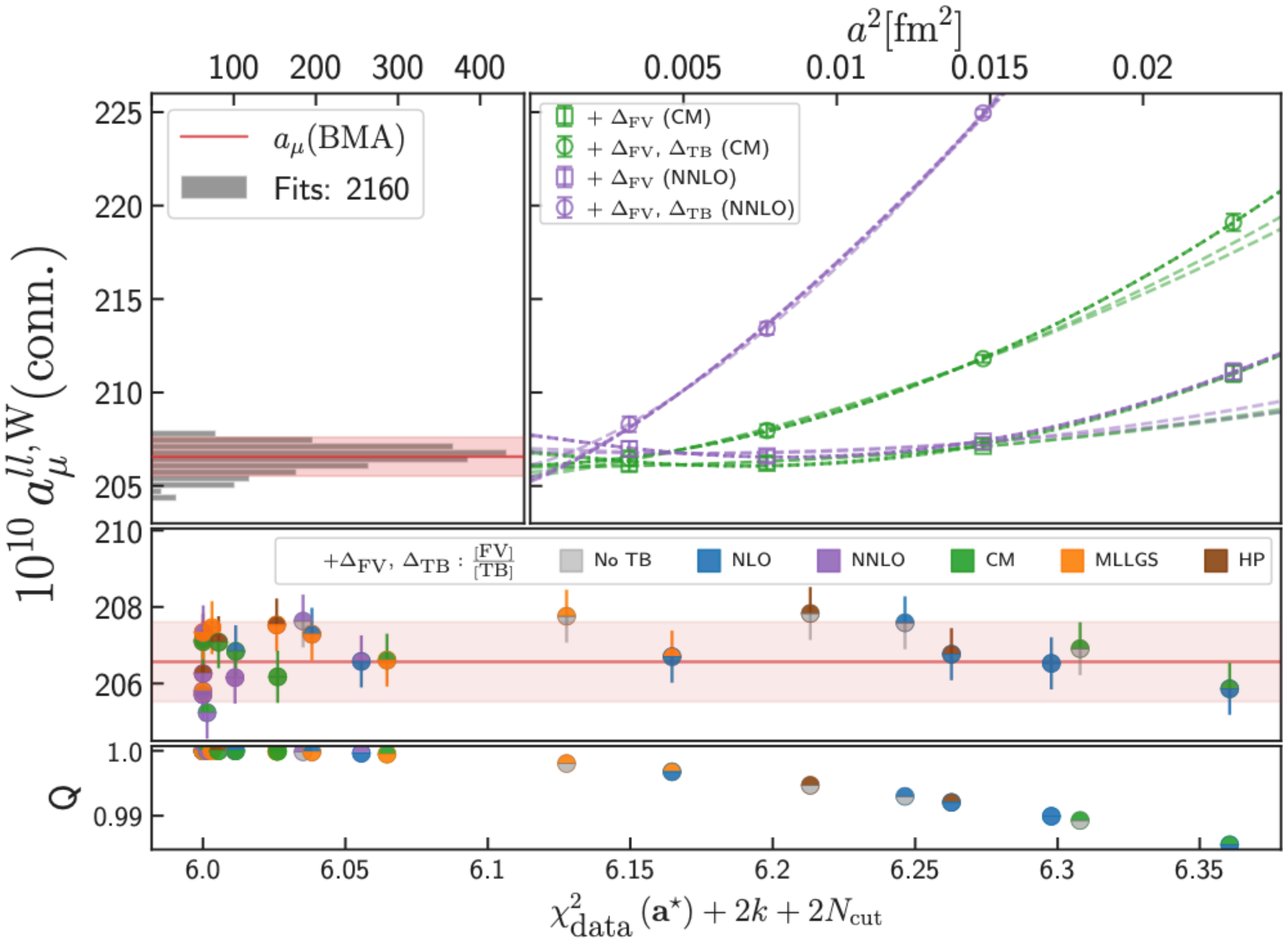
(Fermilab/HPQCD/MILC collaborations, arXiv:**2301.0874**)

arXiv:2008.01069 (with Will Jay)
arXiv:2208.14983 (with Jake Sitison)
arXiv:2305.19417 (with Jake Sitison)



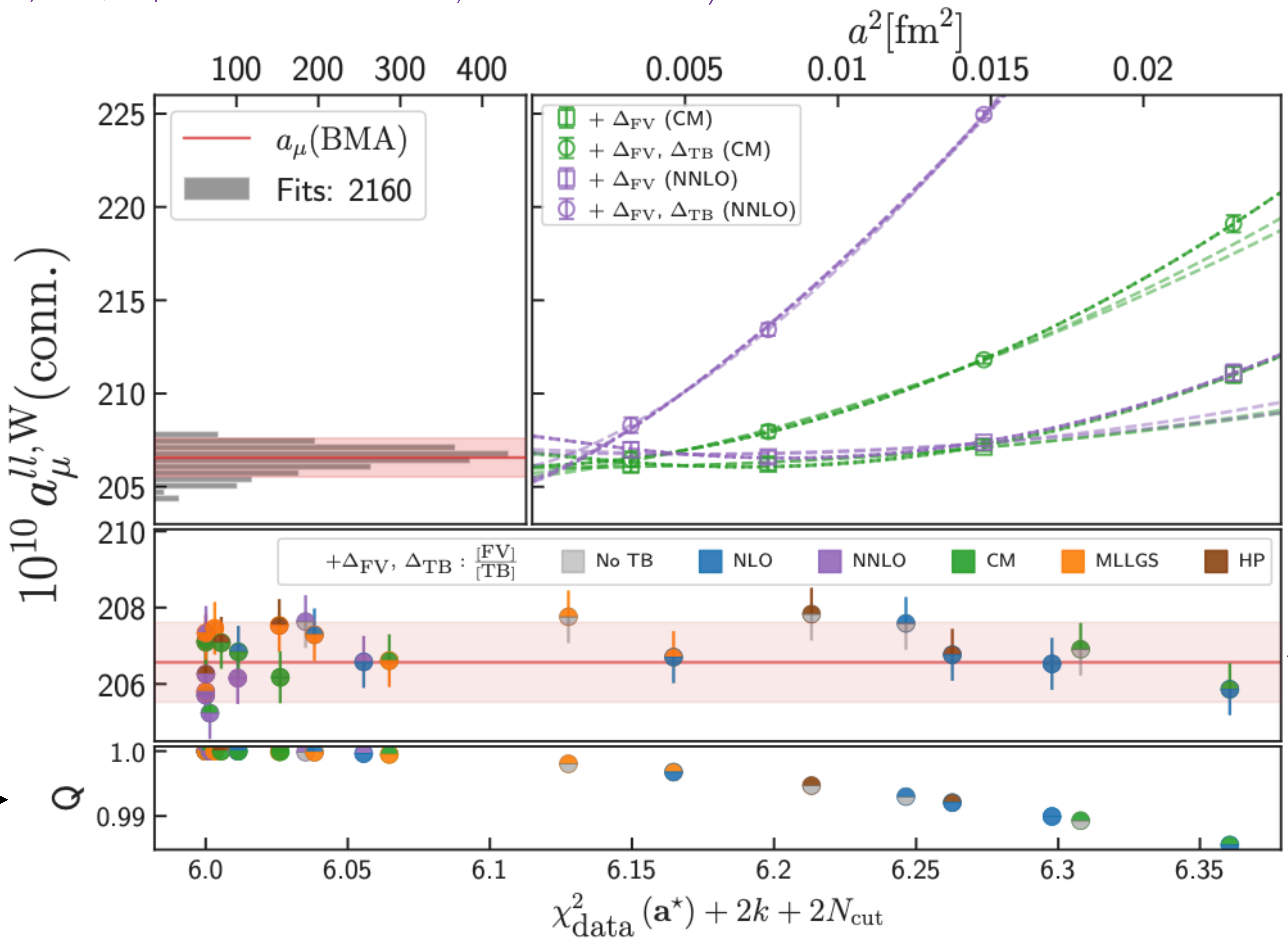
Bayesian model averaging: an overview

Ethan T. Neil (Colorado)
PDFLattice 2024 @ JLab
11/18/24

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

histogram of model results; weighted mean + err

fit Q-value (p-value, from chi-squared)



individual models vs. data

model results sorted by probability weight (bottom axis)

- **Model averaging:** account for systematic error due to model choices. Include all sensible model variations; compile results by model; average together, weighted by model probability.
- Above example has **2160** model variations - discretization, finite volume, mass corrections… model average gives a final combined estimate + error bar for continuum $a_\mu^{ll,W}$.

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

Bayesian model averaging: key ideas

- Bayesian model averaging: key formula is that any expectation value is a weighted average over model space {$\mathbf{M}_\mu$}, given data set **D**:

$$\langle O \rangle = \sum_\mu \langle O \rangle_\mu \ \mathrm{pr}(M_\mu | D)$$

- Usually, models are parametric: we have some parameter vector **a**, taken to be common to all models (model M_μ can have extra $\mathbf{a}_m$, marginalized over.) Expectation values are functions of parameters:

$$\langle f(\mathbf{a}) \rangle = \sum_\mu \langle f(\mathbf{a}) \rangle_\mu \mathrm{pr}(M_\mu | D)$$

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

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

$$= \sum_\mu \sigma^2_{f(\mathbf{a}_\mu)} \mathrm{pr}(M_\mu | \{y\}) + \sum_\mu f(\mathbf{a}^*_\mu)^2 \mathrm{pr}(M_\mu | \{y\}) - \left(\sum_\mu f(\mathbf{a}^*_\mu) \mathrm{pr}(M_\mu | \{y\}) \right)^2,$$

average stat. error   model-variation systematic

(EN and W. Jay, arXiv:**2008.01069**)
(EN and J. Sitison, arXiv:**2208.14983**)

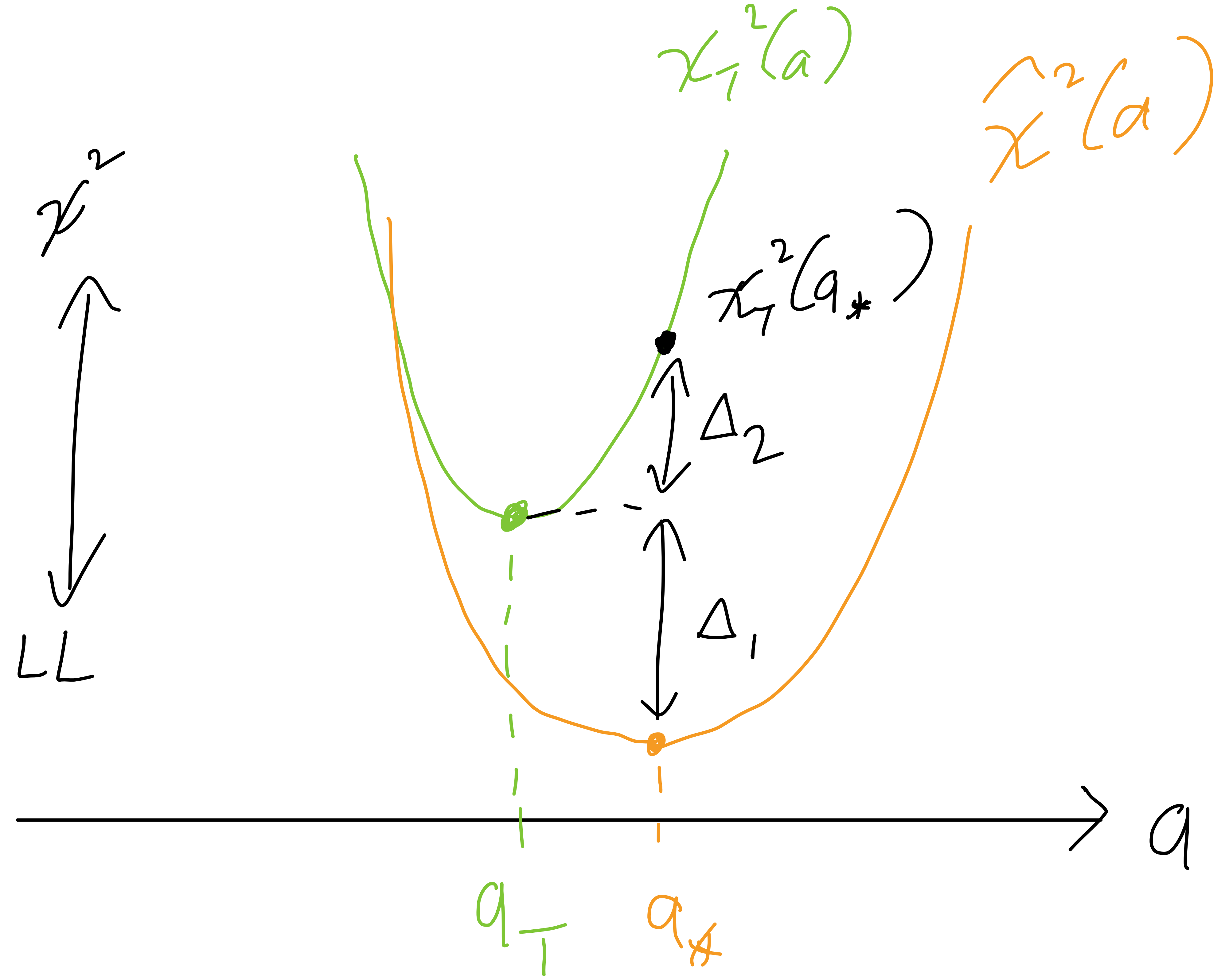
(sketch adapted from S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling*, Springer Series in Statistics, 2008)

Model probability weights

- Asymptotically correct model weights pr(M|D) from the (Bayesian) Akaike information criterion (AIC): (note, $\hat{\chi}^2$ is only data chi-squared, no explicit priors!)

$$-2 \log \mathrm{pr}(M_\mu | D) = -2 \log \mathrm{pr}(M_\mu) + \mathrm{BAIC}$$
$$\mathrm{BAIC} = \hat{\chi}^2(\mathbf{a}^*) + 2k$$

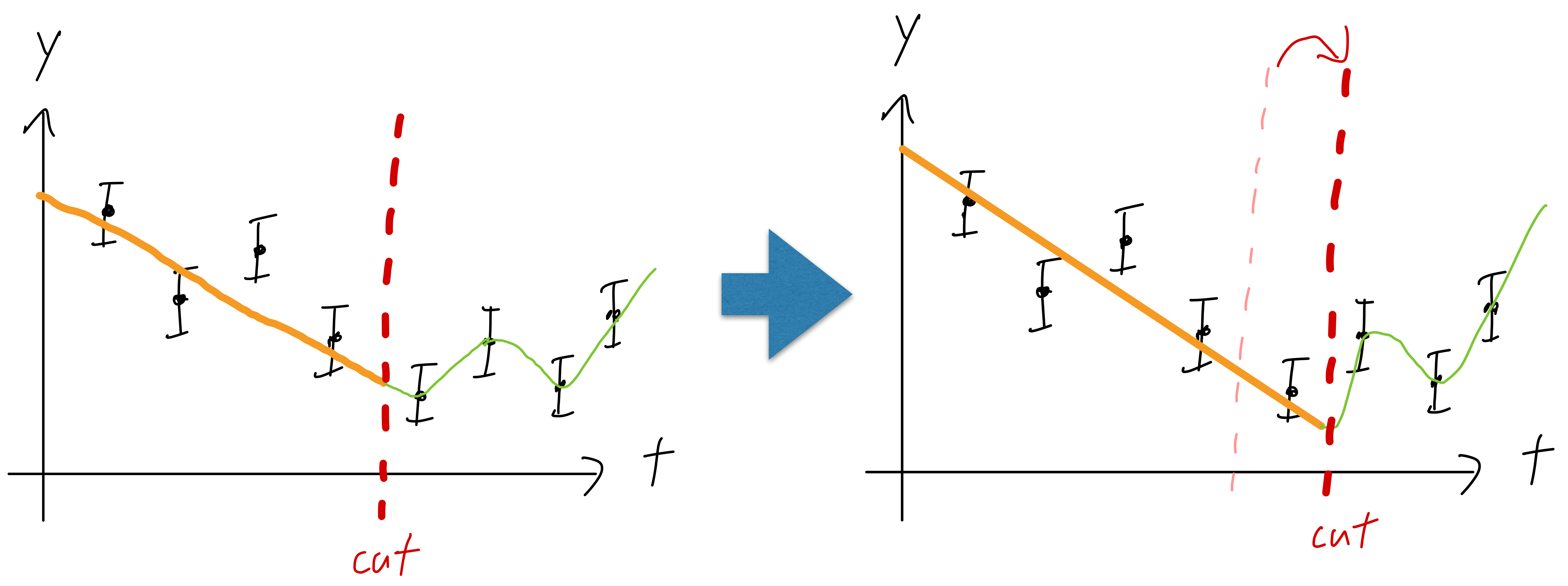
- pr(M) is *model prior probability*; if you don't know this, ignore it (take as flat prior pr(M) = 1/ N_M.)
- "Occam's razor" penalty term +2k appears, where k = # of model parameters.
- Penalty *emerges naturally* from theoretical considerations as asymptotic bias correction.



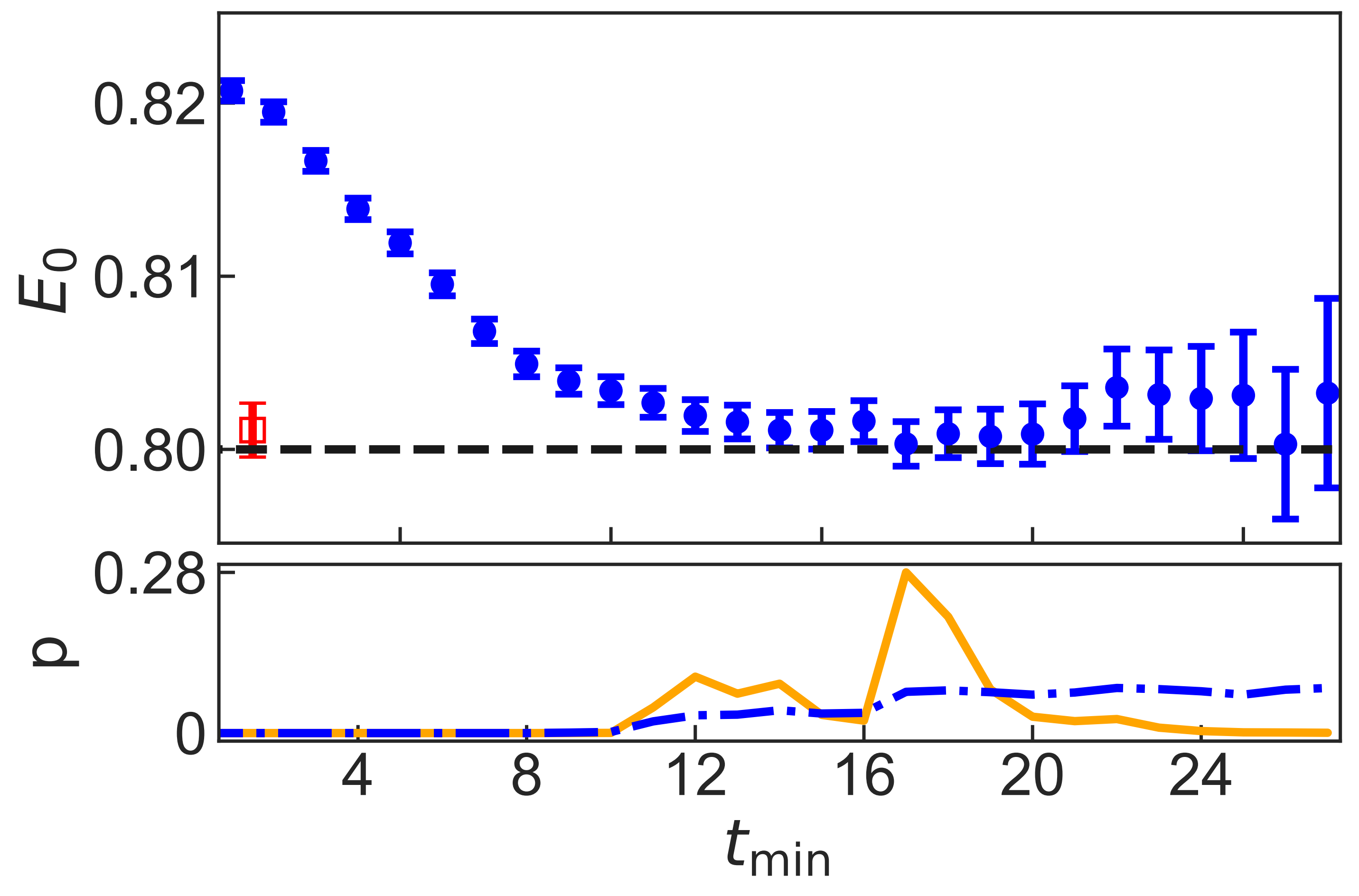
- Briefly: sample best-fit **a*** is an unbiased estimator for true parameter $\mathbf{a_T}$. But fluctuations of **a*** above and below $\mathbf{a_T}$ both overestimate likelihood (underestimate χ^2.) Correction of +2 (per dimension of **a**) —> **+2k.**

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

Data subset selection



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



- Model averaging can also be adapted to handle *data selection systematic effects* (i.e. "data cuts".)
- Imagine piecewise model, with removed data fit to "perfect model" (e.g. order d_C polynomial); contributes χ^2=0 exactly.
- But, bias correct: add subset selection penalty = 2*(# of data points removed).

Model averaging and functions

- This is a PDF workshop, so the expectation values of interest are *functions* and not just single values.

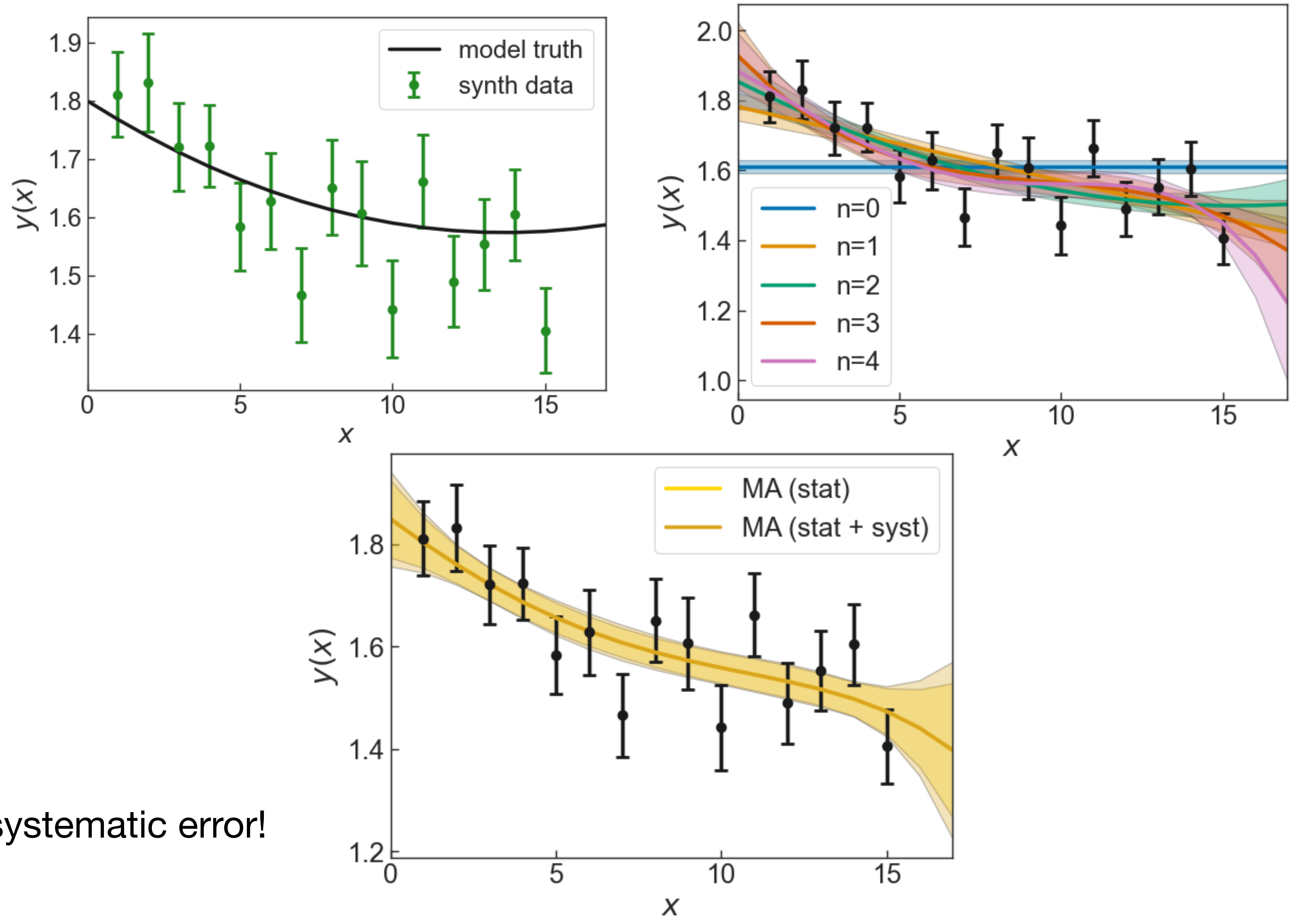
- Easy to extend the formalism to functions of independent variables:

$$f_{\text{avg}}(x) = \sum_{\mu} f_{\mu}(a_{\mu}^{\star}, x) \text{pr}(M_{\mu}|D)$$

$$\sigma_{\text{avg}}^2(x) = \sum_{\mu} \sigma_{\mu}^2(a_{\mu}^{\star}, x) \text{pr}(M_{\mu}|D)$$

$$+ \sum_{\mu} f_{\mu}(a_{\mu}^{\star}, x)^2 \text{pr}(M_{\mu}|D) - f_{\text{avg}}(x)^2$$

- (<u>Important</u>: don't omit model-space systematic error! Small here, but not always…)



(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:

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

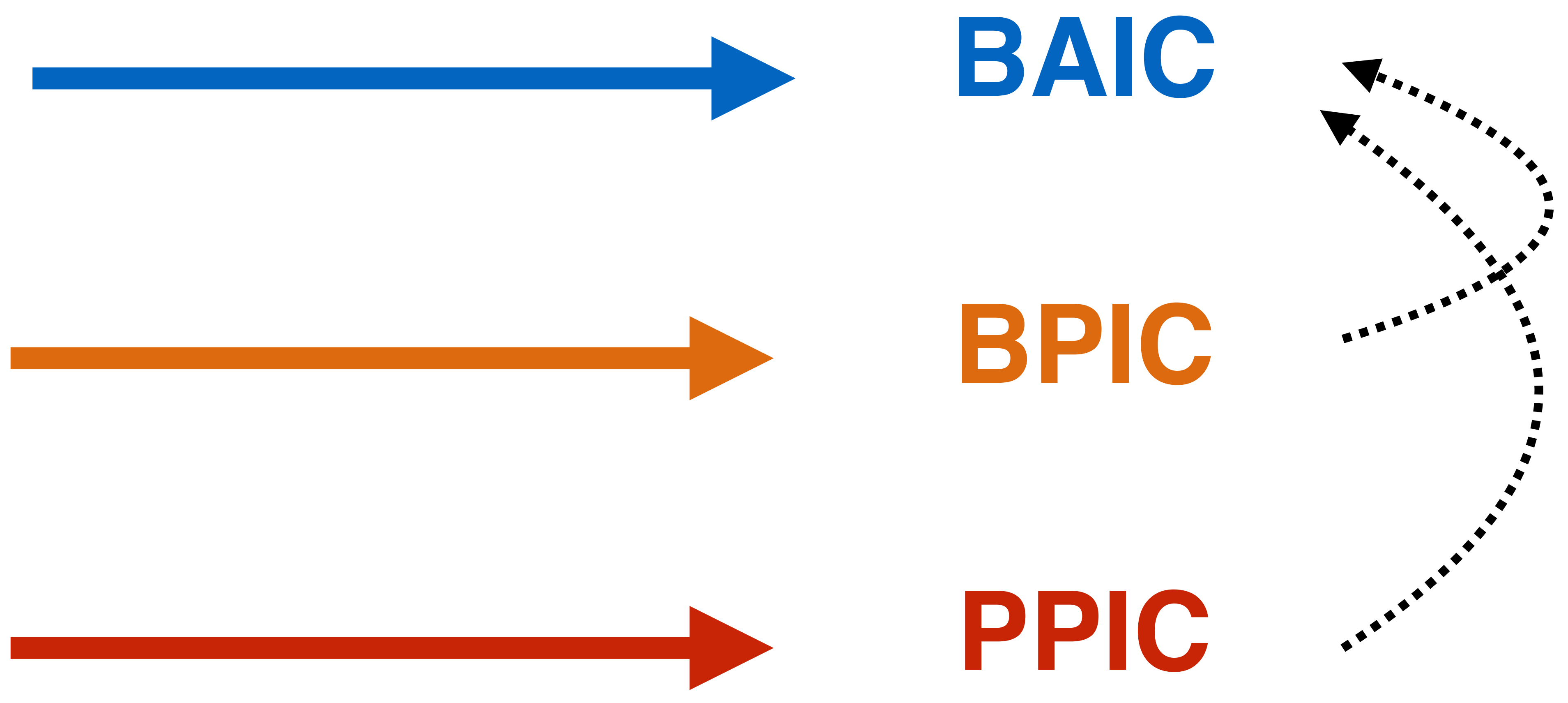
("plug-in")

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

("posterior average")

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

("posterior predictive")



(sample size N -> ∞)

https://github.com/jwsitison/improved_model_avg_paper

Complete formulas

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

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

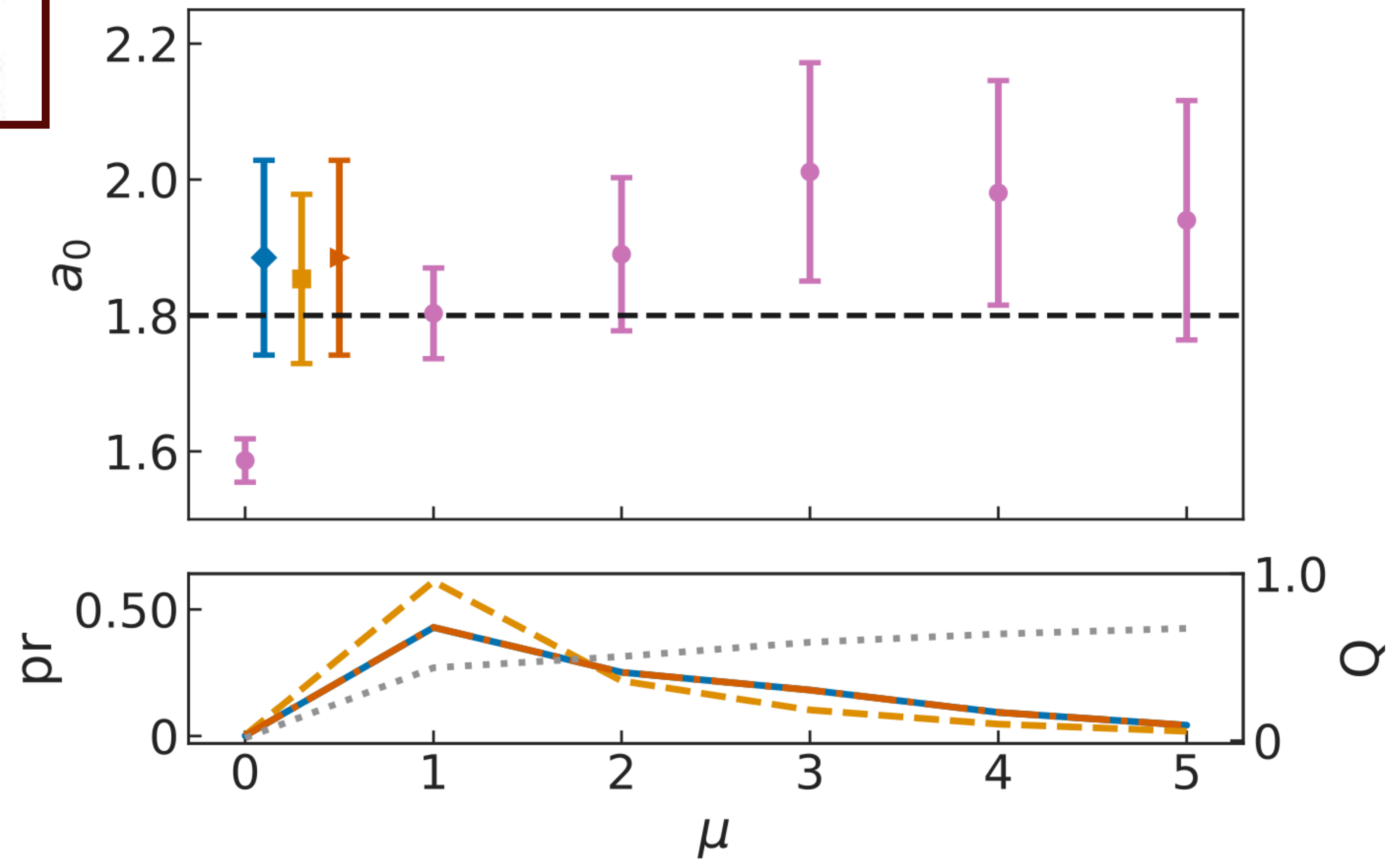
$$\mathrm{PPIC} \approx \hat{\chi}^2(\mathbf{a}^*) + 2k + d_C + N d_C \log\left(1 + \frac{1}{N}\right) - 2\sum_{i=1}^{N} \log\left[1 + \frac{1}{2}\left(\frac{1}{4}(g_i)_b(g_i)_a - \frac{1}{2}(H_i)_{ba}\right)(\Sigma^*)_{ab} + \frac{1}{4}(g_i)_d T_{cba}(\Sigma_2^*)_{abcd}\right]$$

- 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.

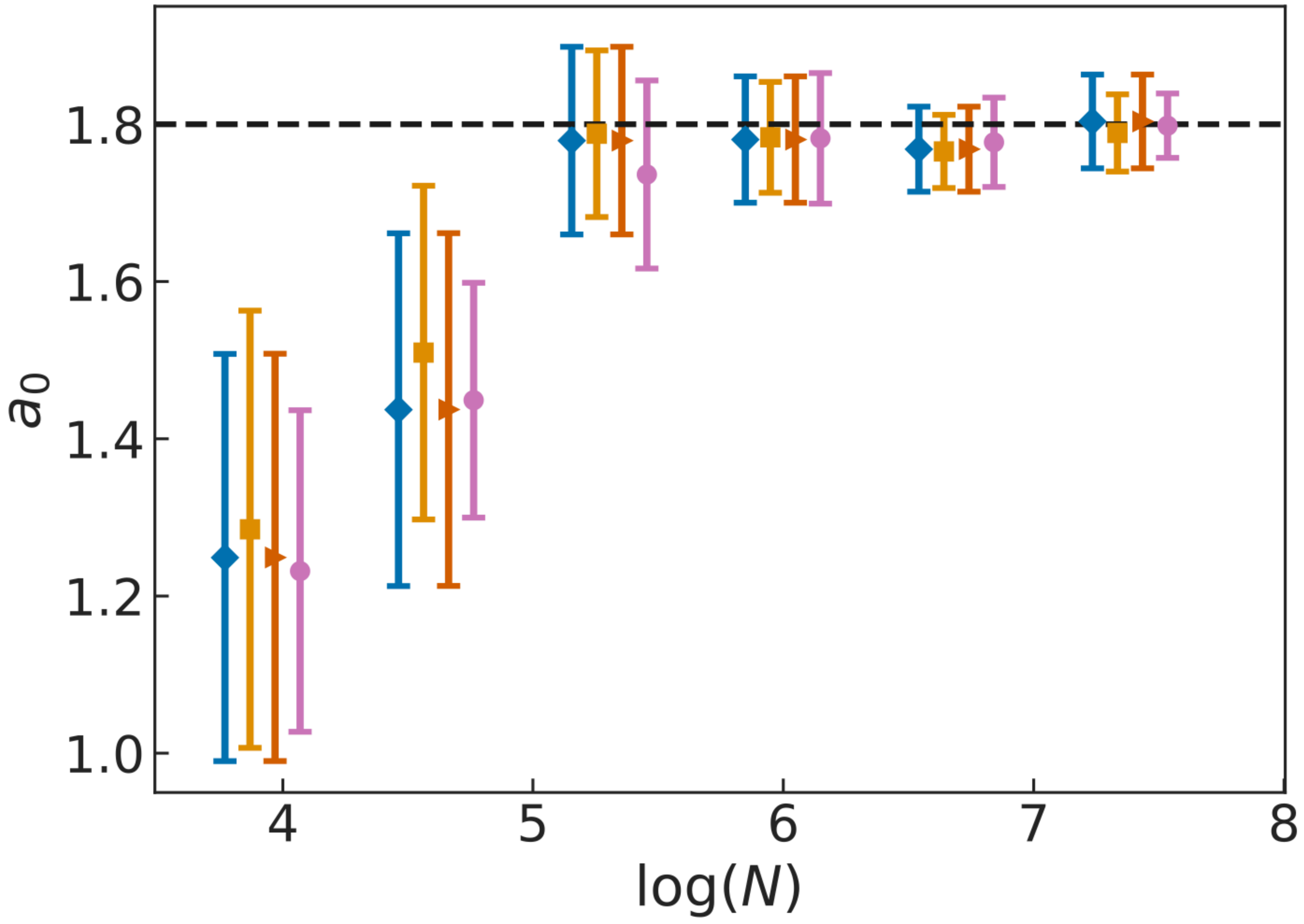
- The above formulas are *approximate*, NLO in large-N expansion (N = data sample size.) PPIC subset penalty is approximately $+2d_C$ plus 1/N corrections. BPIC has larger bias from posterior avg.

- 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

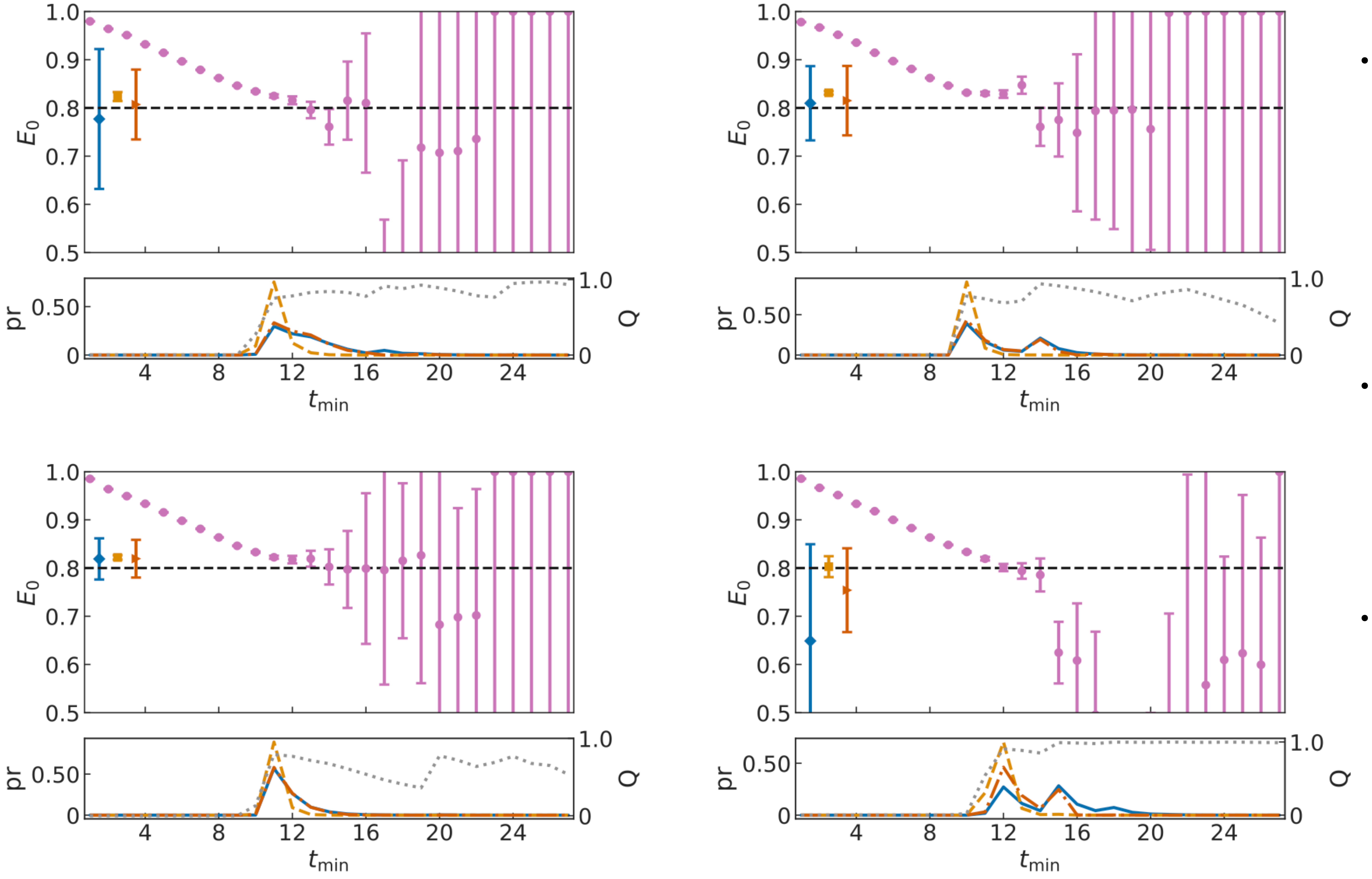


- Quadratic model truth, extract constant term a_0.
- **Left:** fits to polynomials of degree μ. Extra parameters are penalized, moreso for BPIC.



- **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…)

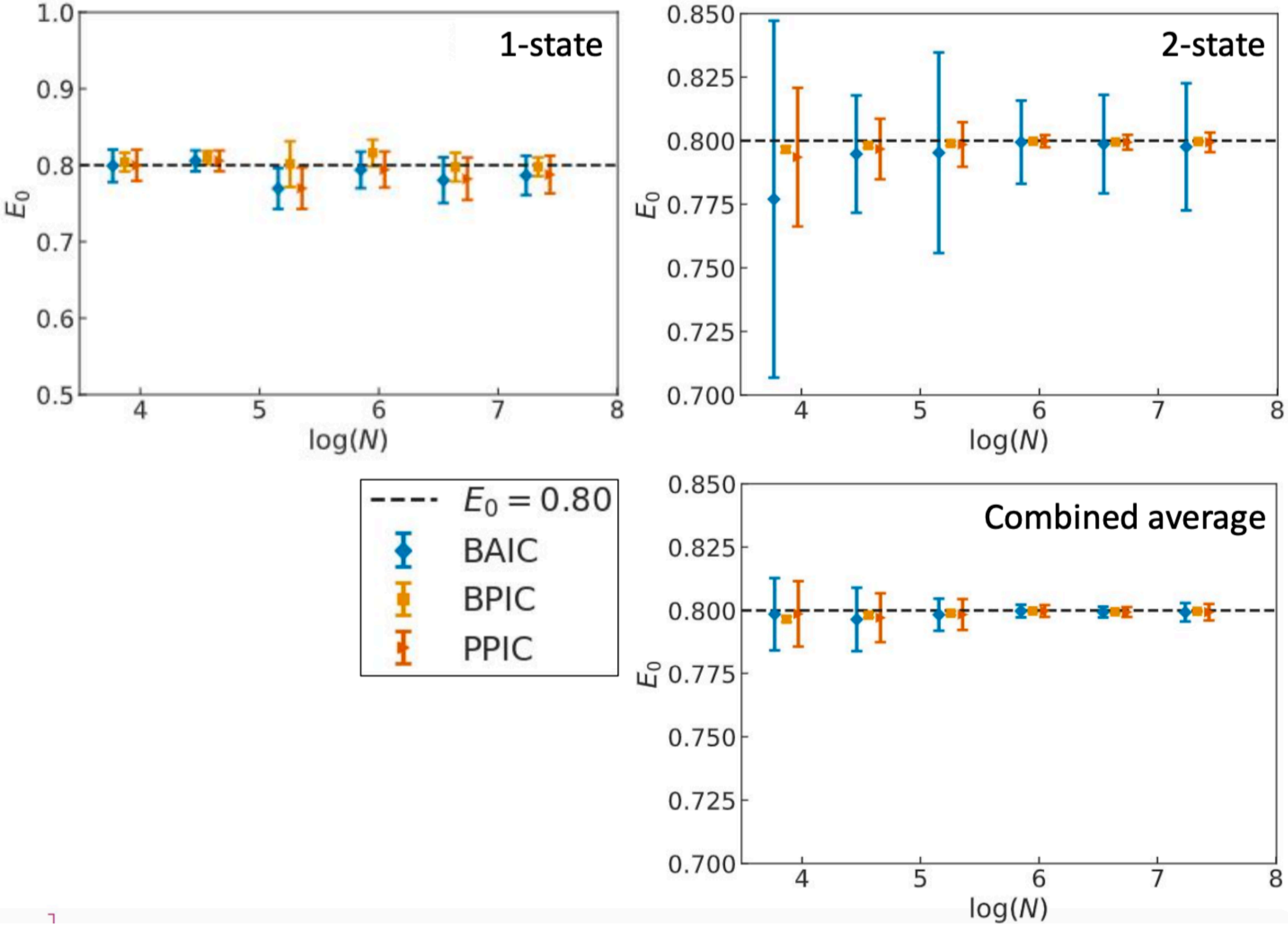
Numerical results: data selection



- 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)



Summary

- Model averaging is a powerful and simple technique for dealing with analysis choices and associated systematic errors. Easy to "plug in" to existing analysis chains where chi-squared fits are done.
- 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.
- Thoughts for PDFs: For methods that aren't chi-squared fits, need to understand *right bias correction* for however you evaluate likelihood of your model being correct… K-L divergence approach? Other issues?

Backup slides

The Kullback-Leibler divergence

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

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

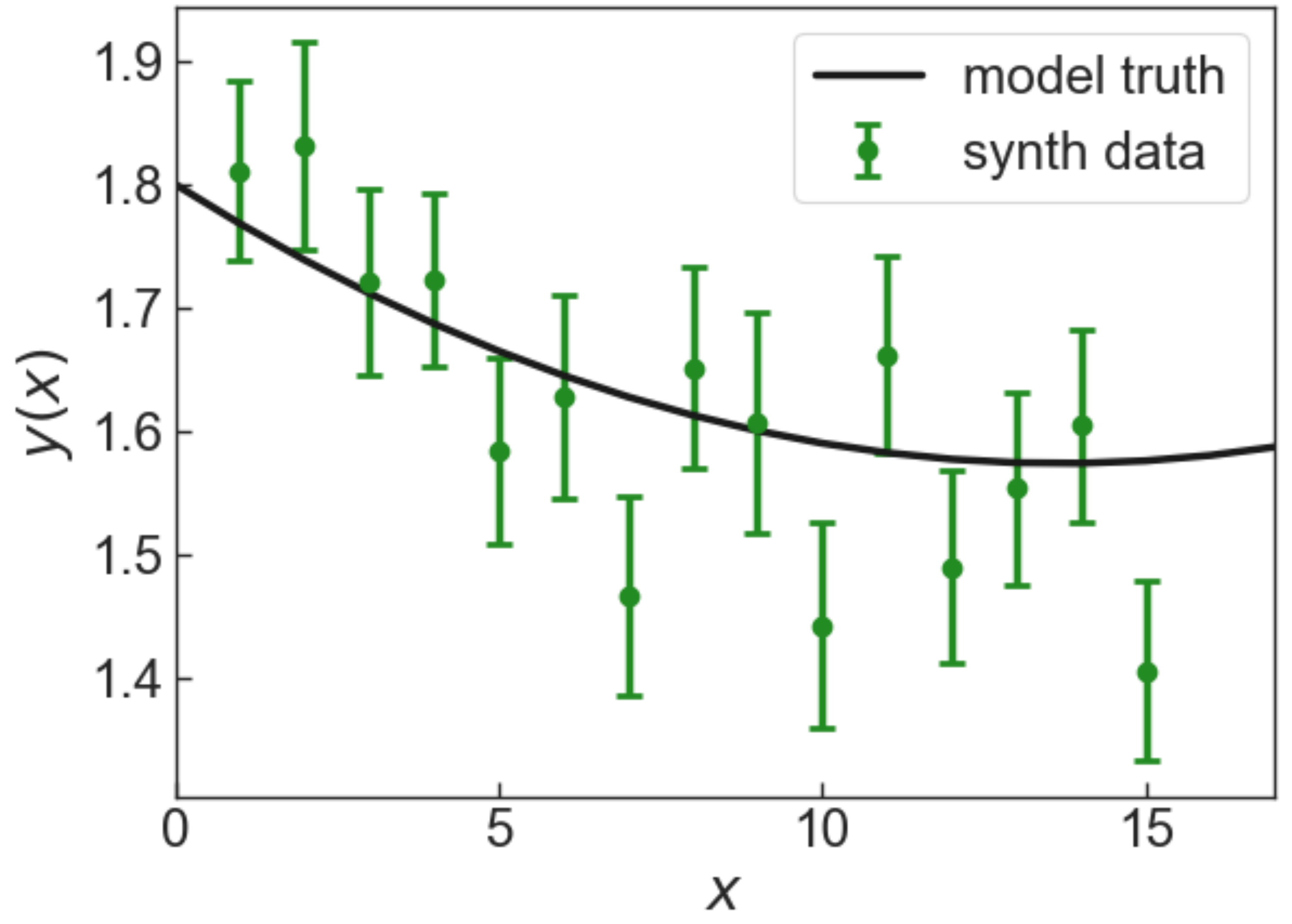
- KL = 0 if the PDFs are equal, positive definite otherwise. Find the "closest" distribution to pr_{MT} by **maximizing** the magnitude of the second term!

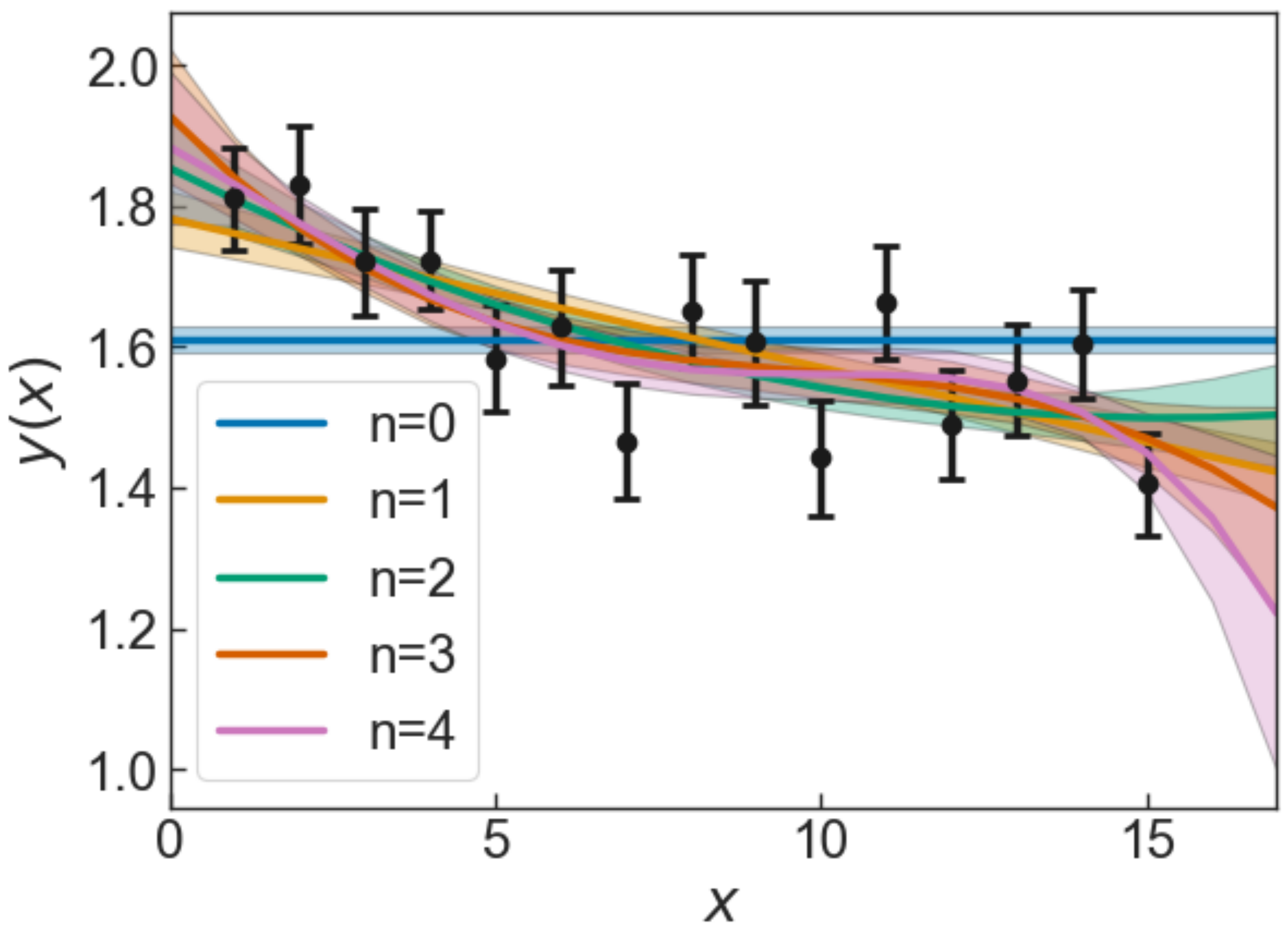
- Introduce model parameters **a**, and this leads to familiar results:

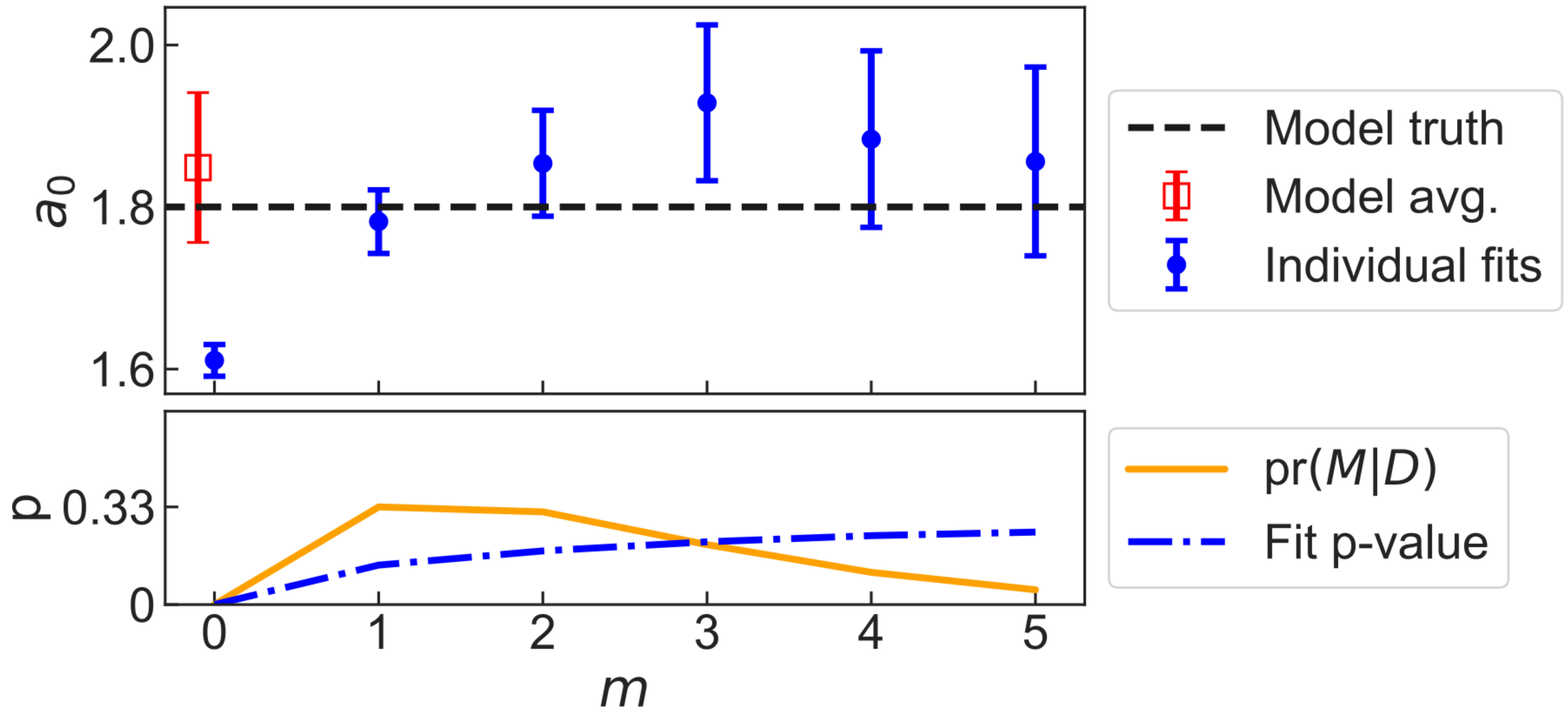
$$E_z[\log \mathrm{pr}(z|\mathbf{a}, M_\mu)] \simeq \frac{1}{N} \sum_i \log \mathrm{pr}(y_i|\mathbf{a}, M_\mu) = \frac{1}{N} \log \mathrm{pr}(\{y\}|\mathbf{a}, M_\mu)$$

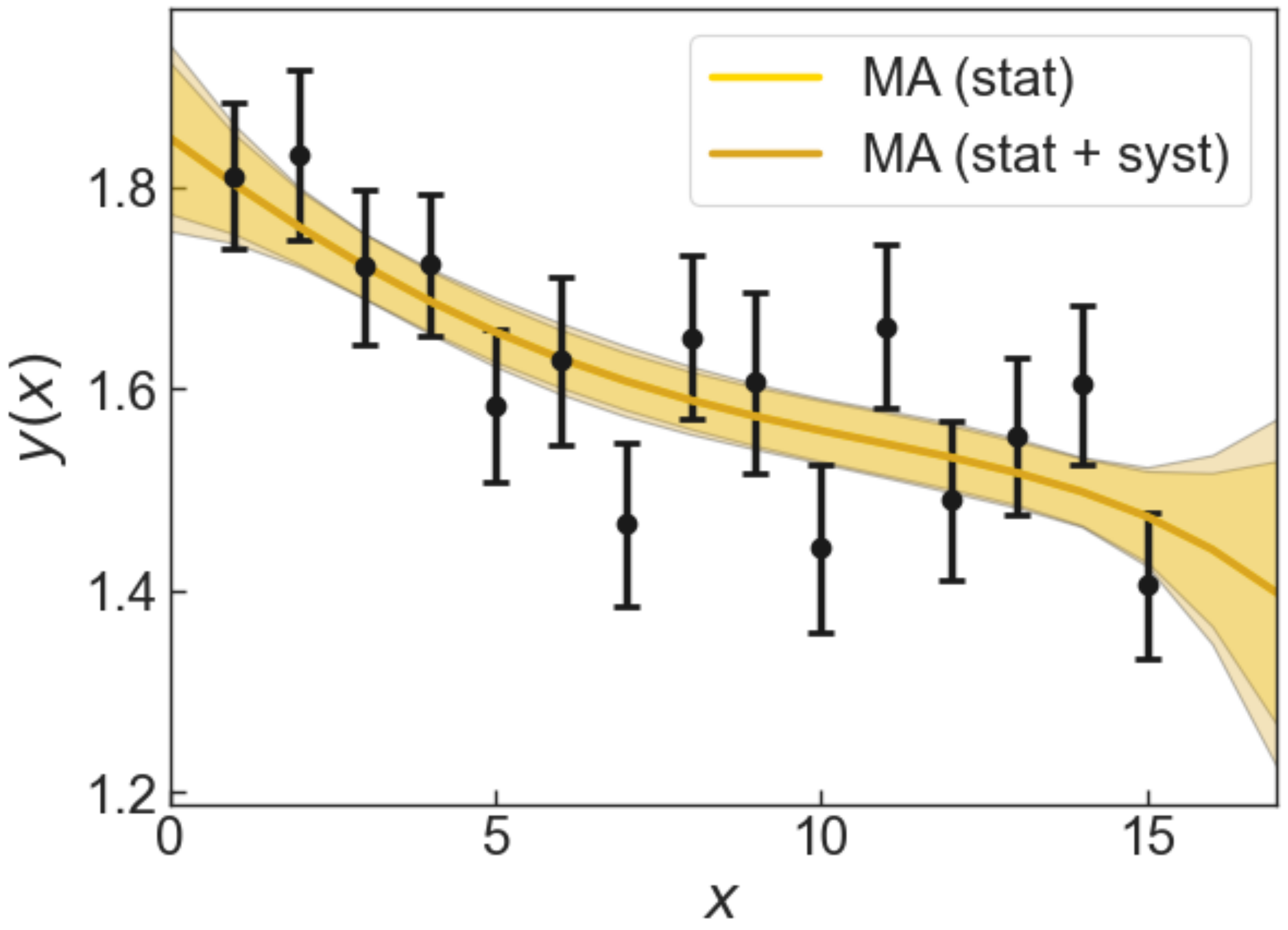
sample log-likelihood, i.e. $-\chi^2/2$

- 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).


model truth
synth data
y(x)
x


n=0
n=1
n=2
n=3
n=4
y(x)
x


Model truth
Model avg.
Individual fits
pr(M|D)
Fit p-value
a0
p
m

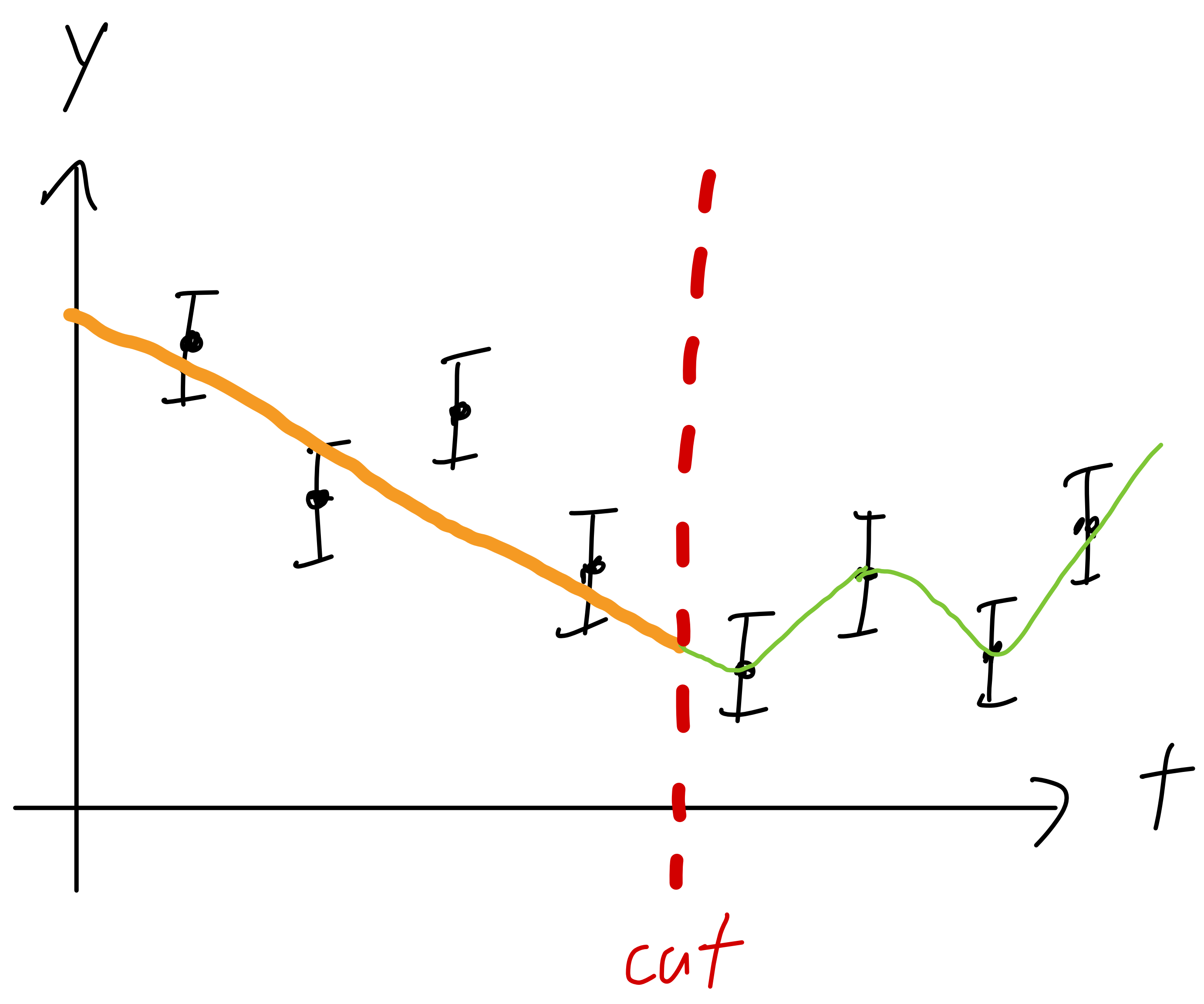

MA (stat)
MA (stat + syst)
y(x)
x

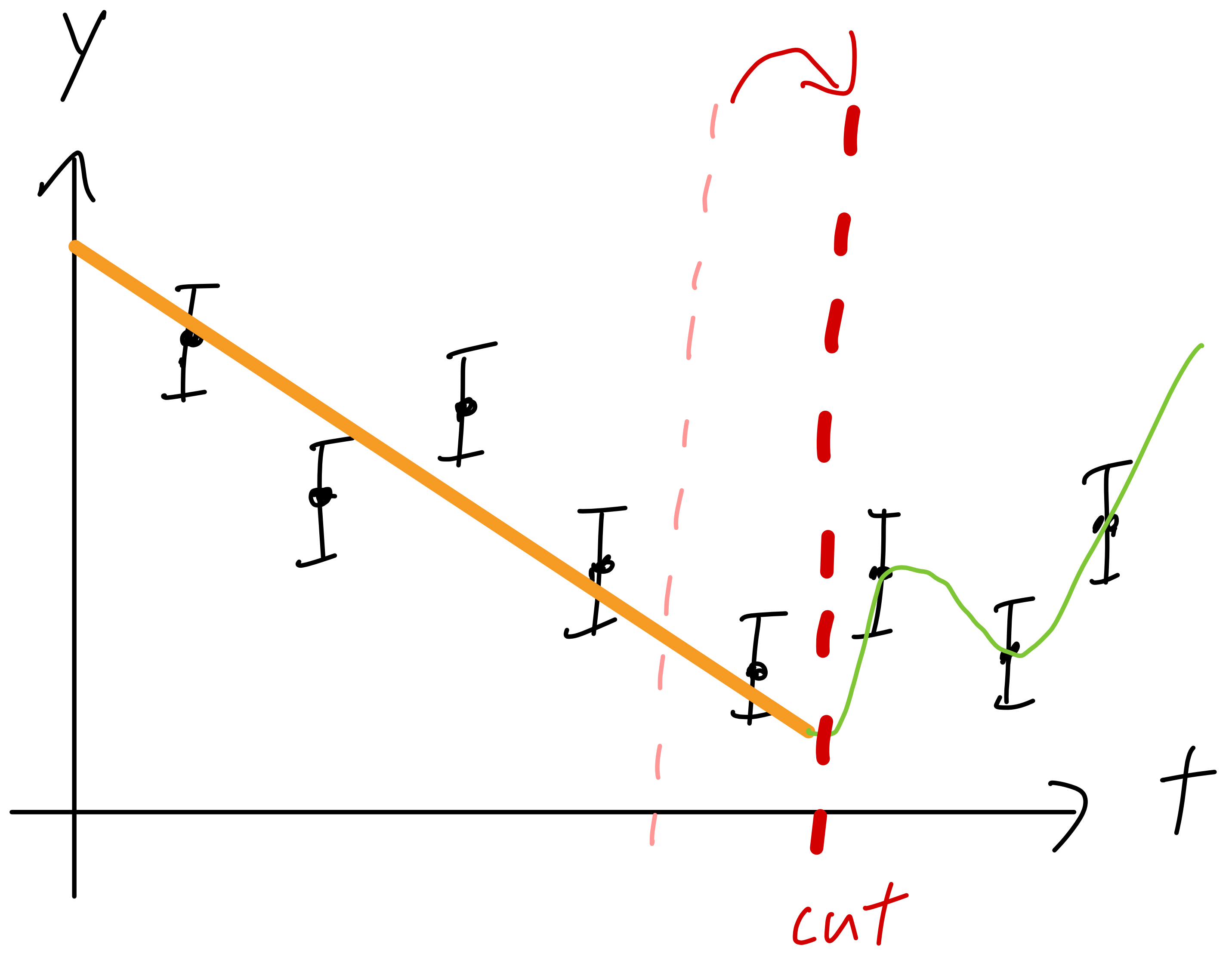
(EN and J. Sitison, arXiv:**2305.19417**)

Two approaches to subset selection

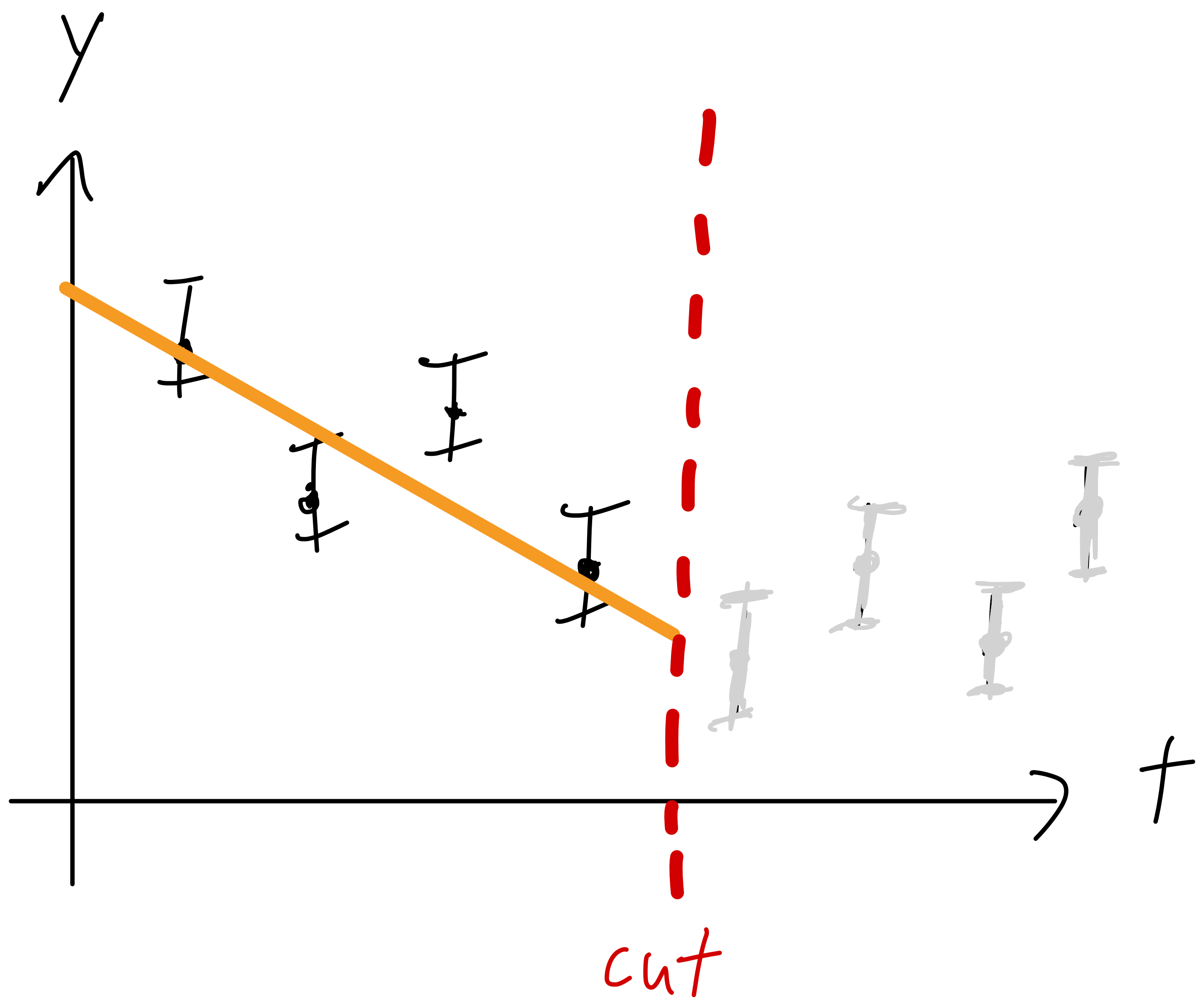
(EN and W. Jay, arXiv:**2008.01069**)

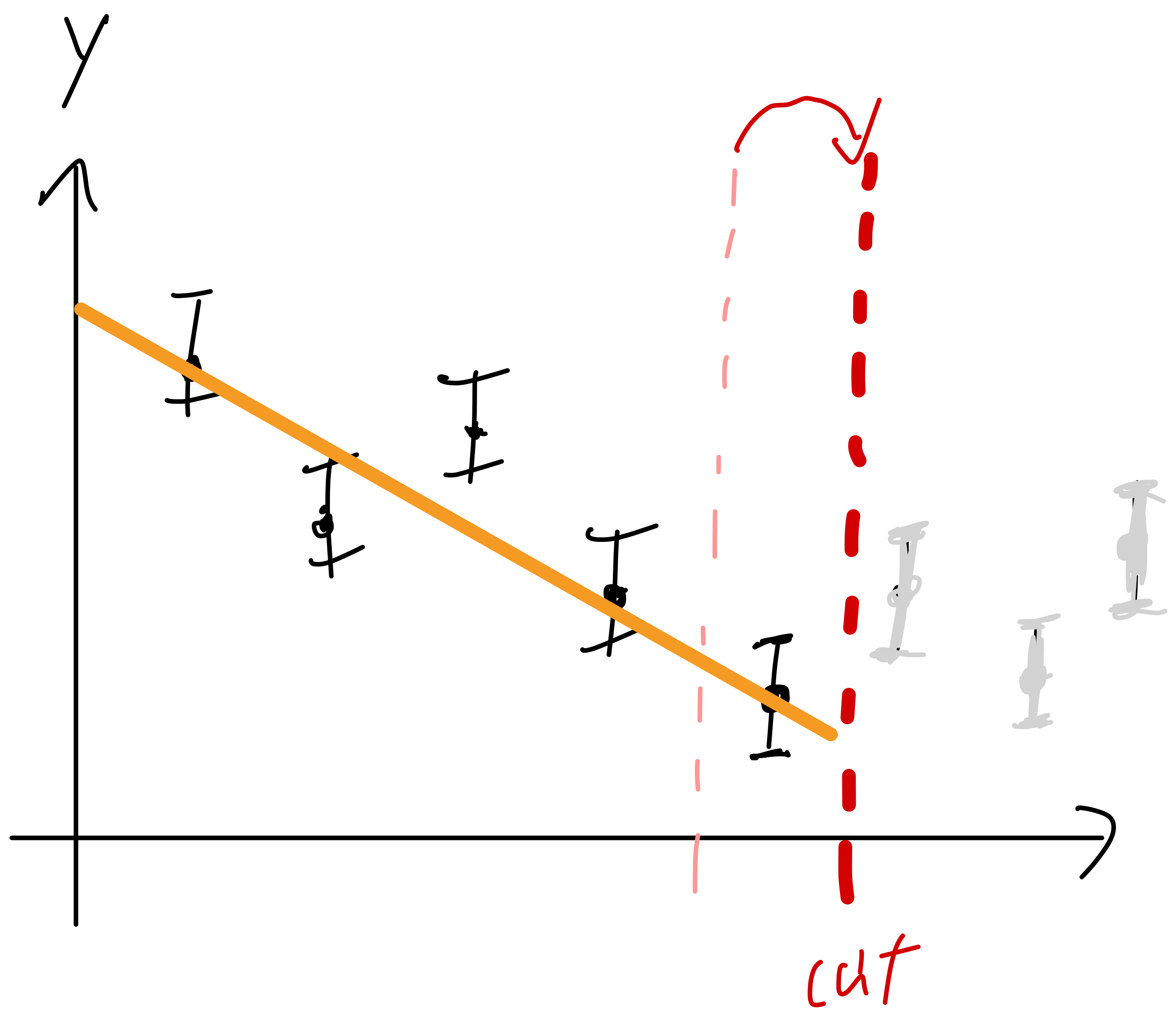
- A common part of lattice analysis is data cutting: "what $[t_{min}, t_{max}]$ should I fit my two-point correlator over?"
- Partition data into kept and cut $[y_K, y_C]$ of size (d_K, d_C). Compute relative model weights, average!
- "Perfect model method": Keep all data. y_C fit to a model with $\chi^2=0$;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

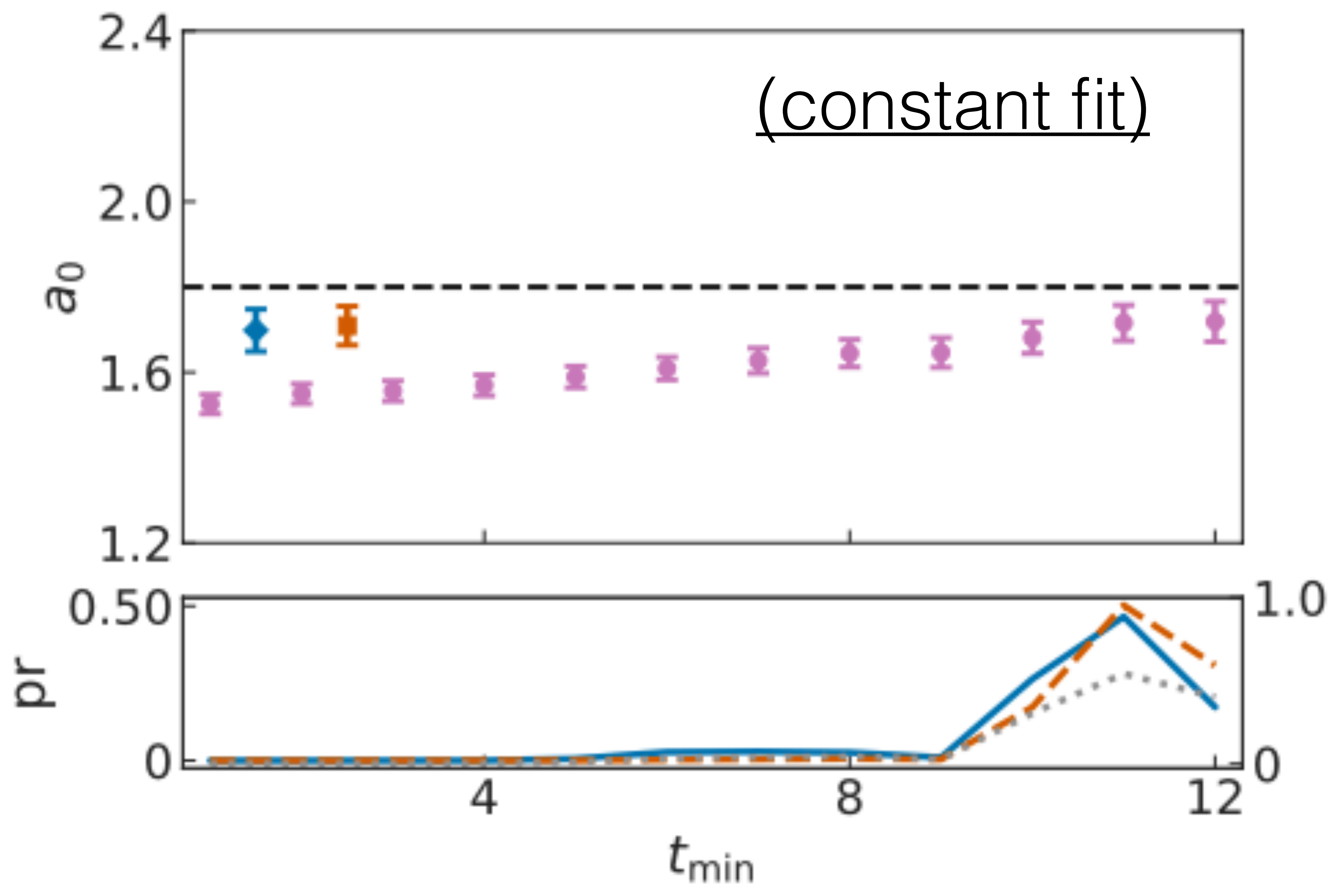


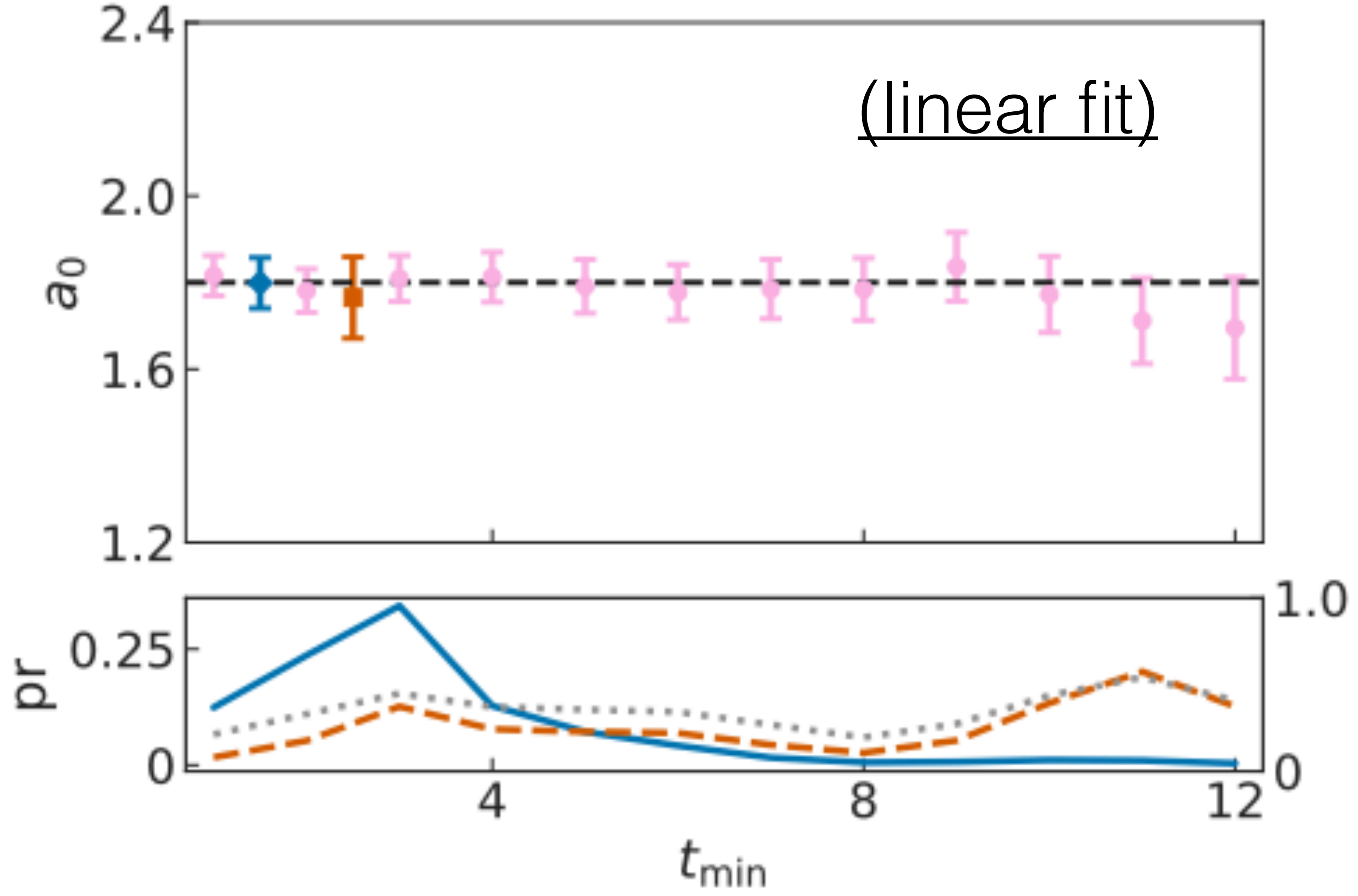


(BMW collab, *Nature* 593 (2021), arXiv:**2002.12347**)









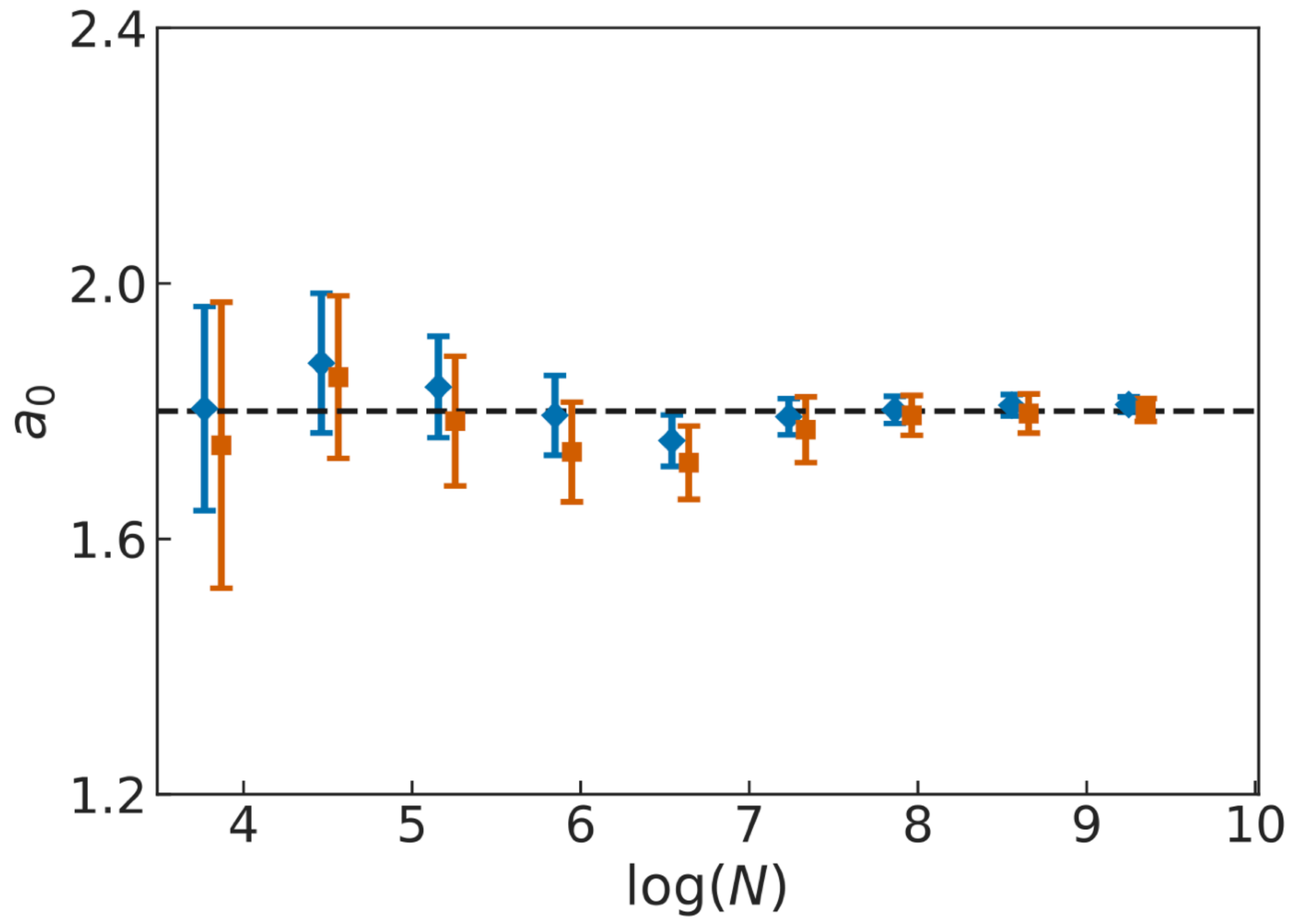
BAIC (perf)
BAIC (subset)

- 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; χ^2 is dominant.
- For linear fit ("true model"), both averages are right, but subset *under-penalizes* cutting 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)



(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, $N_{dof}=d_K-k$:

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

- For a bad fit with large Ndof and $1 < \chi^2 < 2$, we can have $\mathrm{AIC}^{sub} >> 0$ but $\mathrm{AIC}^{perf} << 0$ (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 AIC^{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 $\mathrm{pr}_T(z)$ 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.**

- *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.

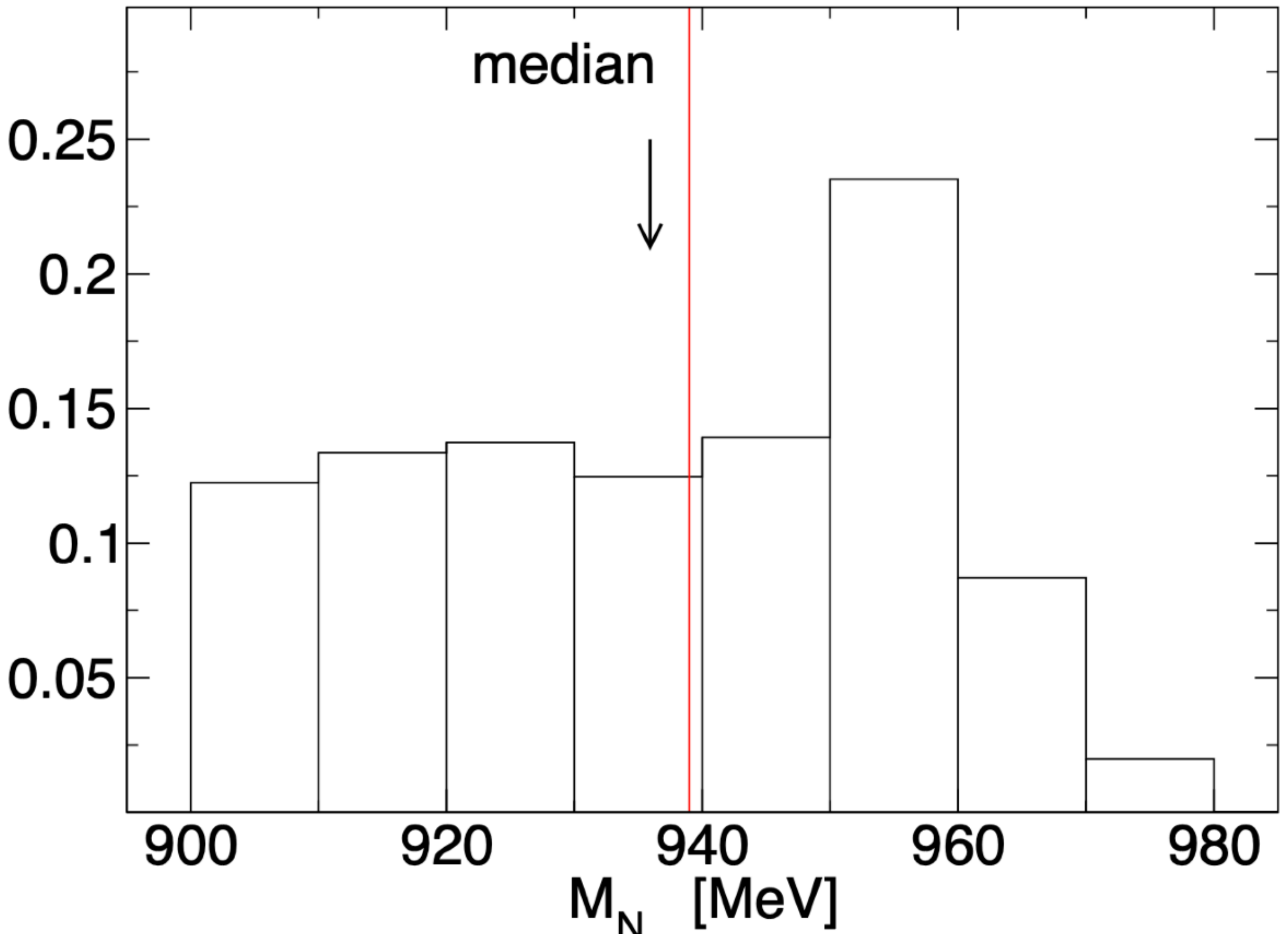
- Several early variations of model averaging/ variation appear in lattice papers: Y. Chen et al. '04, **BMW '08**, **HPQCD '08**, **FNAL/MILC '14**, BMW '14…however, many old papers use *ad hoc* averaging prescriptions.

- 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.)

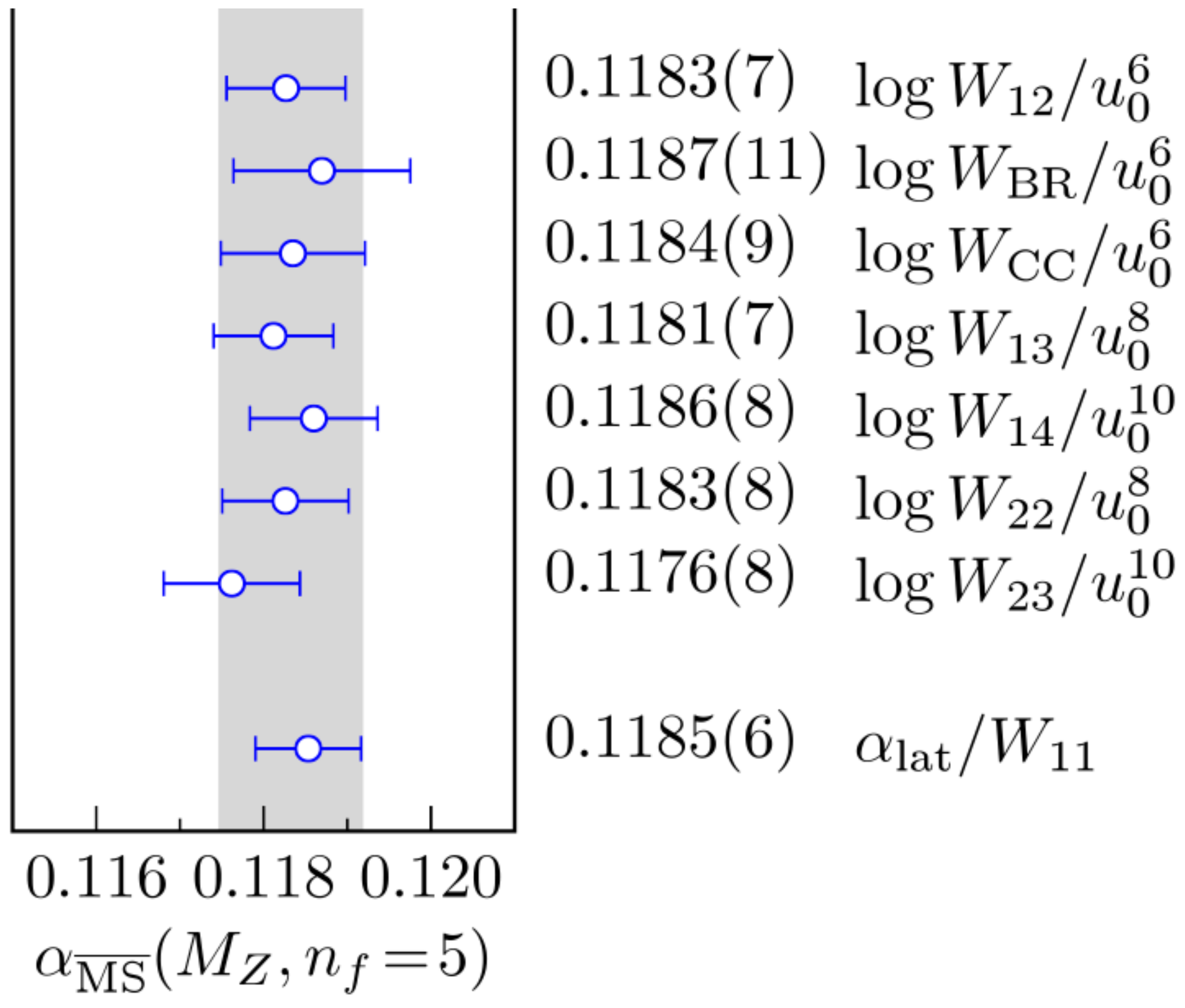
- First use of AIC with data penalty is BMW '21 (although I will argue for a *corrected* version of their formula here.)

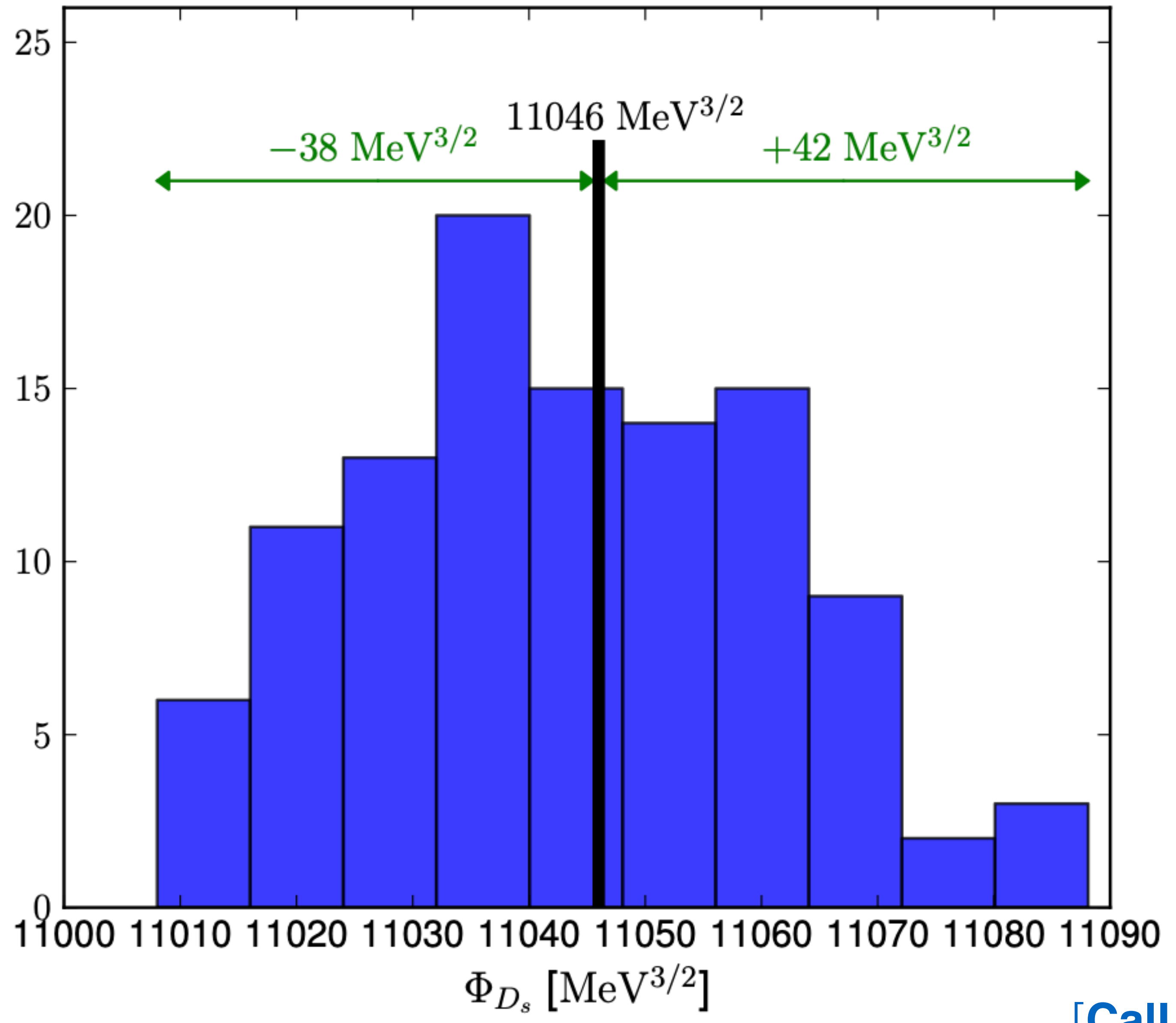
[Y. Chen et al '04]: arXiv:**hep-lat/0405001**
[BMW '14]: PRD 90 (2014), arXiv:**1310.3626**
[BMW '15]: Science 347 (2015), arXiv:**1406.4088**
[Rinaldi et al. '19]: PRD 99 (2019), arXiv:**1901.07519**
[CalLat '20]: PRD 102 (2020), arXiv:**2005.04795**
[BMW '21]: Nature 593 (2021), arXiv:**2002.12347**

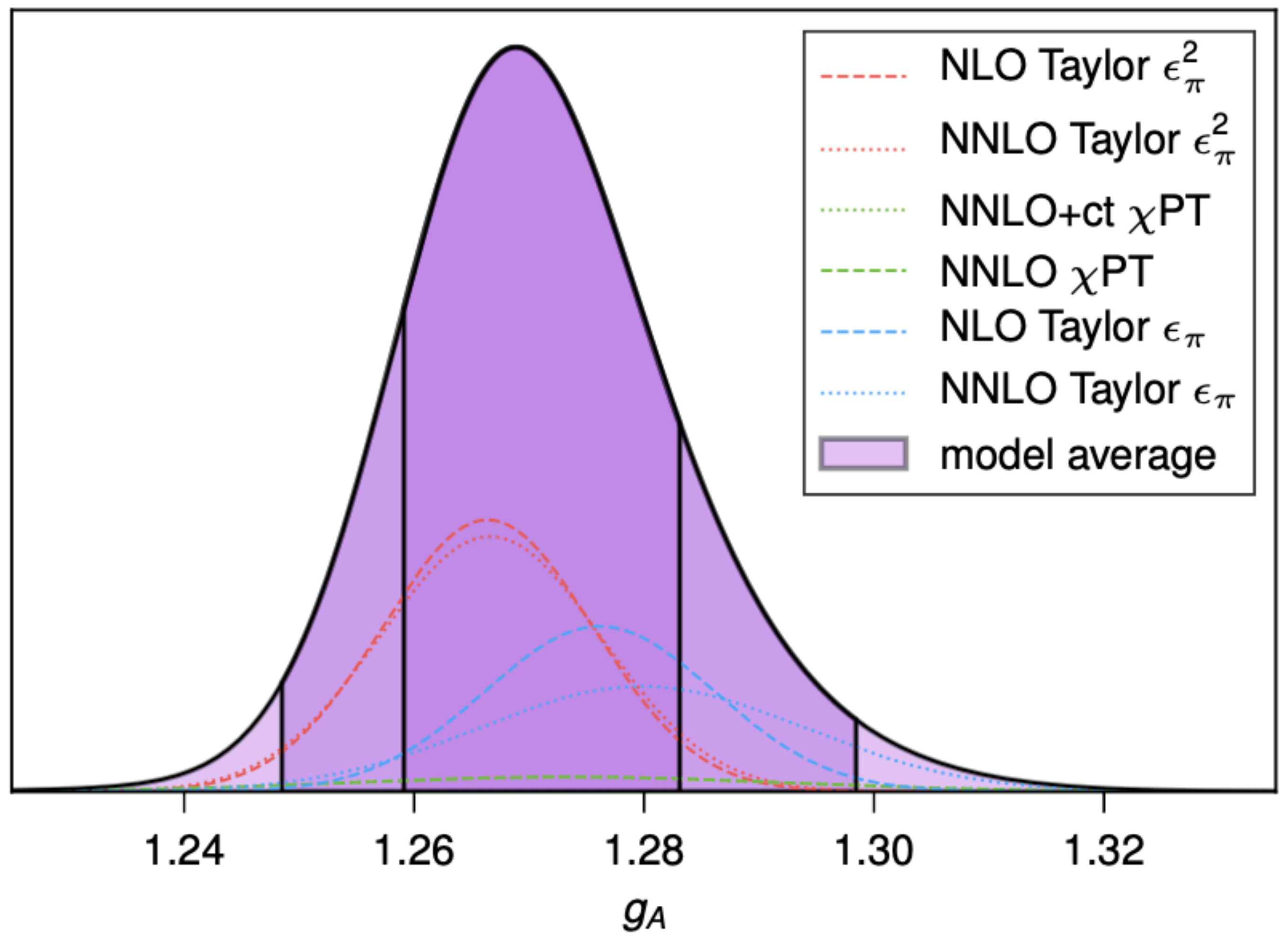
[**BMW '08**]: (BMW collaboration, *Science* 322 (2008), arXiv:**0906.3599**)



[**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**)