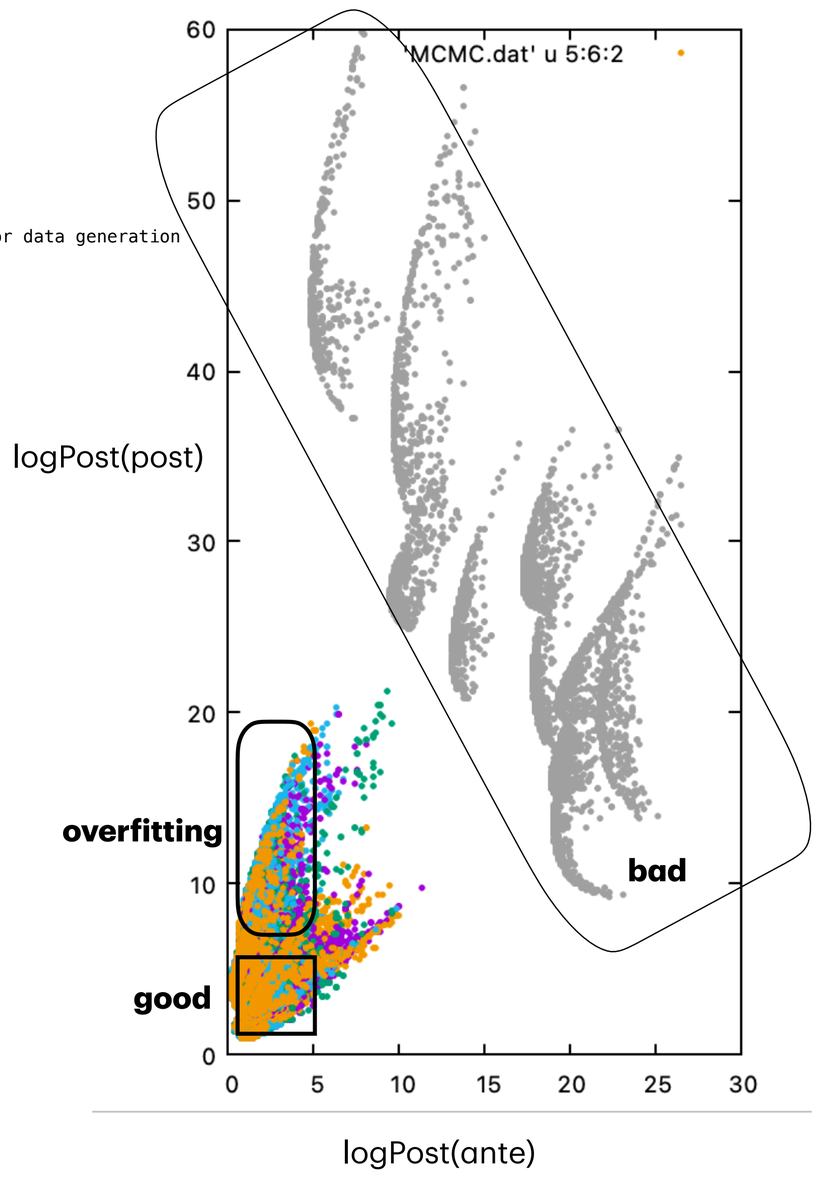


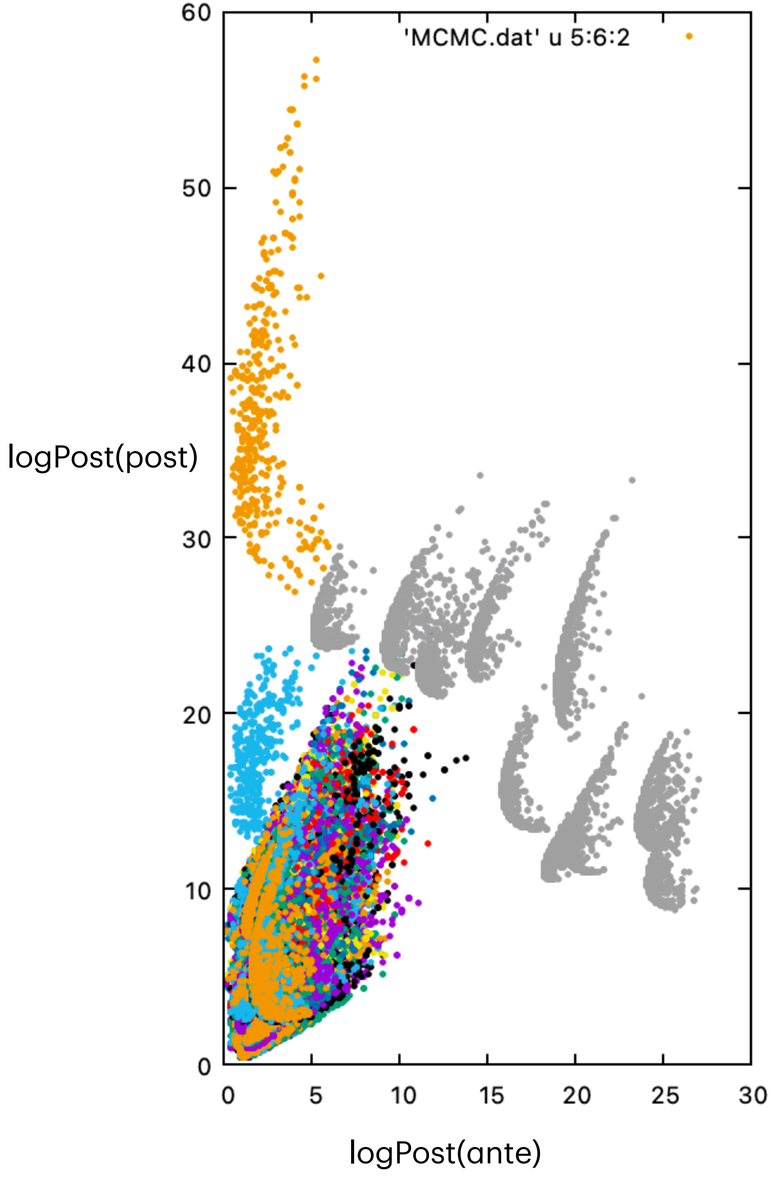
f(0) in MCMC runs



a Simple Example

```
./MA10
 enter verbose [0/1], nthreads, seed, max model size, number of data points, sig for data generation
0 10 98123 14 20 2.
 enter priorWidth, penalty power
3. 1.
 enter NMC, MCMC step size, initial model size
500 .7 2


 acceptance: 0.834
 acceptance: 0.784
 acceptance: 0.83
 acceptance: 0.786
 acceptance: 0.788
 acceptance: 0.772
 acceptance: 0.844
 acceptance: 0.784
 acceptance: 0.812
 acceptance: 0.804


<f(0)> = 1.0484+/-1.40866 (0.0089093)
```



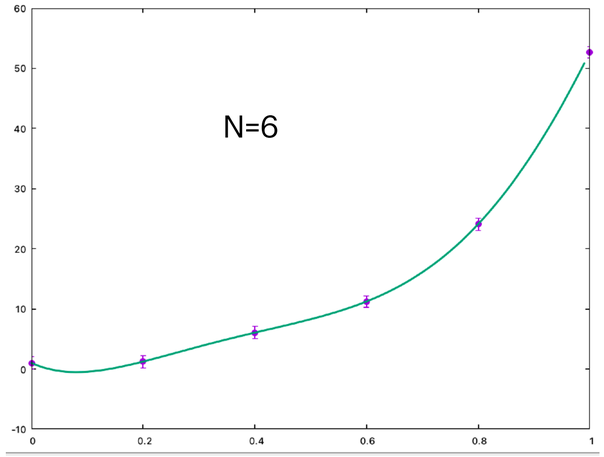


a Simple Example

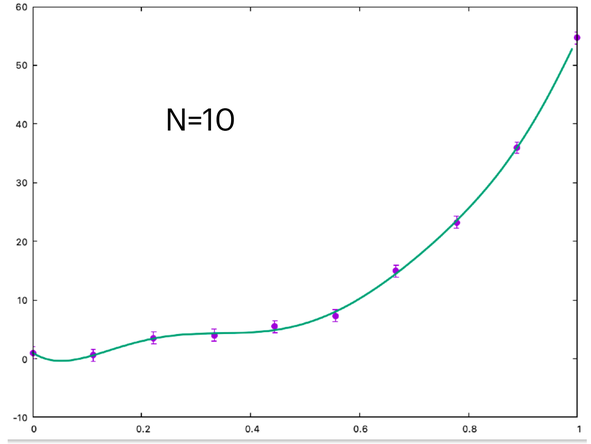
A note on overfitting in practice: Minuit2::migrad is not able to do it past ~10 parameters, so this amounts to an effective regularization!

<more iterations does not help!>

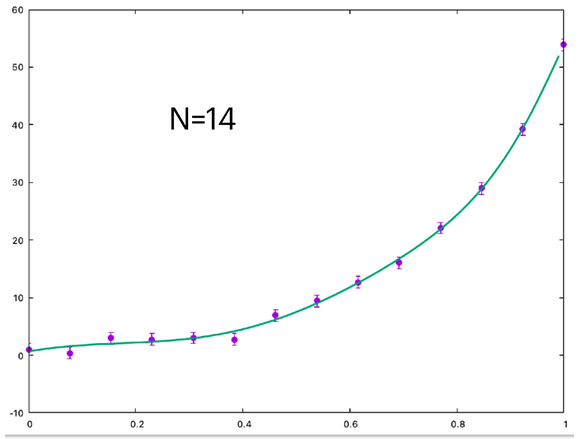
```
X2= 6.9064e-24
params: 1.00336 -42.4893 353.833 -839.63 860.278 -280.321
```



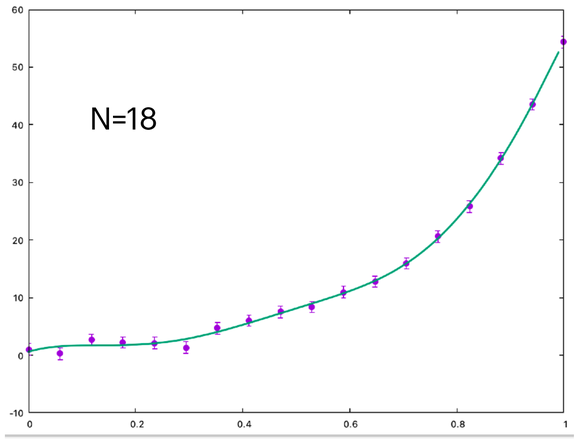

```
1.24668
params: 1.00538 -60.3631 789.056 -2984.12
4126.72 -90.5914 -3377.75 -255.063 3525.54
-1619.76
```




```
5.98328
params: 0.634421 16.7123 -71.2681 121.371 64.9345 -39.0555
-89.3732 -64.1847 -3.0411 57.2922 82.7591 61.633 12.6474
-97.173
```

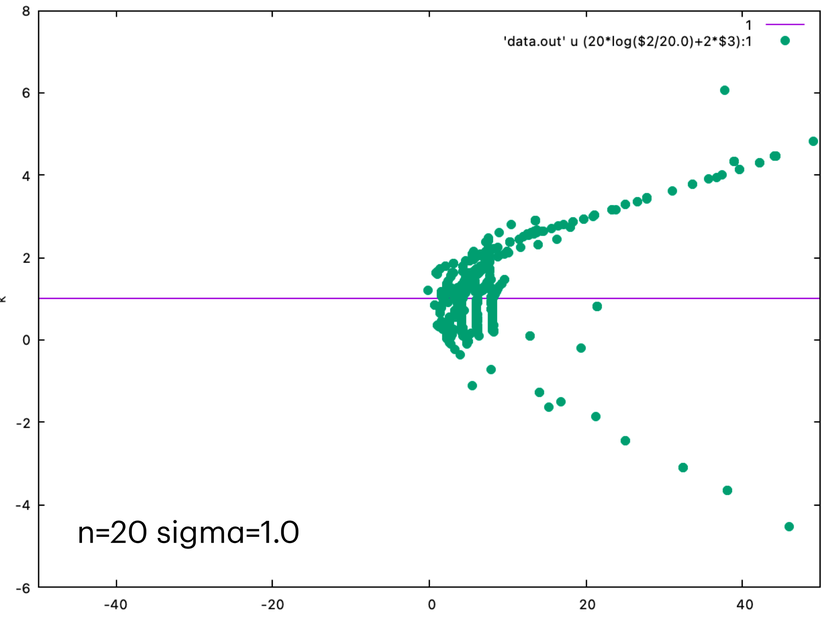



```
6.5082
params: 0.610105 26.2358 -214.995 662.609 -329.239
-666.706 34.2594 579.669 540.124 104.077 -353.569
-518.84 -370.369 -22.6922 333.833 478.568 256.046
-485.291
```

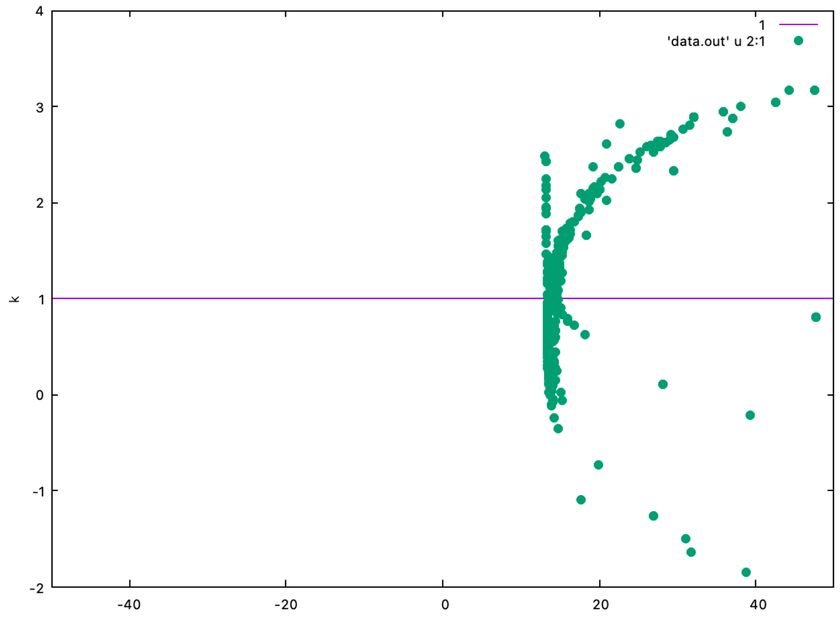


a Simple Example

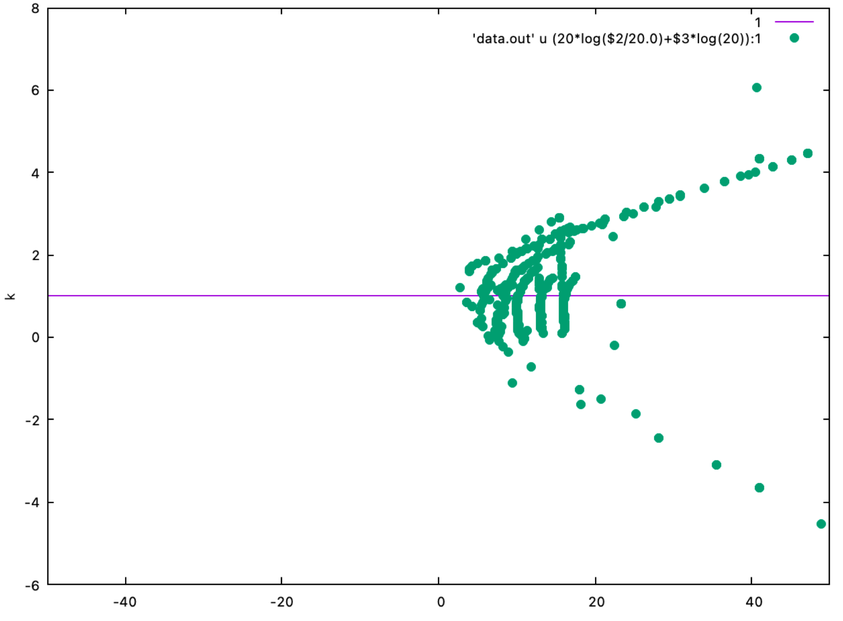
Comparison of <f(0)> with 100 polynomials models, weighted by AIC(unbinned), X2, BIC, AIC(binned)



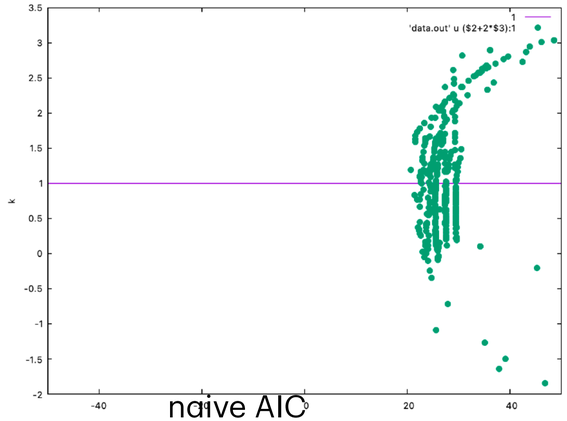
AIC



X2



BIC

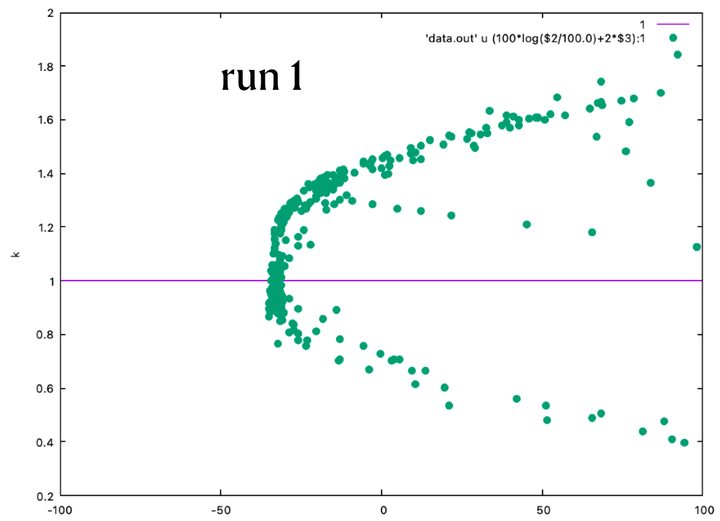


naive AIC

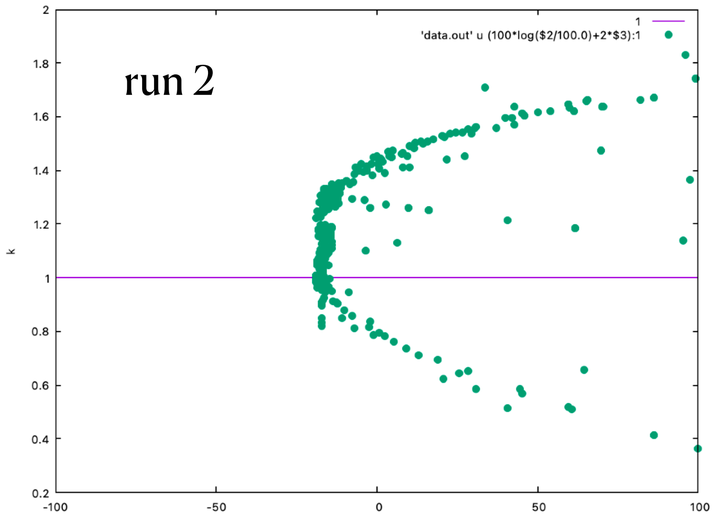
```
 ./MAsimple
 enter verbose [0/1], seed, model space size, max params, number of throws/model size, number of data points, sig for data generation
0 6217 12 8 100 20 1.
```

a Simple Example

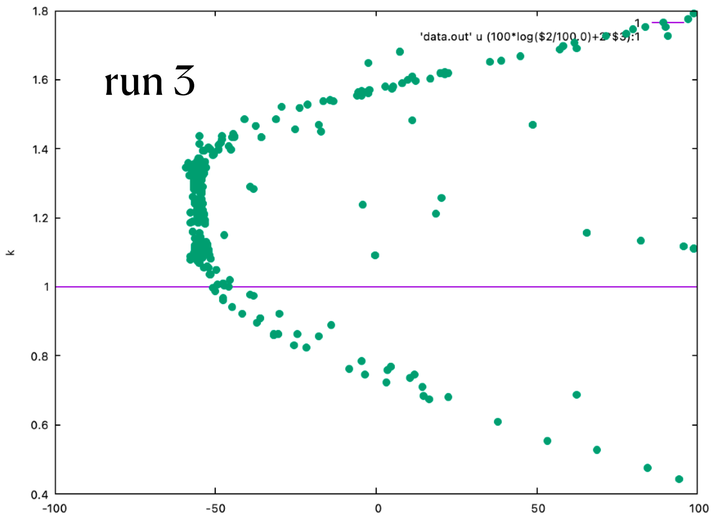
Fit accurate data, plot <f(0)> vs AIC(binned)



AIC



AIC



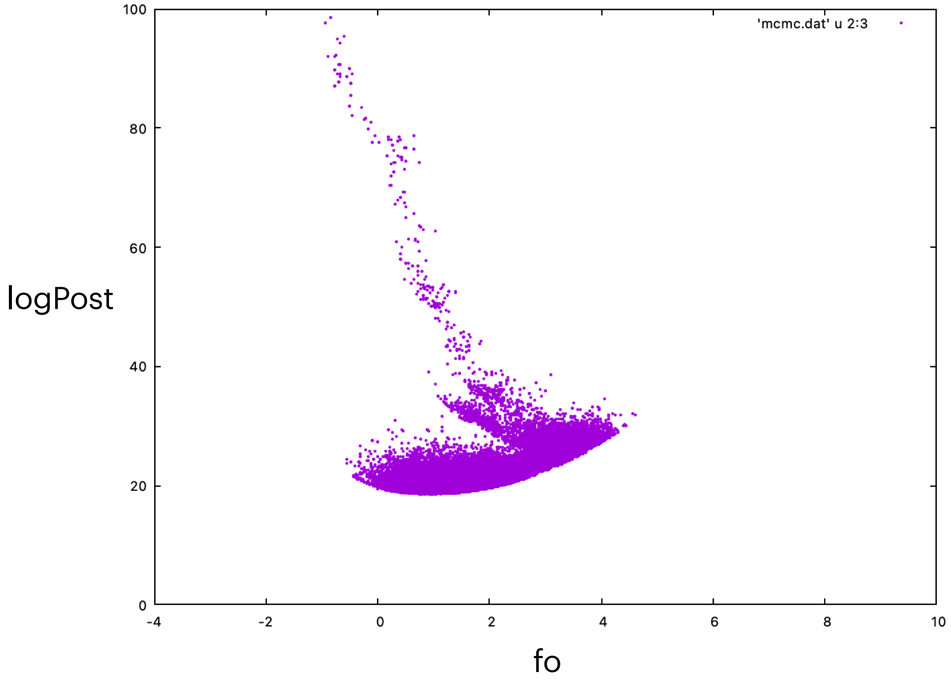
AIC

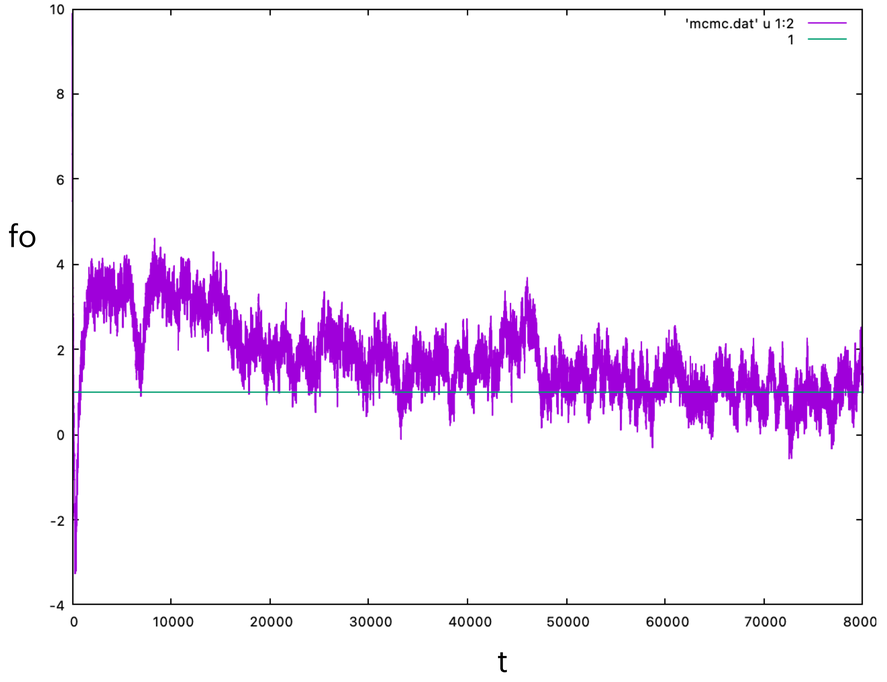
<<multiple runs show the importance of fluctuations in the dataset, even for large n and small sigma!

```
./MAsimple
 enter verbose [0/1], seed, model space size, max params, number of throws/model size, number of data points, sig for data generation
0 2891 12 8 100 80 0.1
```

a Simple Example

starting at a bad point (10,10,10,10), MCMC *is* able to solve the problem, which is nice. This logPost is about X2_min ✔



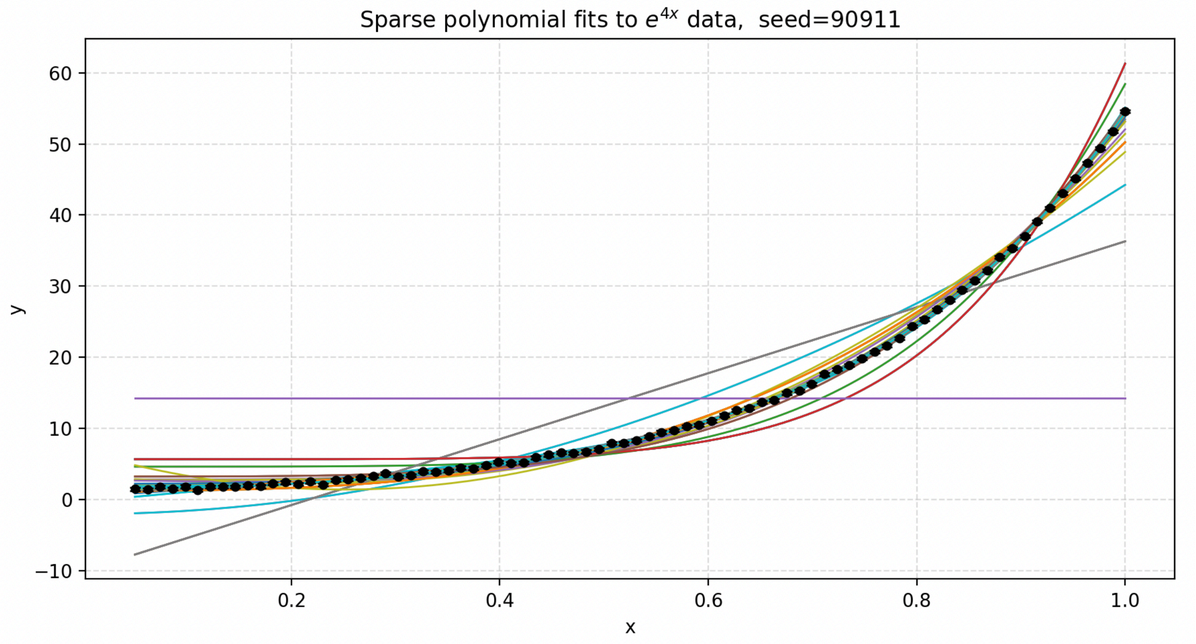


Of course, we are doing NO model averaging here, this is the result for a single model, and we will do no better than a decent minimizer... (IS THIS RIGHT? we could consider the parameter space as a model space ... so we are now comparing estimation with $\hat{\mathscr{L}}$ and some model average...

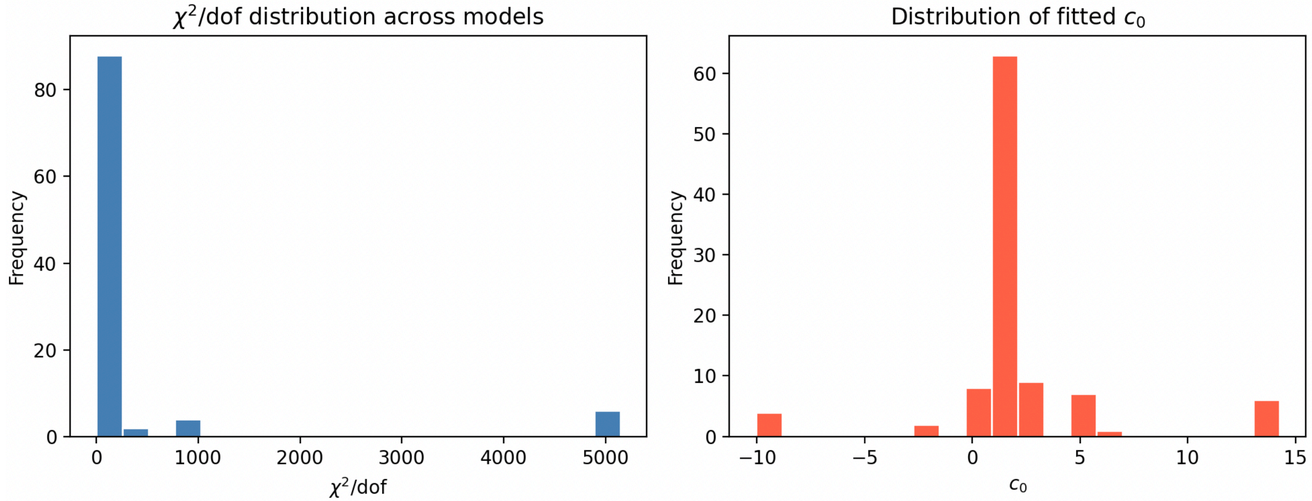
BUT we gain the variable distribution (see ff)

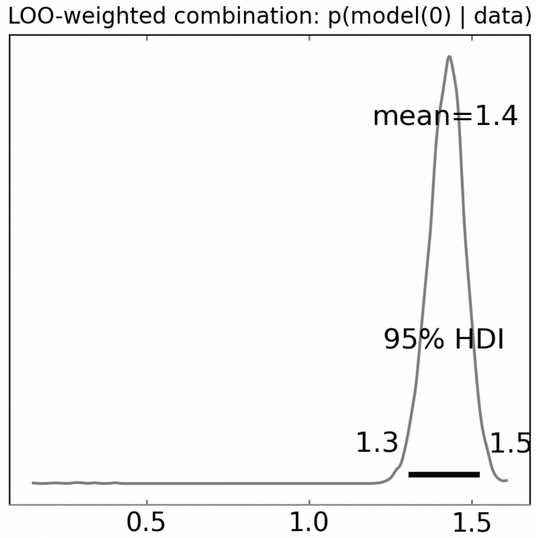
```
 ./BMA
seed, model space size, max params,  number of data points, sig for data generation, lambda, lambda2
9012901 4 4 20 1. 0.0 0.
 enter NMC, stepSize, priorWidth for MCMC
80000 .3 10000.
projector: 1 1 1 1
 X2 = 17.045 f0 = 0.638855
params: 0.638855 13.663 -45.0682 84.0646
 MCMC acceptance = 0.55025
 MCMC points in mcmc.dat
 MCMC f0 = 1.38318
```

a Simple Example



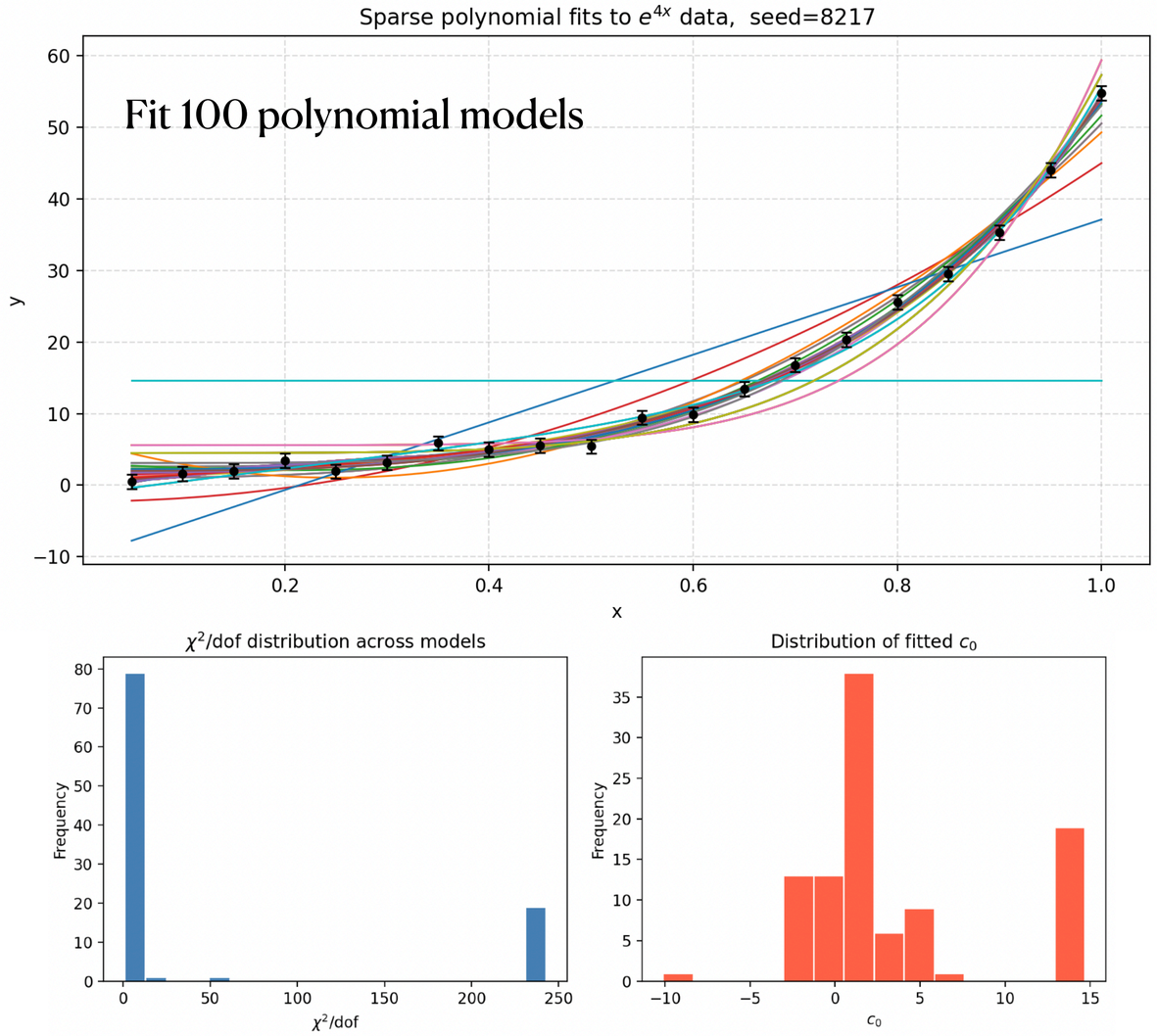

```
time python3 poly4.py
Enter seed: 90911
Enter N (number of points): 80
Enter sigma: .2
Enter nMod (number of polynomial models): 100
Enter nMax (maximum polynomial degree): 6

Chi^2-weighted average of c[0]:
  Weighted c[0] = 1.5532 +/- 0.1319

AIC(binned)-weighted average of c[0] :
  Weighted c[0] = 1.4784 +/- 0.1320


 LOO-weighted c0: mean=1.4192, 95% HDI=[1.3050, 1.5239]
```



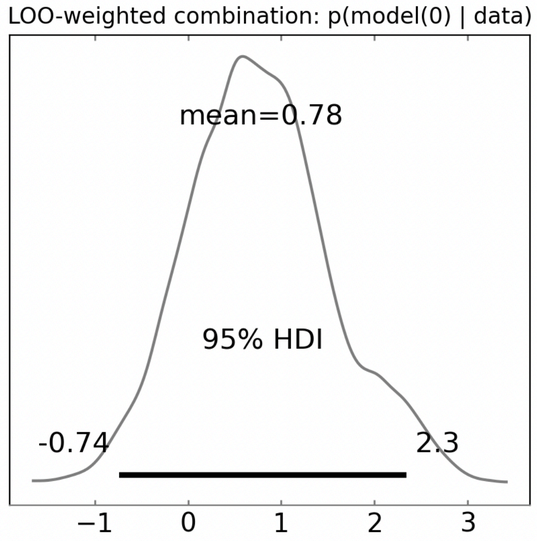


a Simple Example

Fit 100 polynomial models; AIC-average; MCMC the posterior and wieght w/ LOO




```
time python3 poly4.py
Enter seed: 8217
Enter N (number of points): 20
Enter sigma: 1.
Enter nMod (number of polynomial models): 100
Enter nMax (maximum polynomial degree): 6

Chi^2-weighted average of c[0]:
  Weighted c[0] = 0.4266 +/- 1.4093

AIC(binned)-weighted average of c[0] :
  Weighted c[0] = 0.5030 +/- 1.2512

LOO-weighted c0: mean=0.7786, 95% HDI=[-0.7381, 2.3462]
```



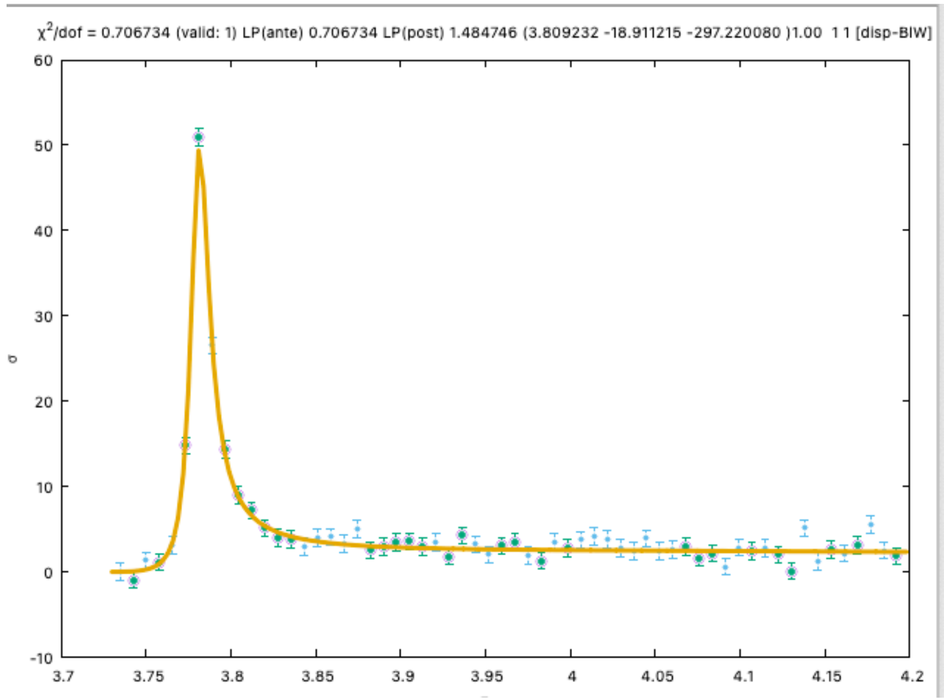
HDI = high density interval (slight generalization of the credible interval)

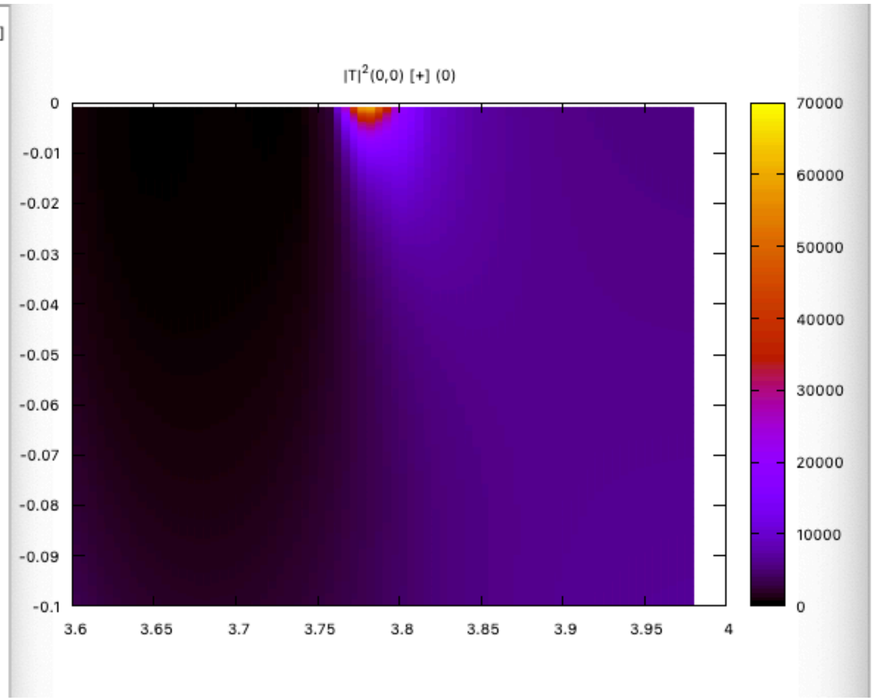
Single Channel K-matrix

Generate synthetic cross section via $\mathscr{M}^{-1} = K^{-1} - i\rho \rightarrow K^{-1} + C$ with $C = -\int_{s_0} \frac{ds'}{\pi} \frac{\rho(s')g^2(s')}{s' - s - i\epsilon}$.

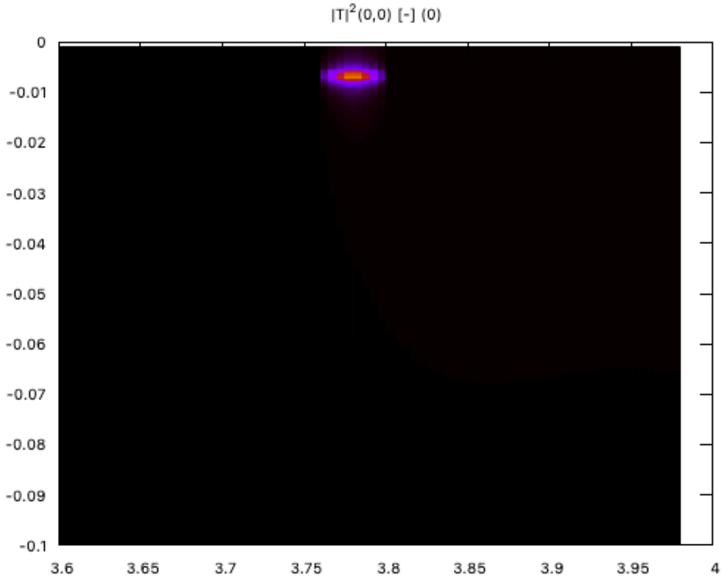
K-matrix parameterization comprises model space

$$M(\vec{\theta}; R, N, Q, C) \rightarrow K_{\mu\nu} = \sum_{r=1}^{R} \frac{g^{(Q)}_{R:\mu} g^{(Q)}_{R:\nu}}{m_R^2 - s} + \sum_{i=0}^{N} c^{(i)}_{\mu\nu} s^i$$





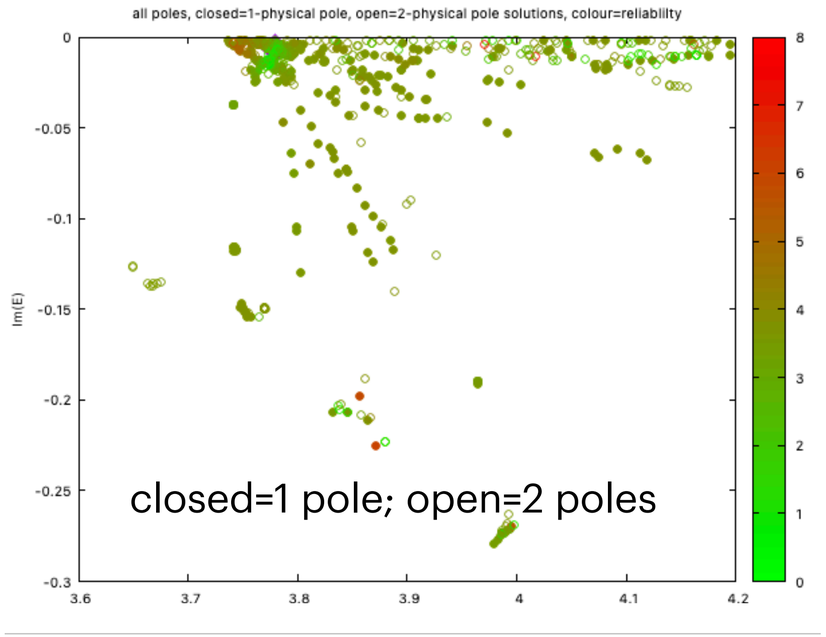
pole location on sheet II

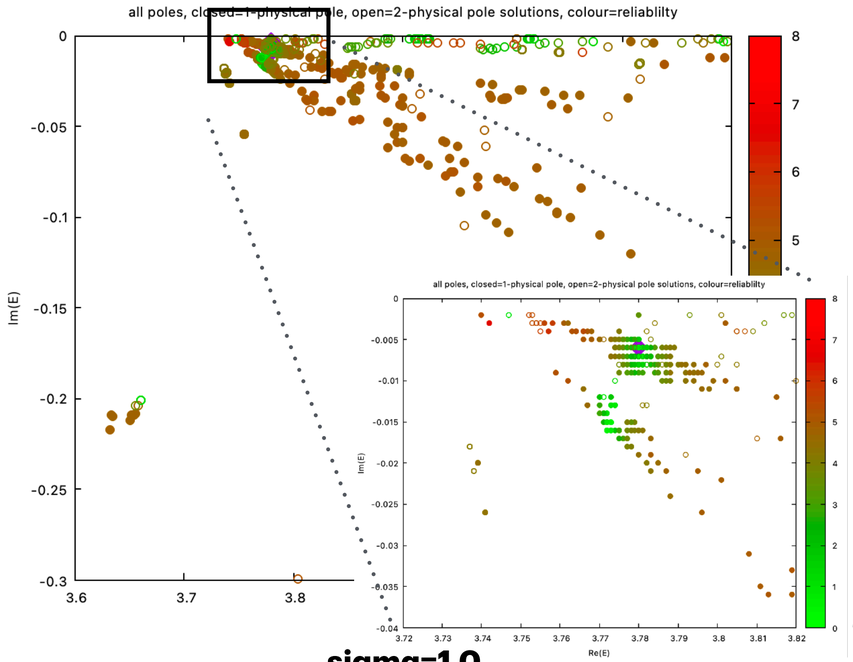


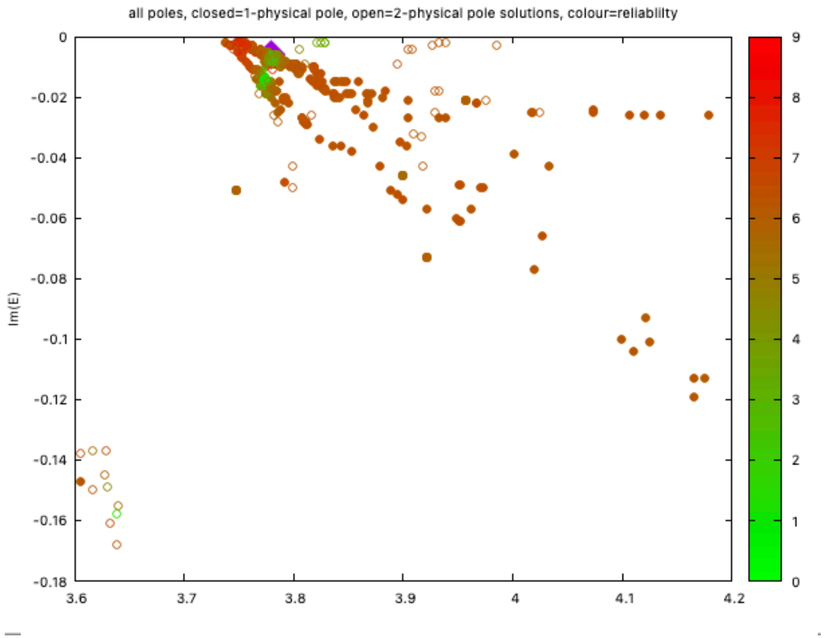
Pole Positions

Model space excludes the generating model.

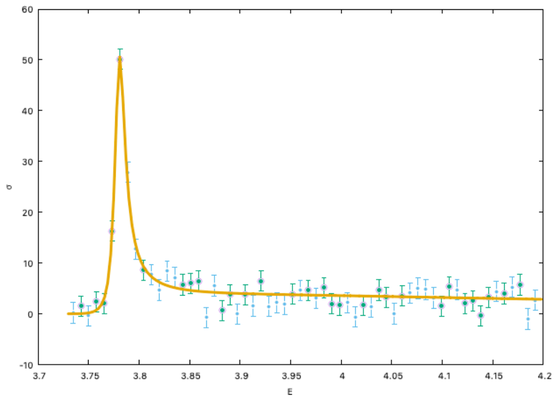
colour = "reliability":= post(ante)*post(post) = exp(-1/2 LP(a)-1.2LP(p)). This is log(1/2LP(a)+1/2LP(p))



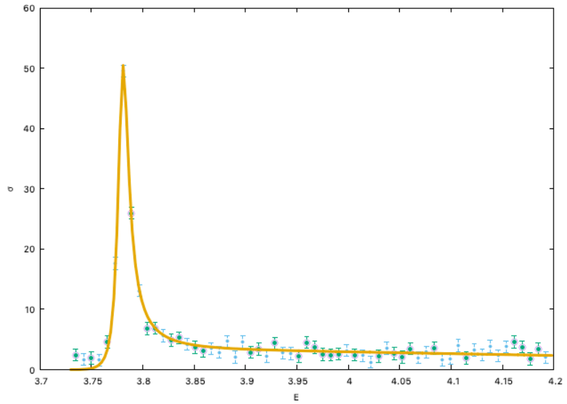




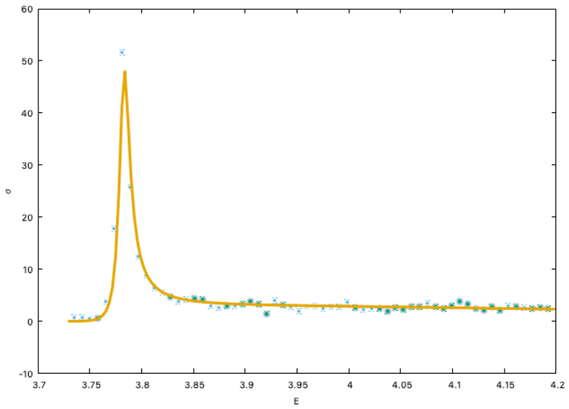
sigma=2.0



sigma=1.0



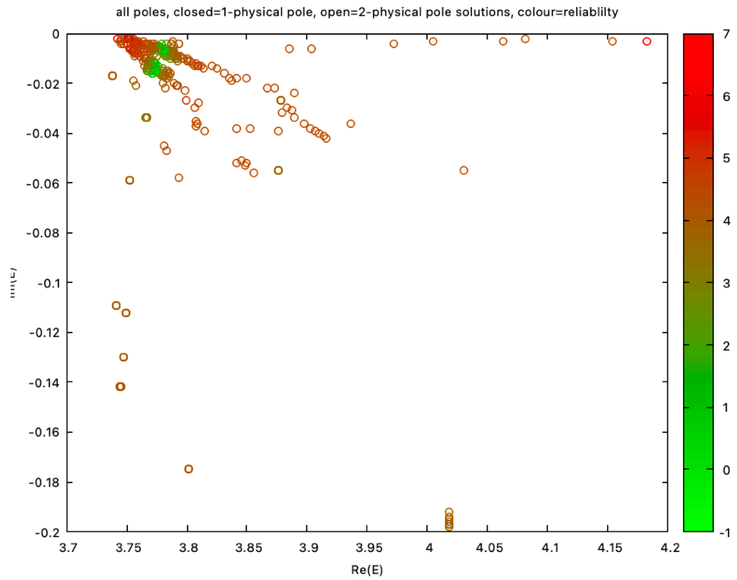
sigma=0.5

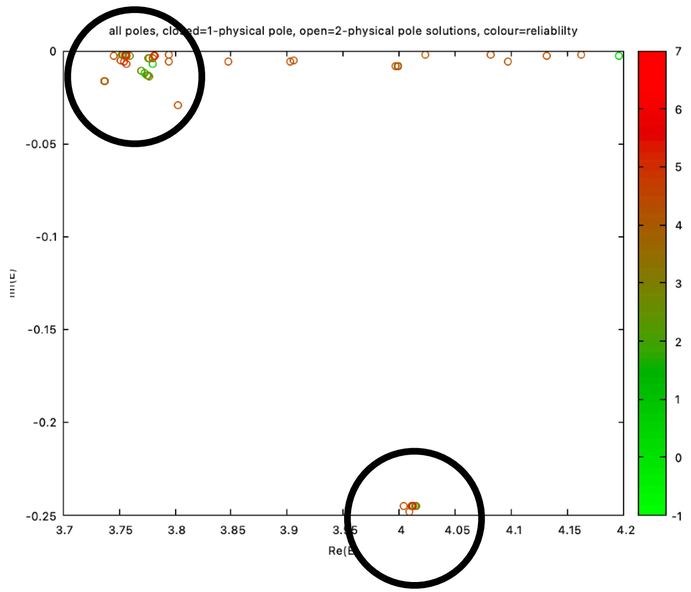


Pole Positions

We seek to say something about one-pole vs 2-pole fits. Here are the optimal reliabilties for each option

```
./MAK4
 number of available threads = 10
 enter verbose [0/1], make plots [0/1], exclude gen [0/1], nthreads, seed, max model size, number of data points, sig for data generation
0 0 1 8 717817 3 60 1.
 enter priorWidth, number of models to throw
1 8
 enter NMC, nMeas (total MCMC=nMeas*NMC)
4 10
 enter the MCMC step size for  masses, couplings, bg's
.05 .1 .1


  global average single pole  3.77694 +/- 0.00370055 (0.000168207)  +i -0.00953407 +/- 0.00374902 (0.00017041)

  one pole optimum reliability: 0.526393 @ (3.78,-0.007) residue: (-13.8443,6.07913)  model:  beta = 1 nPoles = 1 nBack = 1 Copt = 1 FFopt = 1

  two pole optimum reliability: 0.239203 @ (3.780,-0.007) residue: (-14.4516,5.85409) model:  beta = 4 nPoles = 2 nBack = 1 Copt = 1 FFopt = 1
  two pole optimum reliability: 0.239203 @ (4.196,-0.003) residue: (0.947862,2.21273) model:  beta = 4 nPoles = 2 nBack = 1 Copt = 1 FFopt = 1
```





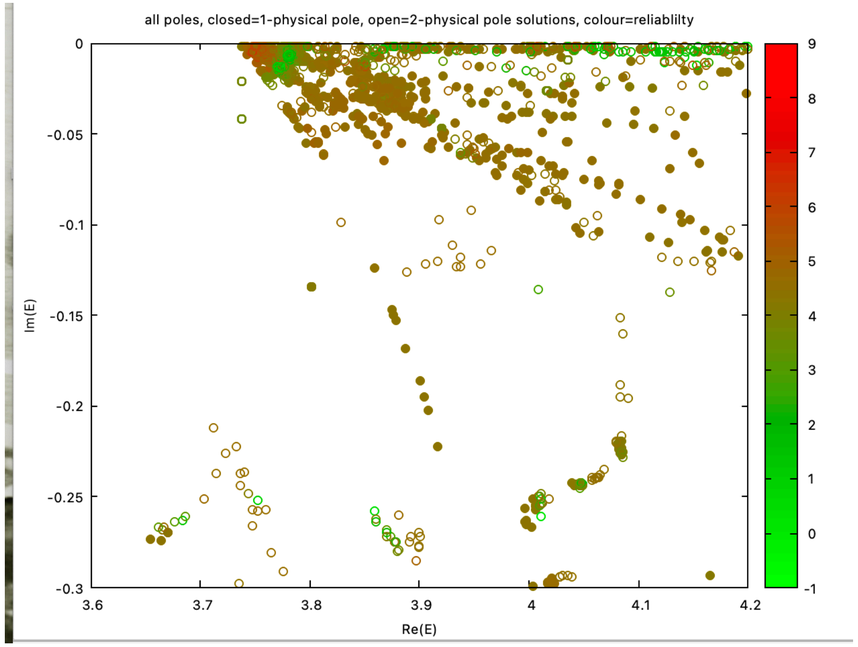
so we see that one pole of the doublet is approx the nominal and the other is off somewhere

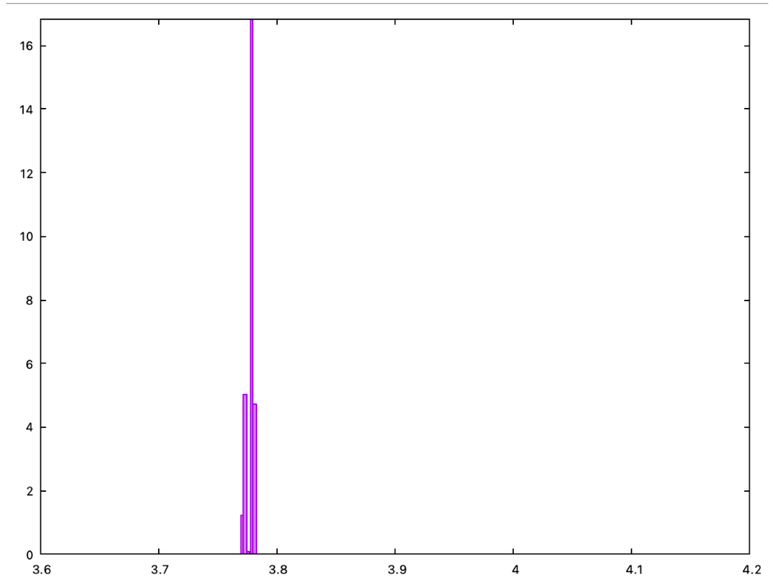
predictiveness/ModAvg/MAK4.cpp

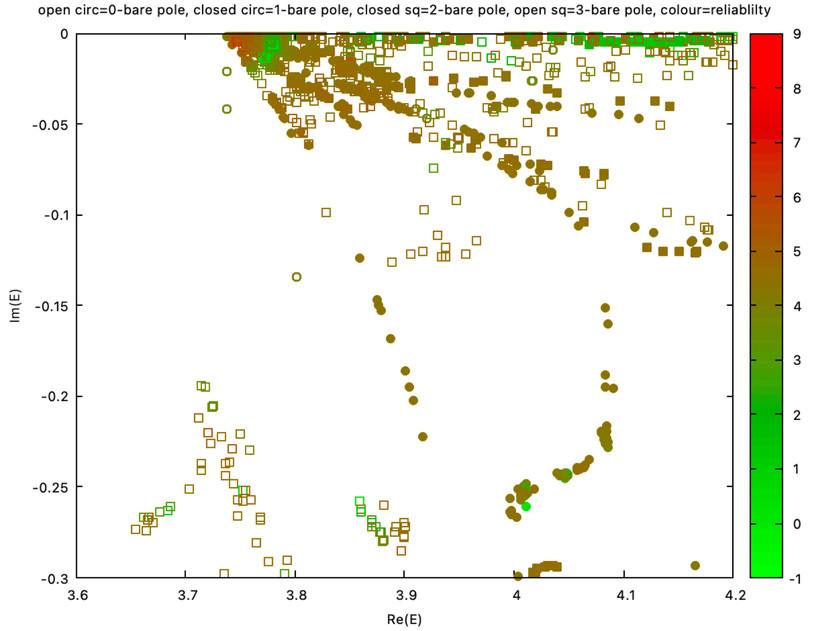
Pole Positions

We seek to say something about one-pole vs 2-pole fits. Here are the optimal reliabilties for each option

```
./MAK4
 number of available threads = 10
 enter verbose [0/1], make plots [0/1], exclude gen [0/1], nthreads, seed, max model size, number of data points, sig for data generation
0 0 1 8 891723 4 60 1.
 enter priorWidth, number of models to throw
1. 12
 enter NMC, nMeas (total MCMC=nMeas*NMC)
4 20
 enter the MCMC step size for  masses, couplings, bg's
.05 .1 .1


global average single pole  3.77841 +/- 0.00292129 (8.44008e-05)  +i -0.00823452 +/- 0.0031752 (9.17367e-05)

one pole optimum reliability: 0.483037 @ (3.78,-0.007) residue: (-13.9922,5.97456)  model:  beta = 2 nPoles = 3 nBack = 0 Copt = 1 FFopt = 1

two pole optimum reliability: 0.373839 @ (3.665,-0.268) residue: (9.54467,-4.69041) model:  beta = 1 nPoles = 3 nBack = 4 Copt = 2 FFopt = 2
two pole optimum reliability: 0.373839 @ (3.773,-0.014) residue: (-5.97938,8.18145) model:  beta = 1 nPoles = 3 nBack = 4 Copt = 2 FFopt = 2
```







big run Mmax=4 *predictiveness/ModAvg/MAK4.cpp*

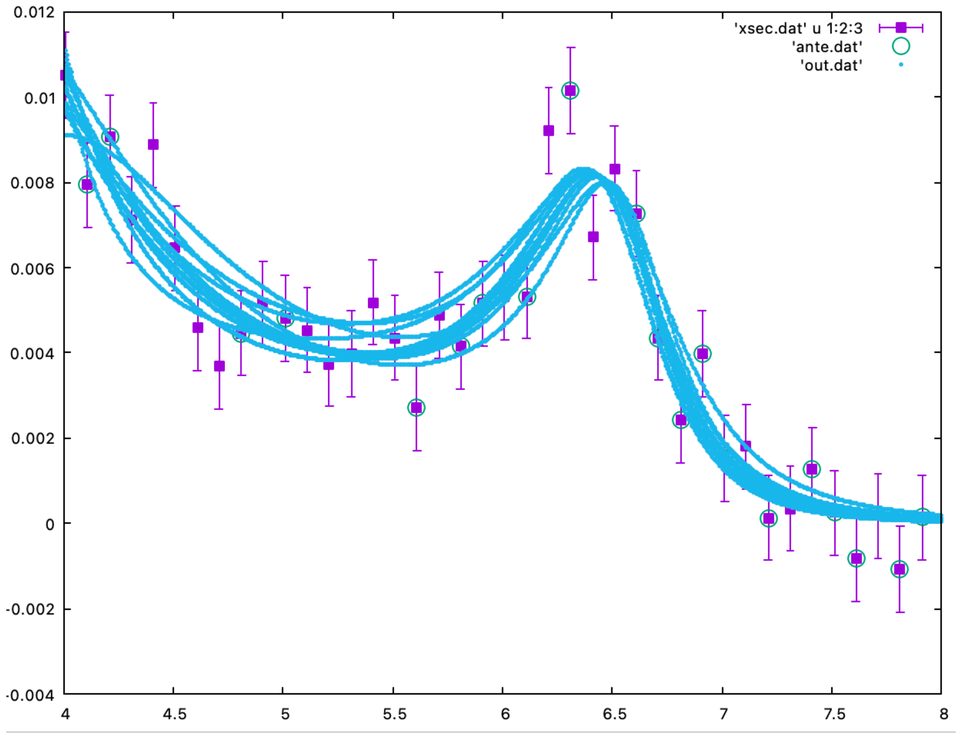
final thoughts

Future

application to unbinned data?

make models completely agnostic with NNs? Cf, use a VAE to model the posterior...

ensemble of 5D fits (10 data splits)



Conclusions

- focus on predictiveness
- de-emphasize fitting and fit quality
- explore a large model space to enhance the reliability of the conclusions (we need to admit model uncertainty!)
- use data realizations to enhance the reliability of the conclusions
- be as agnostic wrt priors and models as possible
- model parameters are not physical
- problems wrt model optimization and finding global minima are obviated/reduced.

✢ ÆRIC MEC HEHT GEWYRCAN