Methods for Bayesian model averaging

Ethan T. Neil (Colorado)
Lattice 2023 @ Fermilab
08/04/23

(image source: me, Fermilab buffalo farm, 2011)

Outline

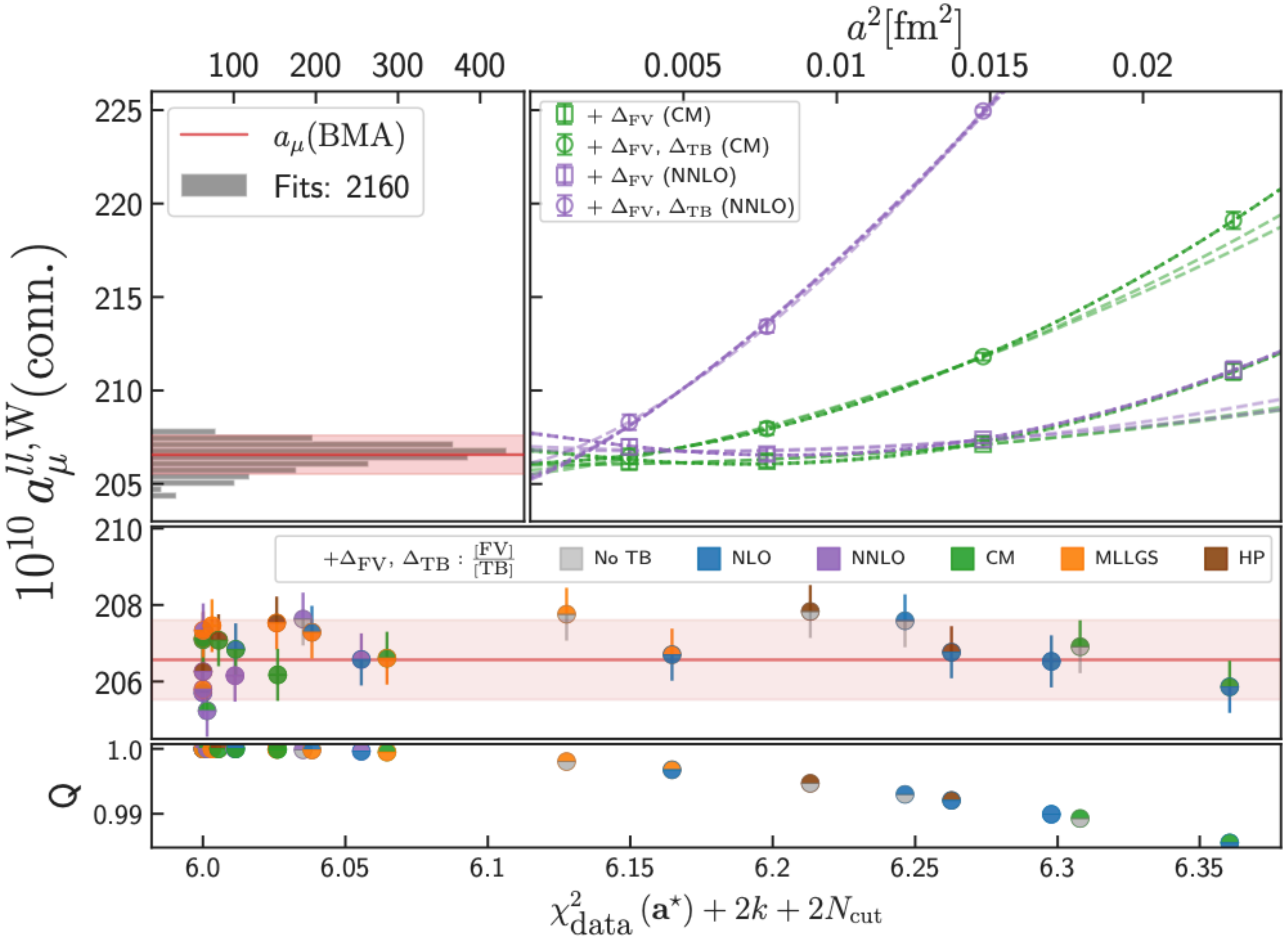
1. Why is model averaging useful?

2. Model averaging basics

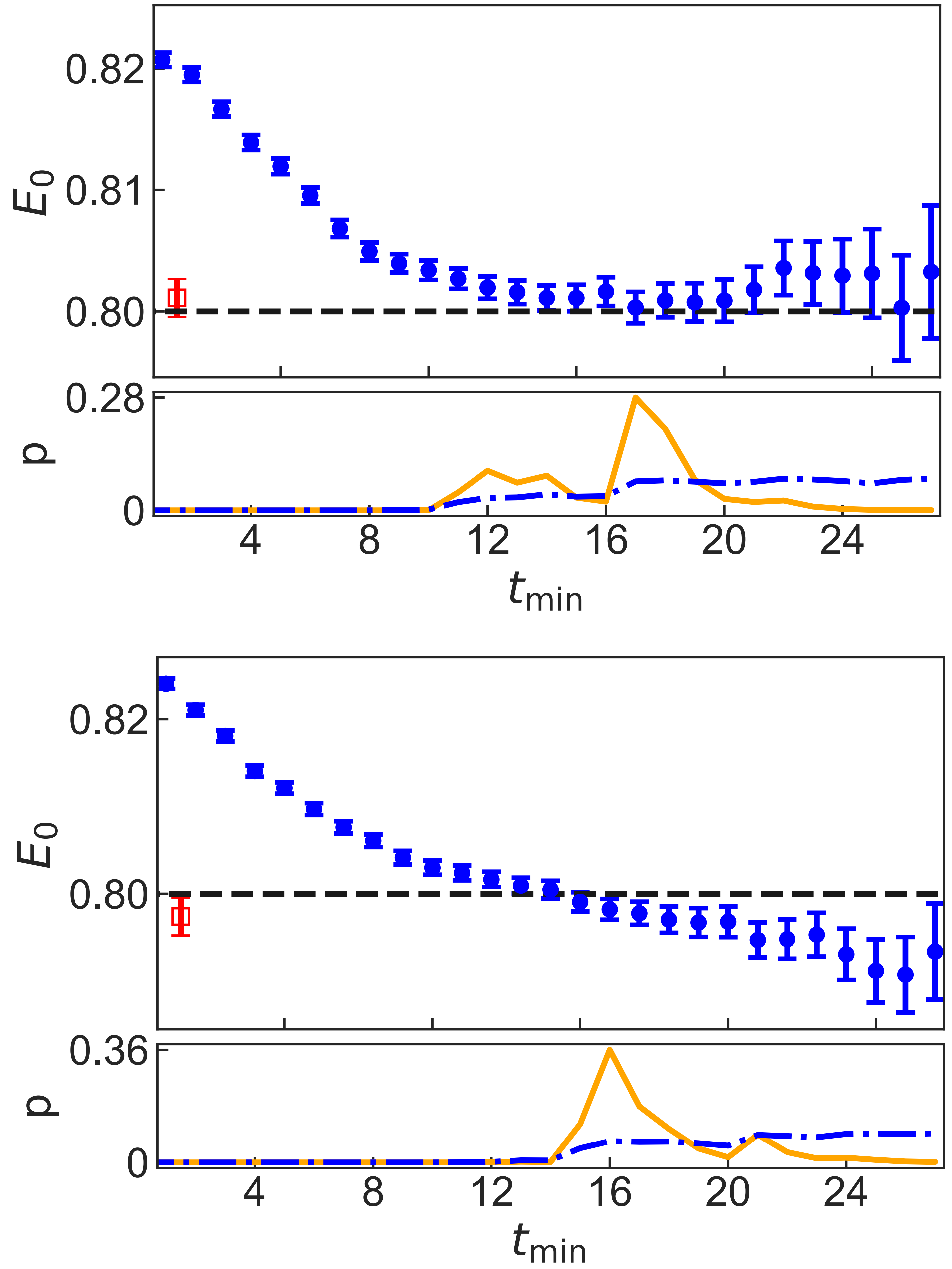
3. Improved information criteria (arXiv:**2208.14983**)

4. Data subset selection - what penalty? (arXiv:**2305.19417**)

(Fermilab/HPQCD/MILC collaborations, arXiv:**2301.0874**)



(EN and W. Jay, arXiv:**2008.01069**)



- **<u>Example 1:</u>** (g-2) HVP intermediate window (see talk by S. Lahert, Tue @ 2:10 PM)
- **2160** fit variations - discretization, finite volume, mass corrections…model average gives a final combined estimate + error bar.

- **<u>Example 2:</u>** synthetic data (fit 1 state to 2-state model truth.)
- Instead of selecting $t_{\rm min}$ by hand, compute model probability for each choice and average together! (Data cuts as model choice.)

- *Some history*: we didn't bring model averaging to lattice, we "added the B" (**Bayesian** MA), found new ICs, and tried to clarify statistical derivations/details.

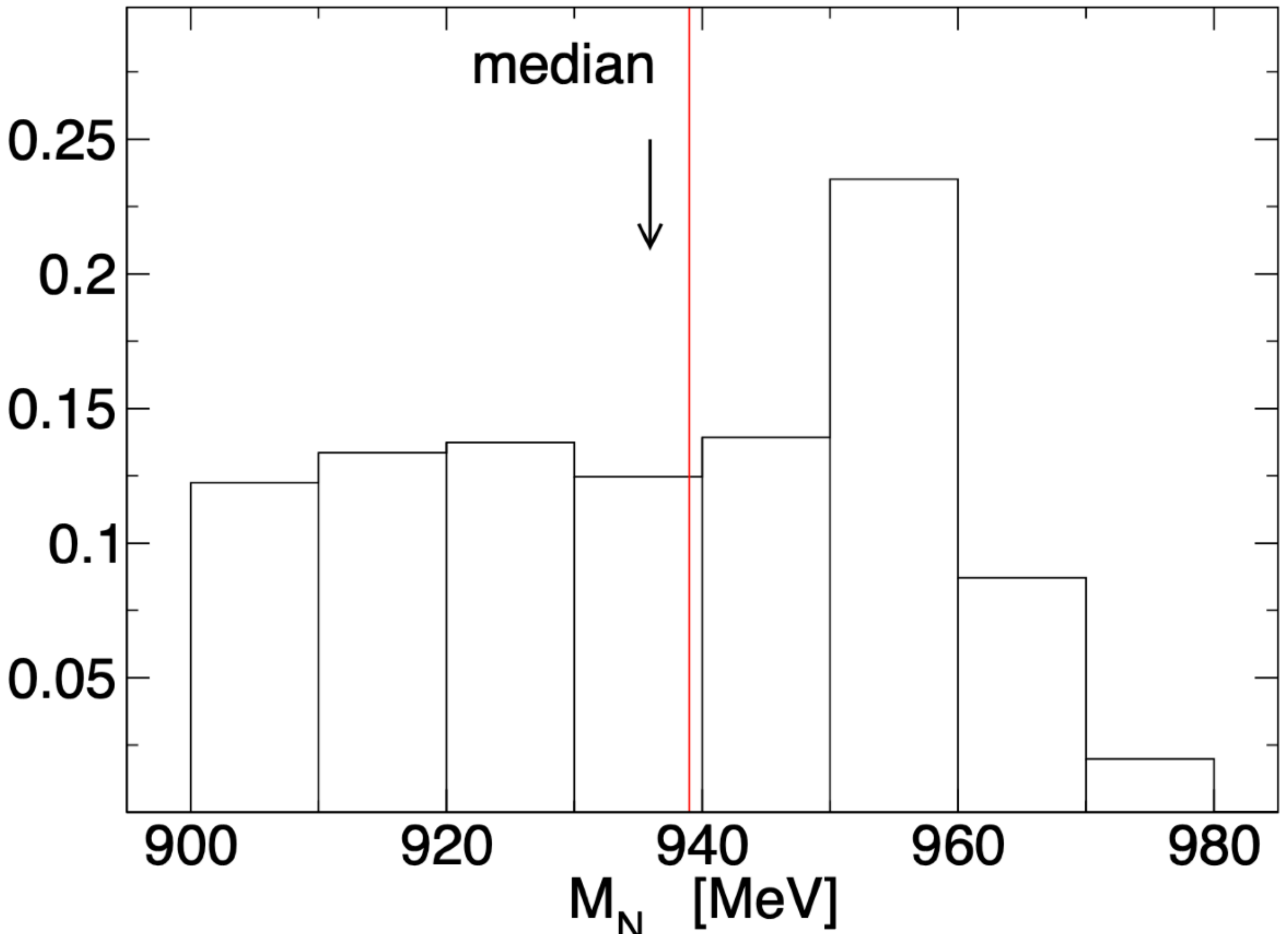
- Several early variations of model averaging/variation appear in lattice papers: Y. Chen et al. '04, **BMW '08**, **HPQCD '08**, **FNAL/MILC '14**, BMW '14…however, many old papers use *ad hoc* averaging prescriptions.

- First use of AIC for lattice is BMW '15; see also **CalLat '18**, '20, Rinaldi et al. '19. (More refs in our paper, including statistics papers back to the '70s.)

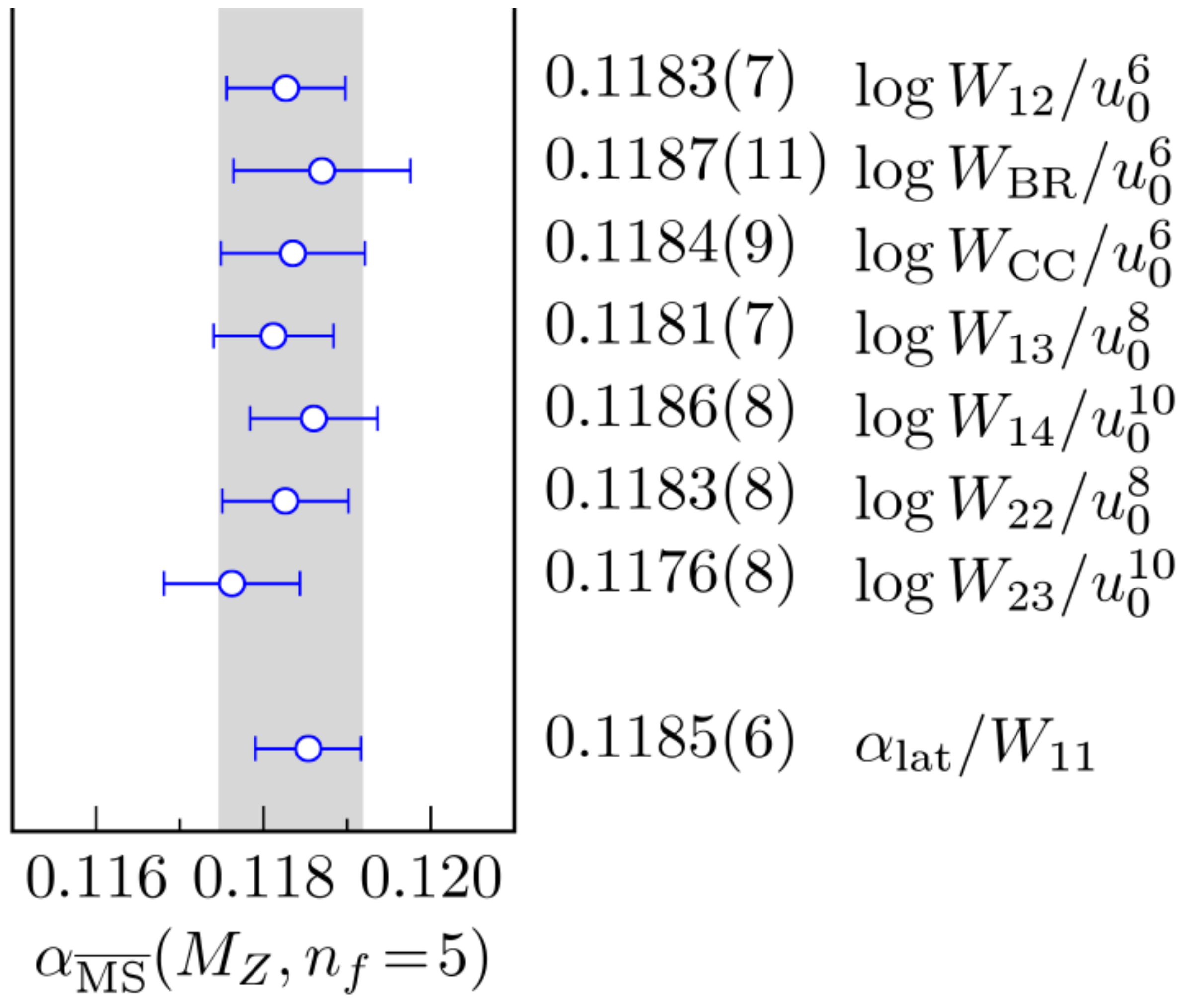
- First use of AIC with data penalty is BMW '21 (although I will argue for a *corrected* version of their formula here.)

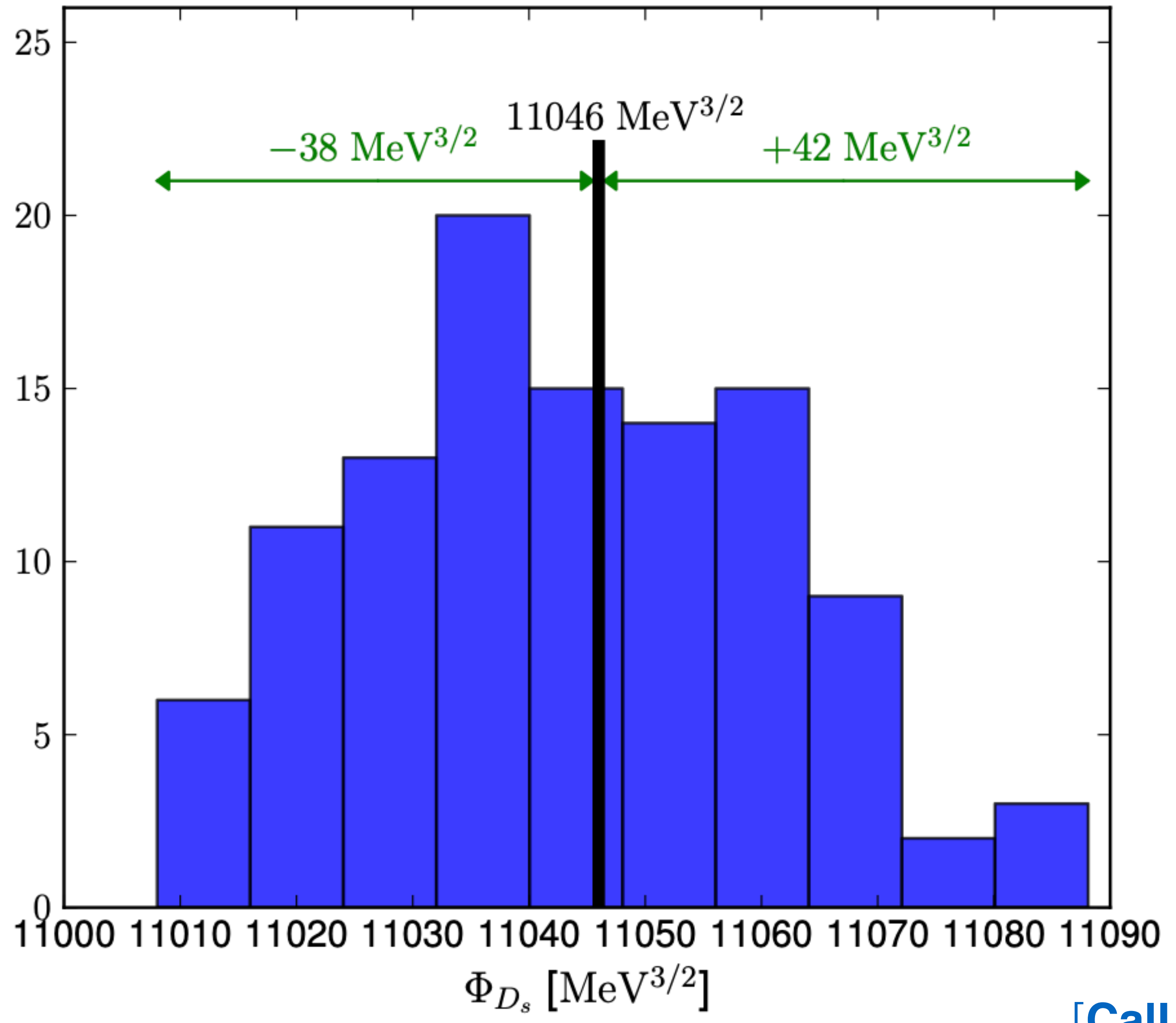
[Y. Chen et al '04]: arXiv:**hep-lat/0405001**
[BMW '14]: PRD 90 (2014), arXiv:**1310.3626**
[BMW '15]: Science 347 (2015), arXiv:**1406.4088**
[Rinaldi et al. '19]: PRD 99 (2019), arXiv:**1901.07519**
[CalLat '20]: PRD 102 (2020), arXiv:**2005.04795**
[BMW '21]: Nature 593 (2021), arXiv:**2002.12347**

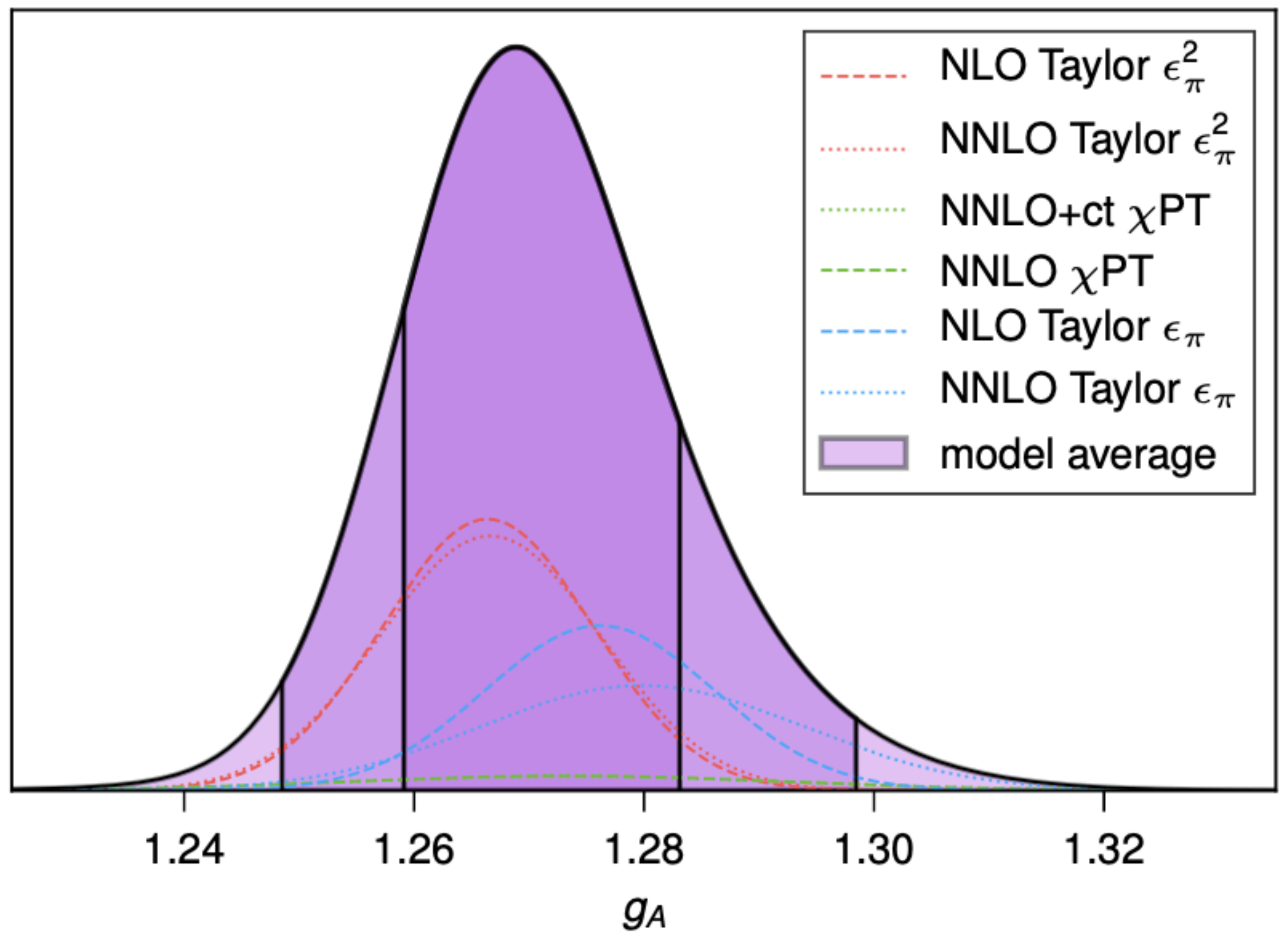
[**BMW '08**]: (BMW collaboration, *Science* 322 (2008), arXiv:**0906.3599**)



[**HPQCD '08**]: (HPQCD collaboration, *PRD* 78 (2008), arXiv:**0807.1687**)







[**CalLat '18**]: (CalLat collaboration, *Nature* 558 (2018), arXiv:**1805.12130**)

[**FNAL/MILC '14**]: (FNAL/MILC collaboration, *PRD* 90 (2014), arXiv:**1407.3772**)

(EN and W. Jay, arXiv:**2008.01069**)

(sketch adapted from S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling*, Springer Series in Statistics, 2008)

Bayesian model averaging: key ideas

- Bayesian model averaging: obtain any expectation value as a weighted average

$$\langle O \rangle = \sum_M \langle O \rangle_M \ \mathrm{pr}(M|D)$$

- Note that this applies at the level of *expectation values*. In particular, for mean and variance we find:

$$\langle f(\mathbf{a}) \rangle = \sum_\mu f(\mathbf{a}^*_\mu)\mathrm{pr}(M_\mu|\{y\}),$$

$$\sigma^2_{f(\mathbf{a})} = \langle f(\mathbf{a})^2 \rangle - \langle f(\mathbf{a}) \rangle^2$$

$$= \underbrace{\sum_\mu \sigma^2_{f(\mathbf{a}_\mu)}\mathrm{pr}(M_\mu|\{y\})}_{\text{average stat. error}} + \underbrace{\sum_\mu f(\mathbf{a}^*_\mu)^2\mathrm{pr}(M_\mu|\{y\}) - \left(\sum_\mu f(\mathbf{a}^*_\mu)\mathrm{pr}(M_\mu|\{y\})\right)^2}_{\text{model-variation systematic}},$$

- Asymptotically correct model weights are given by the (Bayesian) Akaike information criterion (AIC):

$$-2\log \mathrm{pr}(M|D) = -2\log \mathrm{pr}(M) + \mathrm{BAIC}$$

$$\mathrm{BAIC} = \hat{\chi}^2(\mathbf{a}^*) + 2k$$

pr(M) is model prior prob - unless you know what this is, take it to be uniform and ignore it.

- This is *not* the same as taking a weighted average of variances (first term), *or* taking the variance of the weighted f(a*).

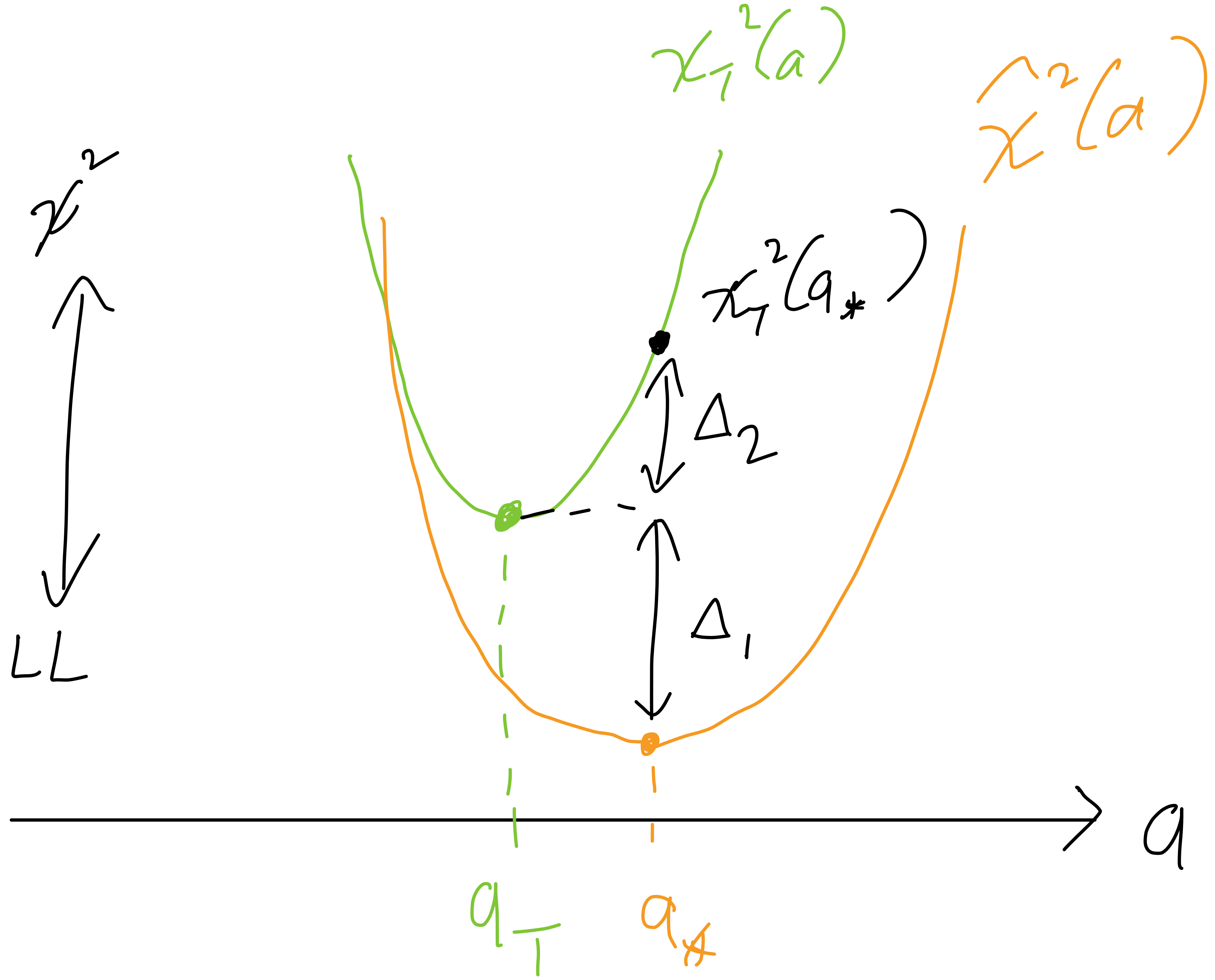
(sketch adapted from S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling*, Springer Series in Statistics, 2008)

Understanding the penalty term

- Looking at the (Bayesian) Akaike information criterion (AIC) again: (yes, this is the correct Bayesian formula, no explicit prior χ^2!)

$$-2 \log \mathrm{pr}(M|D) = -2 \log \mathrm{pr}(M) + \mathrm{BAIC}$$

$$\mathrm{BAIC} = \hat{\chi}^2(\mathbf{a}^*) + 2k$$



- "Occam's razor" penalty term +2k, where k = # of model parameters.

- Penalty *emerges naturally* from theoretical considerations as asymptotic bias correction.

- Briefly: sample **a*** is an unbiased estimator for true parameter a_T. But fluctuations of **a*** above and below $\mathbf{a}_T$ both overestimate likelihood (underestimate χ^2.) Correction of +2 (per dimension of **a**) —> **+2k.**

(EN and J. Sitison, arXiv:**2208.14983**)

Improved information criteria

(EN and J. Sitison, arXiv:**2208.14983**)

(S. Zhou, *Bayesian model selection in terms of Kullback-Leibler discrepancy,* PhD thesis, Columbia, 2011)

(S. Zhou, arXiv:**2009.09248**)

Using the Kullback-Leibler divergence

- **KL divergence** ("relative entropy") gives a path to Bayesian information criteria*. Basic definition:

$$\mathrm{KL}(M_\mu) = E_z[\log \mathrm{pr}_{M_\mathrm{T}}(z)] - E_z[\log \mathrm{pr}_{M_\mu}(z)]$$

  - Second term proportional to -log[pr(M|D)]. This is **non-parametric,** good - data should determine parameters. But there are multiple ways to obtain the above from a parametric model!

  - Three options are natural and give interesting ICs:

$$E_z[\log \mathrm{pr}_{M_\mu}(z)] \sim E_z[\log \mathrm{pr}_{M_\mu}(z|\mathbf{a}^*)]]$$

BAIC

("plug-in")

$$E_z[\log \mathrm{pr}_{M_\mu}(z)] \sim E_z[E_{\mathbf{a}|\{y\}}[\log \mathrm{pr}_{M_\mu}(z|\mathbf{a})]]$$

BPIC

("posterior average")

$$E_z[\log \mathrm{pr}_{M_\mu}(z)] \sim E_z[\log E_{\mathbf{a}|\{y\}}[\mathrm{pr}_{M_\mu}(z|\mathbf{a})]]$$

PPIC

("posterior predictive")

(sample size N -> ∞)

https://github.com/jwsitison/improved_model_avg_paper

Complete formulas

$$\mathrm{BAIC} = \overbrace{\hat{\chi}^2(\mathbf{a}^*)}^{\text{Goodness of Fit}} \overbrace{+2k}^{\text{Model Complexity}} \overbrace{+2d_C}^{\text{Data Truncation}}$$

$$\mathrm{BPIC} \approx \hat{\chi}^2(\mathbf{a}^*) + 3k + 3d_C \overbrace{- \frac{1}{2}\tilde{H}_{ba}(\Sigma^*)_{ab} + \frac{1}{2}\tilde{g}_d T_{cba}(\Sigma_2^*)_{abcd}}^{\text{Higher-Order GoF}}$$

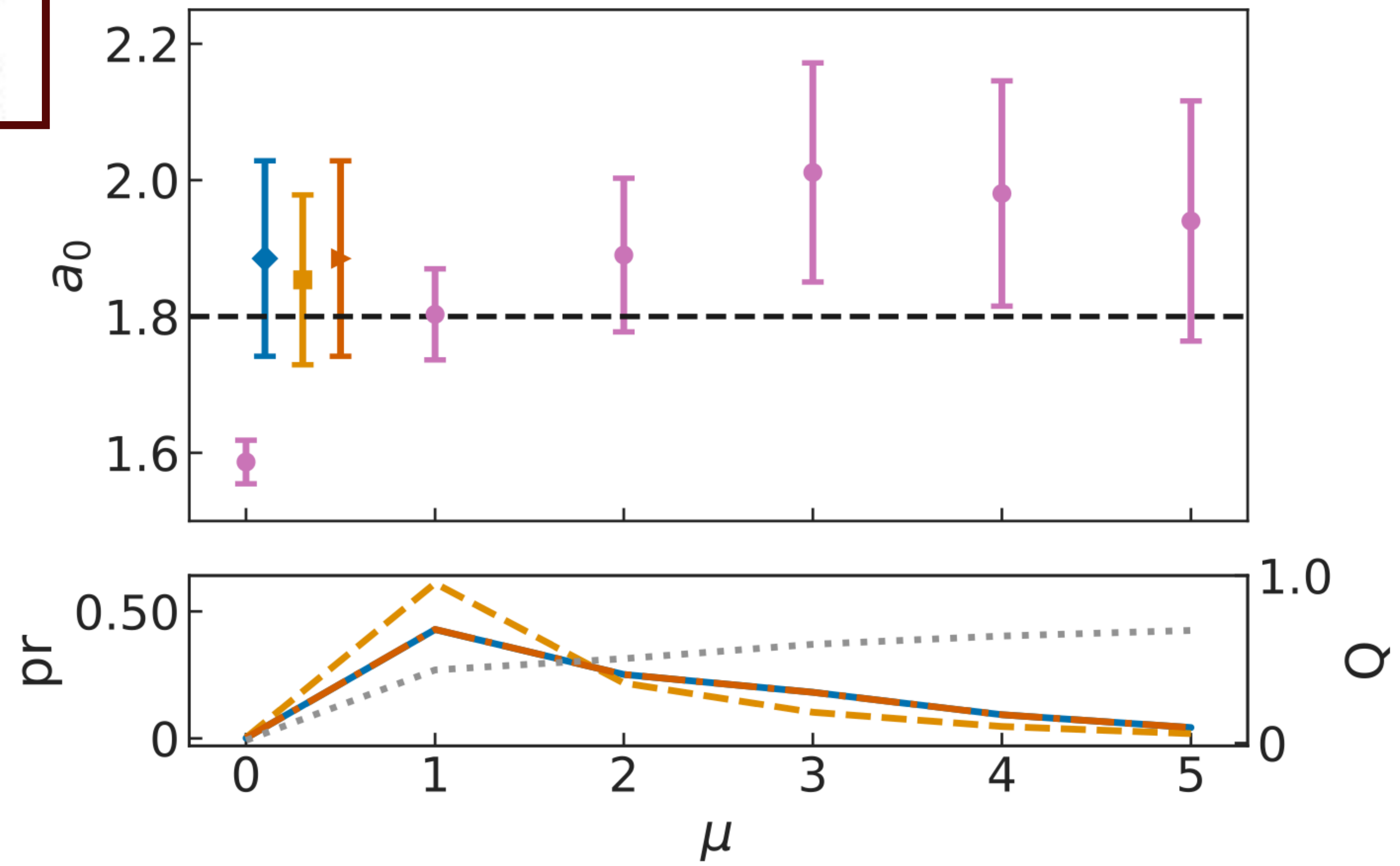
$$\mathrm{PPIC} \approx \hat{\chi}^2(\mathbf{a}^*) + 2k + d_C + N d_C \log\left(1 + \frac{1}{N}\right) - 2\sum_{i=1}^{N} \log\left[1 + \frac{1}{2}\left(\frac{1}{4}(g_i)_b(g_i)_a - \frac{1}{2}(H_i)_{ba}\right)(\Sigma^*)_{ab} + \frac{1}{4}(g_i)_d T_{cba}(\Sigma_2^*)_{abcd}\right]$$

- Various g, H, T, Σ are all *tensors of derivatives of chi-squared functions* - see our paper **2208.14983**, sec. IV. Numerical code available in Python + JAX (gradients/JIT compilation), although the code is *not polished* - just companion code for our paper.

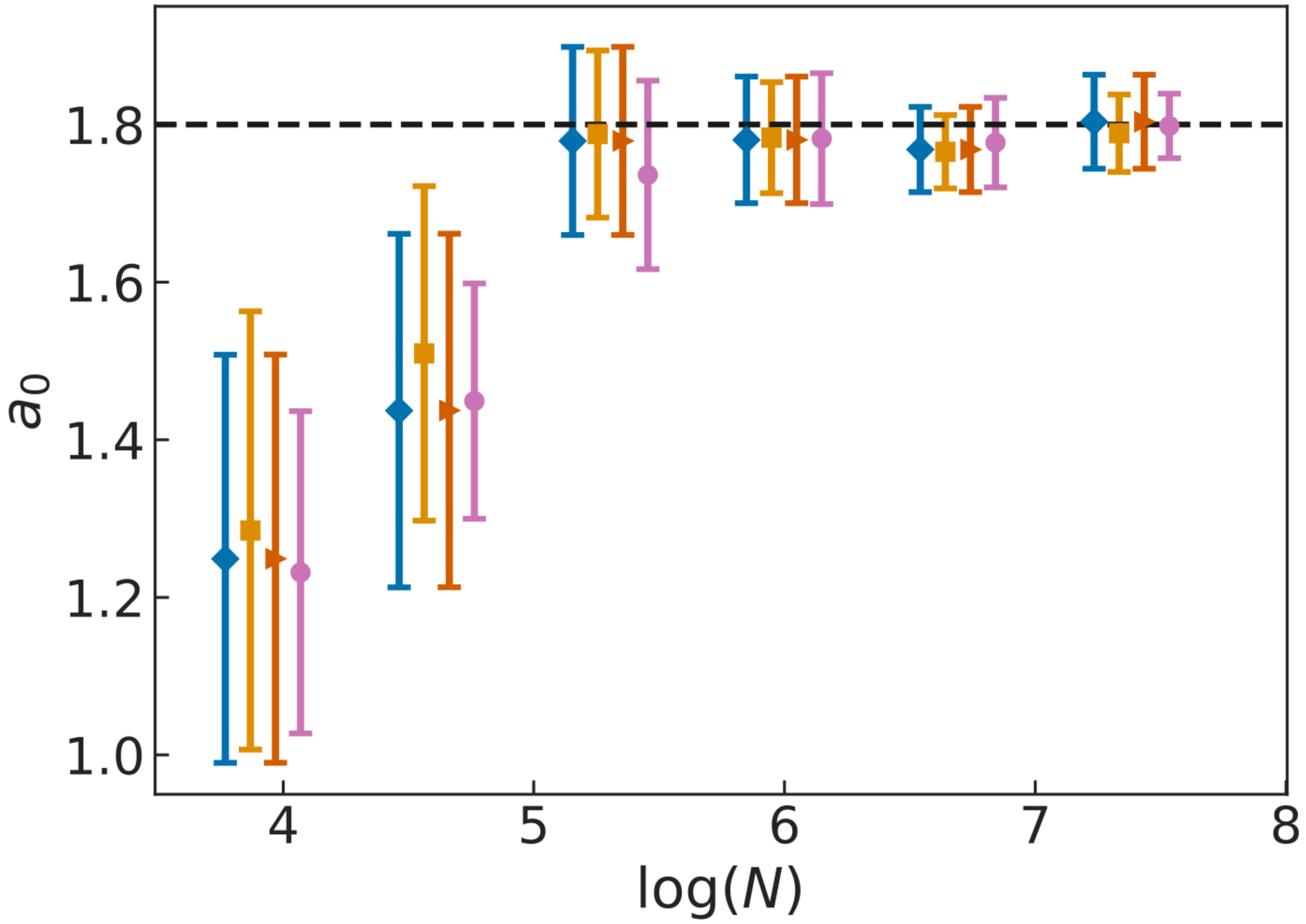
- The above formulas are *approximate*, NLO in large-N expansion (N = data sample size.) PPIC subset penalty is approximately $+2d_C$ plus 1/N corrections. BPIC has larger bias from posterior avg.

- We advocate use of **optimal truncation**, which replaces NLO —> LO when NLO terms are too large. (Fixes a potential numerical problem with log(…) in PPIC.)

Numerical results: fixed data

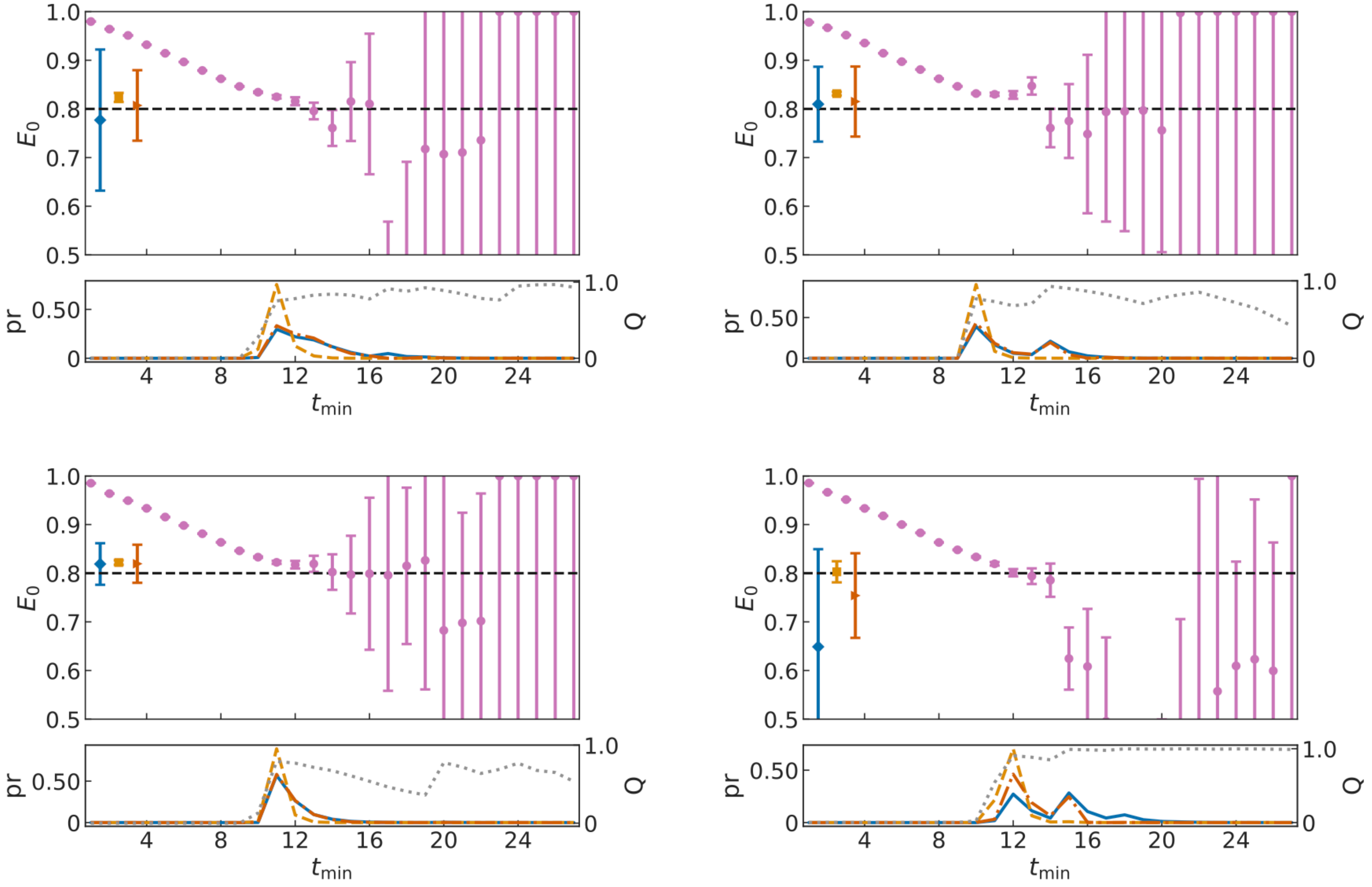


- Quadratic model truth, extract constant term a_0.
- **Left:** fits to polynomials of degree μ. Extra parameters are penalized, moreso for BPIC.



- **Right:** MA vs. sample size log(N). BPIC does slightly better in general, similar to fixed quadratic model.
- (This is sort of a special case since the “true model” is nested within the more complex μ>2 models…)

Numerical results: data selection



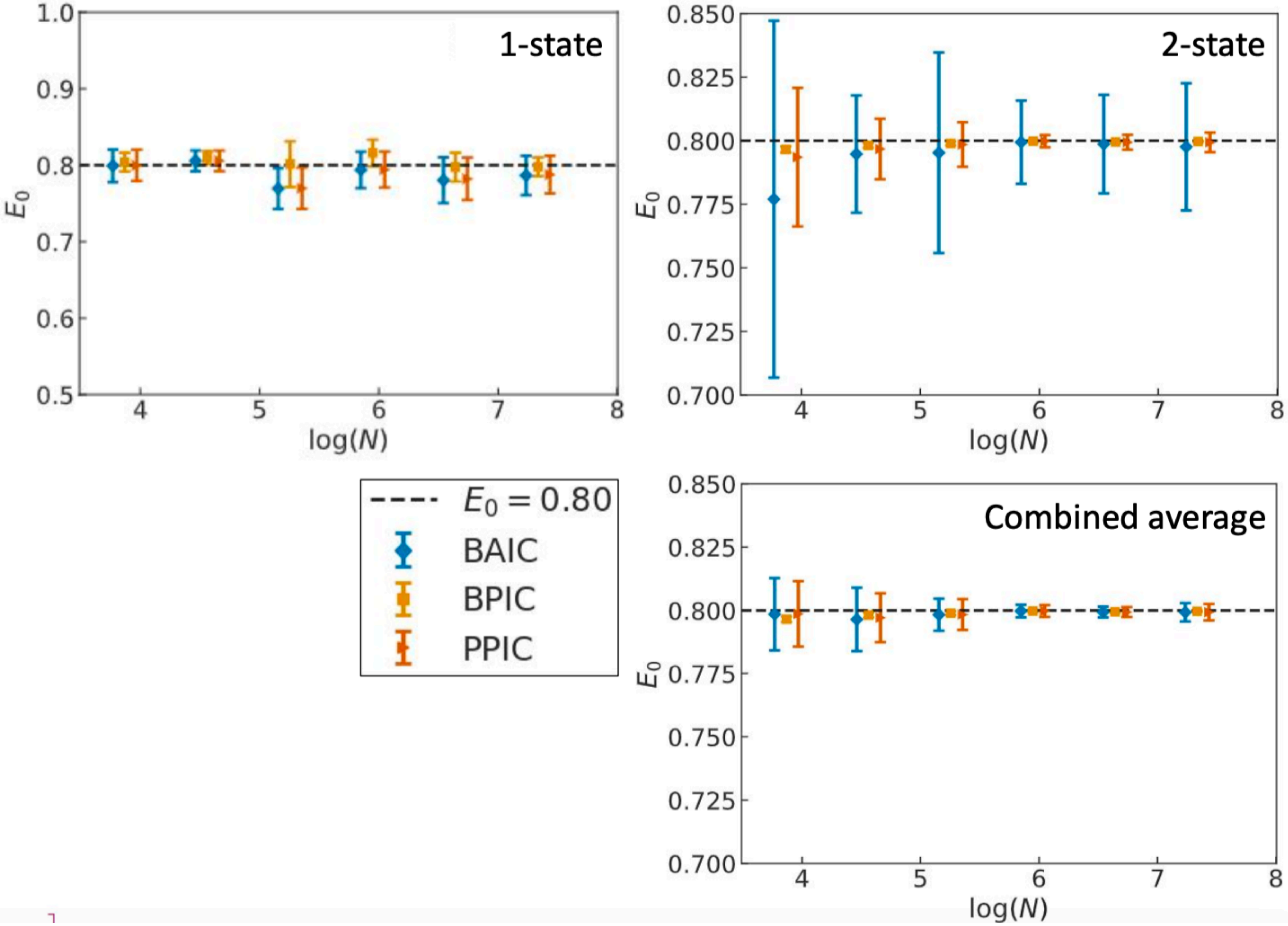
- BPIC cuts aggressively - often overly so (bias-variance tradeoff!) But it does fairly well when fitting the true model or with lots of data.

- PPIC is more robust against noise, otherwise performing similarly to BAIC (no excessive bias)

- BAIC is reliable and simplest to compute; we advocate PPIC generally, but nothing wrong with AIC!

Numerical results: data selection (2)

- Scaling results vs. N, similar conclusions to previous slide: we prefer PPIC, robust results and tends to give smaller error than BAIC, particularly w/noise
- BPIC has smallest error but can be too aggressive, particularly for subset selection.
- See paper for many more numerical results, including tests on real LQCD nucleon data (courtesy of JLab/W&M/MIT/LANL)



(EN and J. Sitison, arXiv:**2305.19417**)

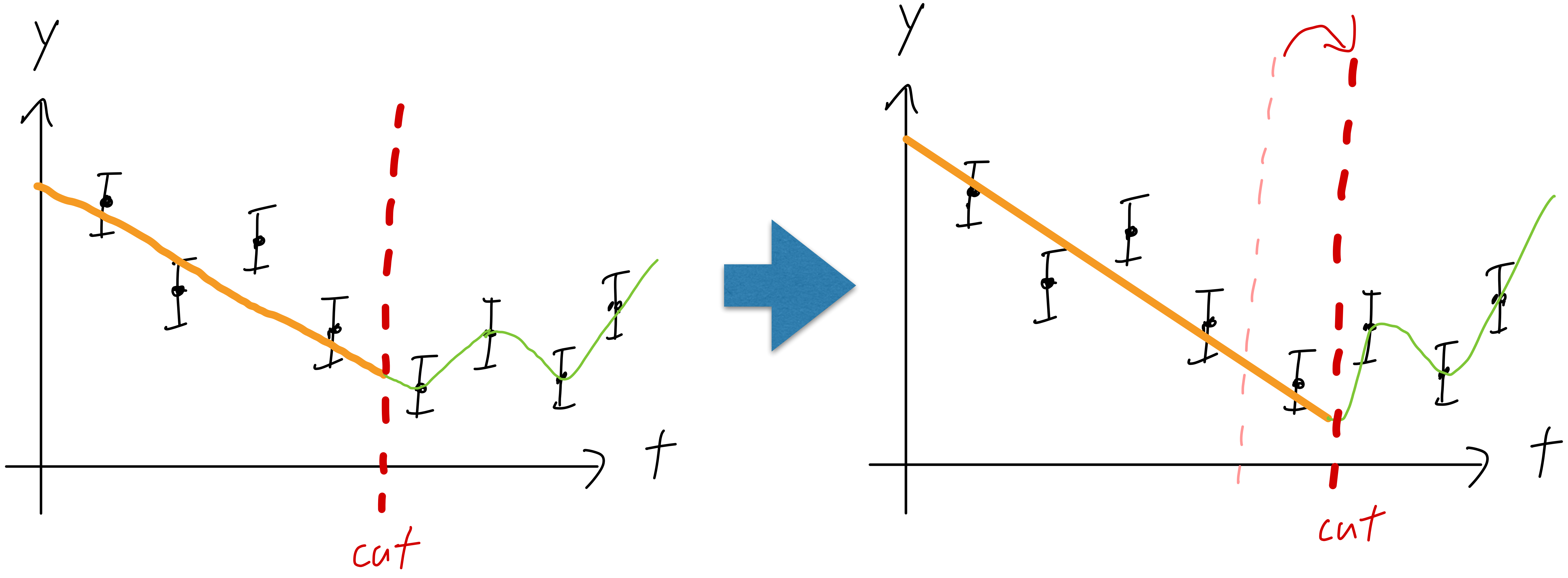
Data subset selection: which penalty?

(EN and J. Sitison, arXiv:**2305.19417**)

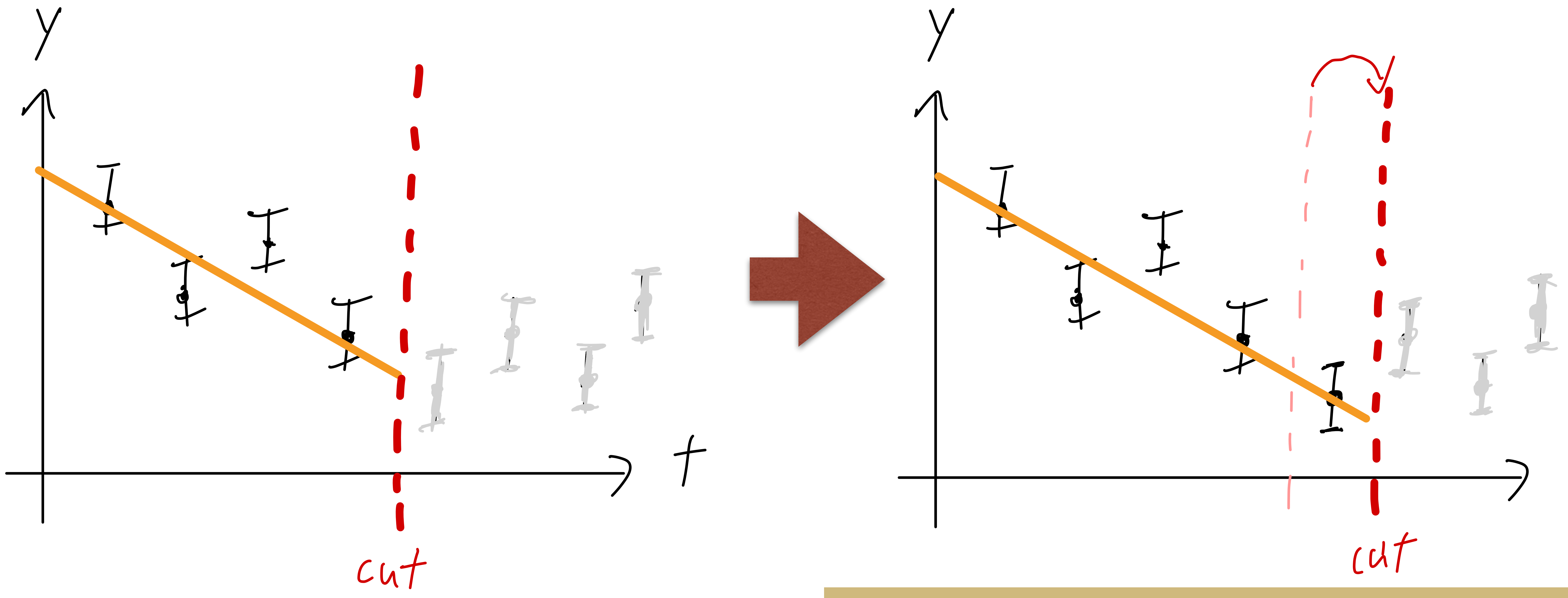
Two approaches to subset selection

(EN and W. Jay, arXiv:**2008.01069**)

- A common part of lattice analysis is data cutting: "what [t_{min}, t_{max}] should I fit my two-point correlator over?"
- Partition data into kept and cut [y_K, y_C] of size (d_K, d_C). Compute relative model weights, average!
- "Perfect model method": Keep all data. y_C fit to a model with χ^2=0; *bias correction* gives **+2d_C** penalty.
- "Subspace method": Discard data in cut partition. Recompute *total* KL divergence, gives **+d_C** penalty.



(BMW collab, *Nature* 593 (2021), arXiv:**2002.12347**)



Comparing the two methods

- We focus on AIC for simplicity (and since subspace proposal is only computed for AIC.)

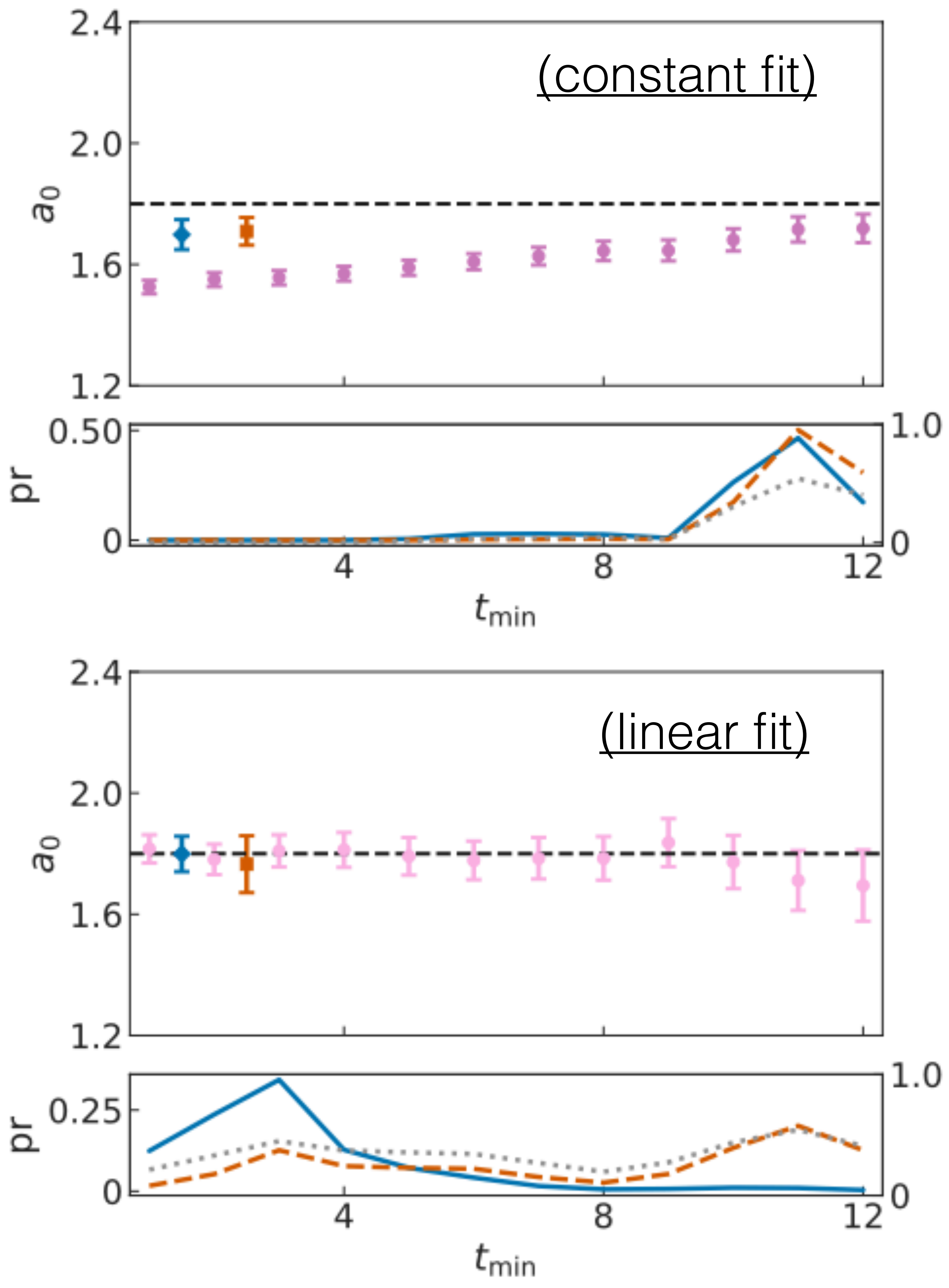
- We argue that AIC (subspace) is *subtly flawed*. By discarding data completely and re-computing the KL divergence, <u>information is thrown away</u>. This leads to <u>inflated errors</u> (with no corresponding bias reduction).

- Aside from the conceptual argument, we prove the identity:

$$\mathrm{KL}^{\mathrm{sub}}(M_\mu, d) \geq \mathrm{KL}^{\mathrm{sub}}(M_\mu, d_{\mathrm{K}})$$

- This behavior is (mildly) pathological - the subspace AIC will *never* choose to fit <u>all</u> of the data over <u>some</u> of the data (asymptotically.) Errors are increased as fits with data cuts are under-penalized.

(EN and J. Sitison, arXiv:**2305.19417**)



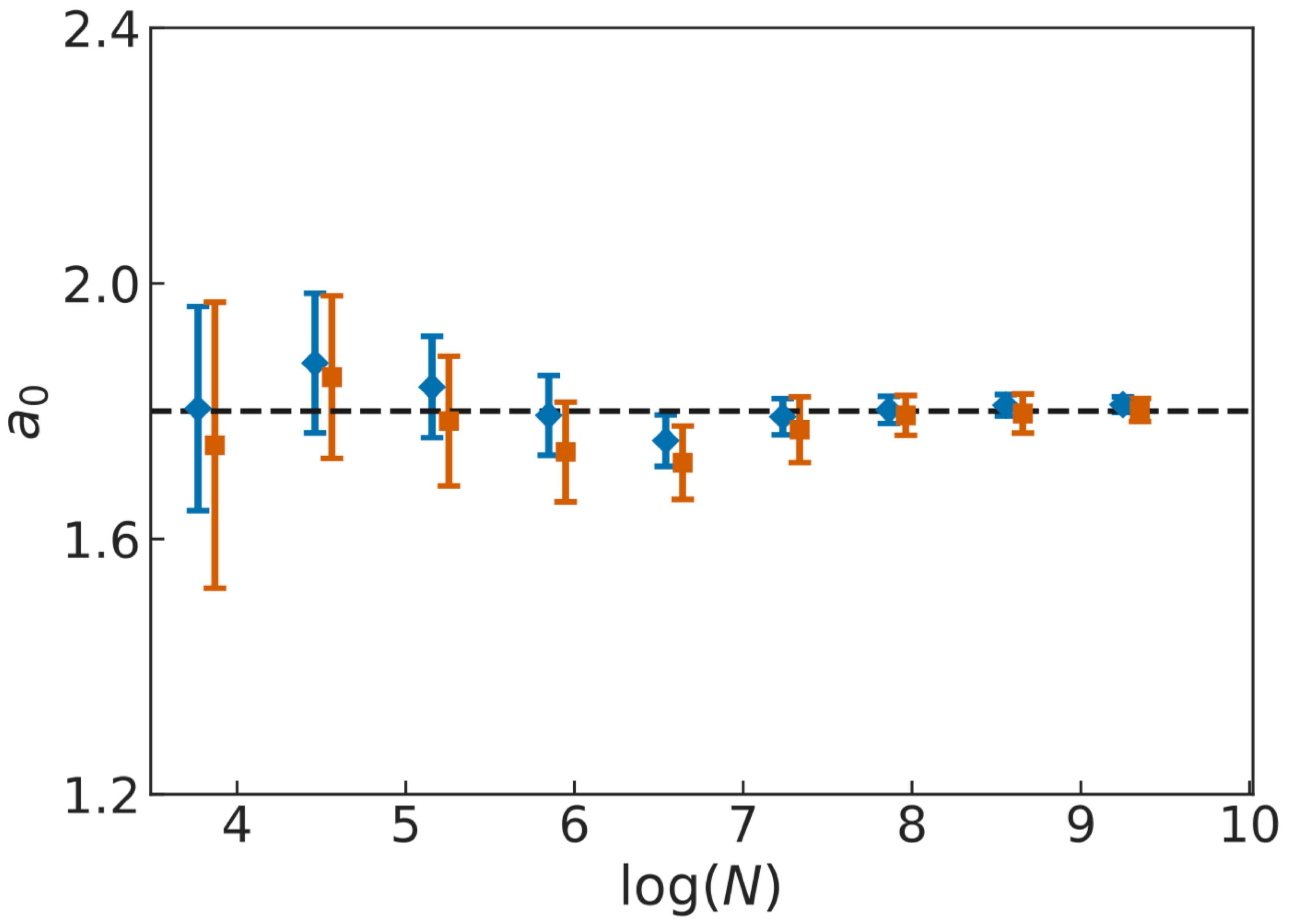
BAIC (perf)
BAIC (subset)

- Toy numerical example: model truth is linear,

$$f_{\mathrm{T}}(t) = 1.80 - 0.53\left(1 - \frac{t}{16}\right)$$

- For constant fit, both criteria are similar; <u>χ^2 is dominant.</u>
- For linear fit ("true model"), both averages are right, but <u>subset *under-penalizes* cutting</u> so has larger error.

- Below: "grand average" (both models @ all tmin) vs. sample size log(N).
- Both ICs agree well w/ model truth for all N; generically larger errors for BAIC (subset)



Summary

- Model averaging is a powerful and simple technique for dealing with analysis choices and associated systematic errors. Not a replacement for full Bayesian treatment (*see talk by J. Frison, this session!*), but easy to "plug in" to existing analysis chains.

- Bayesian + KL divergence perspective suggests two new ICs:

  - PPIC is more robust against noise and performs well in all tests.

  - BPIC uses Occam's Razor more aggressively, smaller error at the price of larger bias.

  - All ($N \to \infty$) roads lead to the (B)AIC, which is simple and effective.

- Data subset selection can also be done ("perfect model" construction.) Use the penalty of **$+2d_C$** for AIC, or analogous penalty formulas for other ICs.

Backup slides

Tips/tricks/FAQs

Q: <u>When should the “model prior” pr(M) be used?</u>

A: Use if you believe (*before* seeing data) that one model is more likely. (e.g., weight an EFT model over an ad hoc one.).

<u>**Do not**</u> use pr(M) to penalize complex models - AIC bias term already does this!

<u>**Do**</u> use pr(M) to deal with classes of similar models. E.g., if you are fitting 20 versions of chiral perturbation theory and one completely distinct model, you might set pr(M) = 1/40 for the variations, so that pr(chiPT) = pr(other model) = 1/2.

Q: <u>How do I use model averaging with strongly-correlated data?</u>

A: “100% correlation” is <u>**built-in**</u>: all pr(M|D) are computed with the <u>same, fixed data **D**</u>. No adjustment needed!

For data subset selection, correlations between cut and kept data can complicate life, particularly for BPIC or PPIC; see our paper **2208.14983** for methods.

Tips/tricks/FAQs

Q: How can I use model averaging with bootstrap/jackknife?

A: No modification needed! Bootstrap and jackknife just give better estimates of expectation values $\langle O \rangle_M$; same MA formulas apply, with same information criteria.

(Using bootstrap to compute ICs/bias directly is an interesting direction for future work!)

Q: Help, my BMA results look weird/I don't believe the MA errors!

A: Model averaging represents a bias-variance tradeoff; accounting for model choice uncertainty generally gives higher variance, but lower bias (your results are more likely to be right.) The discreteness of BMA can give strange-looking behavior, such as *increased* error when more data are added.

You should take this seriously, as long as you trust all of the inputs! ("Garbage in, garbage out…")

The Kullback-Leibler divergence

- KL divergence: "relative entropy" between PDFs, true model M_T vs. candidate model M_μ.

$$\mathrm{KL}(M_\mu) = E_z[\log \mathrm{pr}_{M_\mathrm{T}}(z)] - E_z[\log \mathrm{pr}_{M_\mu}(z)] \equiv \int dz \left[\mathrm{pr}_{M_\mathrm{T}}(z) \log \mathrm{pr}_{M_\mathrm{T}}(z) - \mathrm{pr}_{M_\mathrm{T}}(z) \log \mathrm{pr}_{M_\mu}(z)\right]$$

- KL = 0 if the PDFs are equal, positive definite otherwise. Find the "closest" distribution to pr_{MT} by **maximizing** the magnitude of the second term!

- Introduce model parameters **a**, and this leads to familiar results:

$$E_z[\log \mathrm{pr}(z|\mathbf{a}, M_\mu)] \simeq \frac{1}{N} \sum_i \log \mathrm{pr}(y_i|\mathbf{a}, M_\mu) = \underline{\frac{1}{N} \log \mathrm{pr}(\{y\}|\mathbf{a}, M_\mu)}$$

sample log-likelihood, i.e. $-\chi^2/2$

- e.g. finding best-fit point **a*** = minimization of KL divergence ("max likelihood".) Same likelihood function gives model probability weights, via Bayes theorem: pr(M|D) ~ pr(D|M).

(EN and J. Sitison, arXiv:**2305.19417**)

χ^2, dof, and subset selection

- Rewrite both forms of AIC in terms of usual number of degrees of freedom, $N_{dof}=d_K-k$:

$$
\begin{aligned}
\mathrm{AIC}^{\mathrm{sub}}_{\mu,d_{\mathrm{K}}} &= N_{\mathrm{dof}}\left(\hat{\chi}^2_{\mathrm{K}}(\mathbf{a}^*)/N_{\mathrm{dof}} - 1\right) + k, \\
\mathrm{AIC}^{\mathrm{perf}}_{\mu,d_{\mathrm{K}}} &= N_{\mathrm{dof}}\left(\hat{\chi}^2_{\mathrm{K}}(\mathbf{a}^*)/N_{\mathrm{dof}} - 2\right).
\end{aligned}
$$

- For a bad fit with large Ndof and $1 < \chi^2 < 2$, we can have $\mathrm{AIC}^{\mathrm{sub}} >> 0$ but $\mathrm{AIC}^{\mathrm{perf}} << 0$ (lower AIC is preferred.) Is this a problem?

- Example by explicit construction in appendix B of paper, but favoring a “bad fit” over a “good fit” in this way requires that a large amount of data are cut for the “good fit”. Rewrite $\mathrm{AIC}^{\mathrm{perf}}$ to see explicitly that the difference is still just data cutting penalty:

$$
\mathrm{AIC}^{\mathrm{perf}}_{\mu,d_{\mathrm{K}}} = N_{\mathrm{dof}}\left(\hat{\chi}^2_{\mathrm{K}}(\mathbf{a}^*)/N_{\mathrm{dof}} - 1\right) + k - d_{\mathrm{K}}.
$$

Asymptotic bias

- When constructing any statistical estimator, one typically worries about bias, defined as follows: for distribution $pr_T(z)$ with property ξ(z), given a finite sample {y} of size N and estimator X({y}),

$$b_z[X(\{y\})] \equiv E_z[X(\{y\}) - \xi(z)] = E_z[X(\{y\})] - \xi(z)$$

- In other words, when averaged over the true distribution (i.e. over many independent samples), a non-zero bias means the estimator is wrong. We can further define asymptotic bias as:

$$b_z[X(z)] = \lim_{N \to \infty} b_z[X(\{y\})]$$

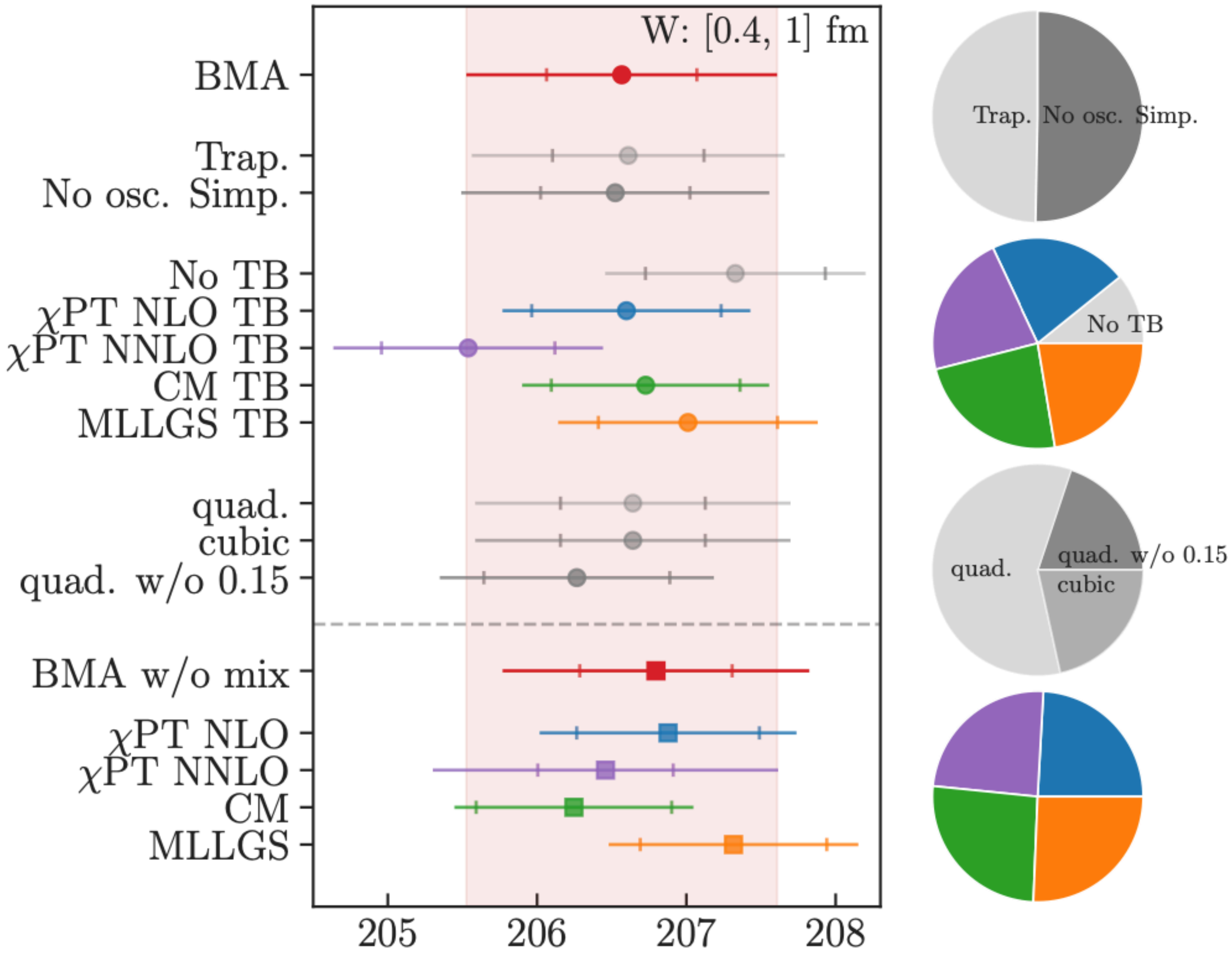
- Asymptotic bias is often easier to calculate than finite-sample bias, and estimators with zero asymptotic bias are at least self-correcting, in the sense that they are correct as N —> ∞.

- It is *not* obvious that an unbiased model probability gives an unbiased model average. But we prove the bias on the model average is bounded:

$$|b_z[\langle f(\mathbf{a})\rangle]| \leq \sum_\mu \left|\langle f(\mathbf{a})\rangle_\mu\right| |b_z[\mathrm{pr}(M_\mu|z)]|$$

assuming that the individual-model estimates <f(a)> are consistent (a slightly stronger version of asymptotically unbiased.) In short: **unbiased model weights give unbiased model averages.**

From (g-2) HVP model averaging analysis - can compare **subsets of model space** to understand systematic effects (center), or use model weights to compute **posterior probabilities** (pie charts)



(Fermilab/HPQCD/MILC collaborations, arXiv:**2301.0874;** talk by S. Lahert, Tue @ 2:10 PM)