


Improved information criteria for Bayesian model averaging in lattice field theory

Ethan T. Neil^{*} and Jacob W. Sitison[†]

Department of Physics, University of Colorado, Boulder, Colorado 80309, USA

 (Received 21 September 2022; revised 24 July 2023; accepted 11 December 2023; published 29 January 2024)

Bayesian model averaging is a practical method for dealing with uncertainty due to model specification. Use of this technique requires the estimation of model probability weights. In this work, we revisit the derivation of estimators for these model weights. Use of the Kullback-Leibler divergence as a starting point leads naturally to a number of alternative information criteria suitable for Bayesian model weight estimation. We explore three such criteria, known to the statistics literature before, in detail: a Bayesian analog of the Akaike information criterion which we call the BAIC, the Bayesian predictive information criterion, and the posterior predictive information criterion (PPIC). We compare the use of these information criteria in numerical analysis problems common in lattice field theory calculations. We find that the PPIC has the most appealing theoretical properties and can give the best performance in terms of model-averaging uncertainty, particularly in the presence of noisy data, while the BAIC is a simple and reliable alternative.

DOI: [10.1103/PhysRevD.109.014510](https://doi.org/10.1103/PhysRevD.109.014510)

I. INTRODUCTION

In many data analysis applications, particularly in lattice gauge theory, model uncertainty is a common challenge. Model uncertainty arises when multiple candidate model descriptions exist for a given set of observations, with the desired analysis results dependent on which model is used. A simple solution to this problem is model selection, i.e., choosing a single model from the available candidates based on some (generally data-driven) criteria. Model selection is appealing due to its relative simplicity: once a model is chosen using some procedure, inference on parameters or prediction of future observations can be done using standard statistical methods within the chosen model.

However, this approach is not always optimal, especially when the primary goal of analysis is parameter inference and model selection is only an intermediate step. By choosing a single “best” model, model selection neglects the effects of model uncertainty compared to other sources of error such as parameter uncertainty from a regression procedure (e.g., least squares) [1–3]. As a result, model selection can lead to overly confident results based on limited statistical information.

To incorporate model uncertainty into statistical analyses, a natural alternative to model selection is model averaging. With model averaging, quantities of interest are determined for each model in a space of candidates, and a final estimate is made by taking a weighted average over

the model-dependent estimates. The weights correspond to how likely each respective model is to describe the observed data. Combining models in this way accounts for the model uncertainties in the overall statistical uncertainty of the analysis. Moreover, the probabilistic weighting of models can yield smaller uncertainties compared to overly conservative procedures such as taking the full difference between plausible model variations as a systematic error, without introducing asymptotic bias.

Bayesian inference gives a natural framework in which to carry out the procedure of model averaging. Specifically, Bayes’ theorem gives a way to construct a posterior distribution over the combined model-parameter product space and allows analysts to incorporate whatever prior information is available. Bayesian model averaging has been well known in the statistics literature for some time [4–8]. The central problem in applying Bayesian model averaging is the estimation of model probability weights, which is generally formulated in terms of quantities known as “information criteria” (ICs). The most well-known information criterion is the Akaike information criterion (AIC) [9–11], which by construction is inherently a frequentist estimator (although a close analog, the “Bayesian AIC,” may be derived in a Bayesian context as we will show).

In this paper, we explore several ICs that may be used to determine model weights for Bayesian model averaging. As a unifying concept, and inspired by the work of Zhou [12,13], we focus on the derivation of information criteria based on the Kullback-Leibler (KL) divergence, which can be thought of as an information-theoretic starting point for

^{*}ethan.neil@colorado.edu

[†]jacob.sitison@colorado.edu

evaluating model probability. Variations on the explicit definition of the KL divergence in the case of parametric models (which are of primary interest for model averaging) are shown to lead to different ICs.

This work is motivated specifically by a need for improved statistical methods in lattice field theory. Bayesian model averaging is well suited for lattice applications because of the notorious stability issues of the functional forms that arise in lattice application (e.g., two-point correlators modeled with an infinite tower of exponentials). As a result, statistical analyses of lattice data typically require model and/or dataset truncation, which introduces systematic errors to a typical model selection procedure. By averaging over models and data subsets (which we will see is equivalent to the general model-averaging framework), this systematic error is accounted for. Furthermore, the firm physical foundation of lattice field theory complements the use of Bayesian inference by giving well-motivated families of models to consider. Other explorations of model averaging in a lattice field theory or effective field theory context include Refs. [14–24]; the AIC was first used in the context of lattice field theory analysis in [25]. Our current work inherits directly from [26], which rigorously studied model averaging for lattice field theory in a Bayesian context.

The remainder of the paper is structured as follows. In the next subsection Sec. IA, we give an overview of our key results. We then review some general results important for model averaging in Sec. II, including the Bayesian framework for model averaging developed in [26] and some general concepts from mathematical statistics; as part of this discussion, we establish how bias on information criteria influences bias of parameter estimates. In Sec. III, we define the KL divergence and give several distinct variations for parametric distributions; here we also introduce the information criteria that result from these variations. We specialize our discussion of Bayesian model averaging to least-squares regression in Sec. IV and derive formulas to approximate the model probability weights from the aforementioned information criteria. In Sec. V, we reformulate the data subset selection problem as one of model variation and derive the corresponding expressions for the information criteria in this case. Section VI gives three numerical examples to demonstrate the performance of each information criterion in model averaging; these include linear least squares applied to a fixed dataset (Sec. VIA), a nonlinear toy problem that resembles fitting a two-point correlation function to demonstrate the effectiveness of model averaging in lieu of manual data subset selection (Sec. VIB), and finally a similar two-point nucleon correlator example on a set of real lattice QCD data (Sec. VIC). Section VII summarizes our findings and gives some concluding remarks.

Appendix A connects the theoretical details of [26] to our updated view in terms of the KL divergence and

provides some additional discussion. In Appendix B, we discuss the asymptotic equivalence of the various information criteria in the limit of infinite data. A bound on the asymptotic bias of model averaging is derived in detail in Appendix C. Another information criterion known as the posterior averaging information criterion (PAIC) was proposed by Zhou [12,13] to generalize and improve the performance of the Bayesian predictive information criterion (BPIC); however, using the same integral approximation as the BPIC and posterior predictive information criterion (PPIC) gives a lower order (in the inverse sample size N^{-1}) approximation to the PAIC and hence worse performance in practice. Therefore, the PAIC is not discussed as thoroughly as the other ICs, but the relevant formulas are given in Appendix D. Appendix E gives a brief derivation of the asymptotic approximation known as Laplace’s method used in Sec. IV as well as some Gaussian integrals used in Sec. V. Appendix F contains an alternative derivation of the data subset selection criteria introduced in Sec. V. Finally, some of the relevant derivative tensors used in the calculations are given in Appendix G.

A. Summary of key results

Since we derive a number of technical results in this paper in some detail, we include here an overview of some of our key findings. Our primary focus is on information criteria, which quantify the (logarithmic) probability of a given model; for a review of the basic formalism of model averaging including a rigorous definition of ICs, see Sec. II A below.

Included in our work are two important clarifications. First, we study the effect of bias in ICs on bias in parameter estimates; our result Eq. (16) establishes that unbiased ICs are important for obtaining unbiased model-averaged parameter estimates. Second, we clarify some key points in how the AIC arises in a Bayesian context compared to the earlier work in [26]; see Sec. III A and Appendix A. We call our revised formula the Bayesian AIC, or BAIC, see Eqs. (35), (65), and (124); the differences between AIC, ABIC_{CV} (our name for the formula defined in [26]) and BAIC are subtle and irrelevant in the limit that the priors do not influence the results.

A central aspect of our work is to approach the problem of data modeling using the Kullback-Leibler divergence as a foundation. Given a “true model” M_T , the KL divergence (or relative entropy) between the true model and a candidate model M_μ is (notation is defined more fully in Sec. III)

$$\text{KL}(M_\mu) = E_z[\log \text{pr}_{M_T}(z)] - E_z[\log \text{pr}_{M_\mu}(z)], \quad (1)$$

where future data z are drawn from the true likelihood $\text{pr}_{M_T}(z)$. Broadly speaking, minimization of this divergence will select the model M_μ that most closely resembles the true model M_T . However, this definition is nonparametric,

and therefore ambiguous when the model likelihood $\text{pr}_{M_\mu}(z|\mathbf{a})$ depends on some fit parameters \mathbf{a} .

This ambiguity allows us to define several variations on $\text{KL}(M_\mu)$, which then lead naturally to different ICs. Each variation can be thought of as representing a choice for how to obtain the nonparametric model predictive distribution $\text{pr}_{M_\mu}(z)$ by starting from a parametric model. For example, we may adopt a “plug-in” estimate using the best-fit parameter value \mathbf{a}^* , leading to a Bayesian version of the well-known Akaike information criterion as described in Sec. III A:

$$E_z[\log \text{pr}_{M_\mu}(z)] \sim E_z[\log \text{pr}_{M_\mu}(z|\mathbf{a}^*)] \simeq -\frac{\text{BAIC}}{2N},$$

where N is the data sample size, “ \sim ” indicates a choice of construction for the nonparametric model predictive distribution, and “ \simeq ” indicates that the IC on the right-hand side may be used to compute an unbiased sample estimate of the term on the left-hand side. The plug-in approach is simple, but unnatural from a Bayesian point of view since it focuses on a single best-fit parameter *value* instead of a posterior *distribution*.

We also explore two alternatives that are more manifestly Bayesian and lead to two other ICs. Specifically, we study the Bayesian predictive information criterion [Eq. (47)] and the posterior predictive information criterion [Eq. (58)]:

$$E_z[\log \text{pr}_{M_\mu}(z)] \sim E_z[E_{\mathbf{a}|\{y\}}[\log \text{pr}_{M_\mu}(z|\mathbf{a})]] \simeq -\frac{\text{BPIC}}{2N},$$

$$E_z[\log \text{pr}_{M_\mu}(z)] \sim E_z[\log E_{\mathbf{a}|\{y\}}[\text{pr}_{M_\mu}(z|\mathbf{a})]] \simeq -\frac{\text{PPIC}}{2N}.$$

These two ICs are not the unique constructions that may be used to estimate the corresponding expectation values; in Sec. III we discuss a wide variety of other ICs that have appeared in the statistics literature before. We emphasize that none of these ICs are new, although we believe that our approach to deriving them in a unified way from variations on the KL divergence is novel.

Although these two constructions look superficially similar, we will find that the form of the PPIC makes it uniquely sensitive to fluctuations within a given data sample, and therefore more robust in the presence of noise. On the other hand, the BPIC is somewhat more aggressive in selecting models with fewer parameters. This can lead to lower variances at the cost of higher bias at finite sample size due to the bias-variance trade-off (see discussion in Sec. II B and explicit demonstration of this effect in our numerical results in Sec. VI).

In Appendix B we demonstrate that the BPIC and PPIC are asymptotically equivalent to the BAIC, so that in the limit of large sample size N all three will give identical results; in this sense, the BPIC and PPIC may be viewed as finite-sample size modifications of the BAIC. Including

the possibility of data subset selection (see Sec. V), our simplified approximate formulas for the case of least-squares fitting (see Sec. IV) are Eqs. (124), (125), and (127); we reproduce these here for convenience:

$$\text{BAIC}_{\mu,P} = \hat{\chi}^2(\mathbf{a}^*) + 2k + 2d_C, \quad (2)$$

$$\begin{aligned} \text{BPIC}_{\mu,P} \approx & \hat{\chi}^2(\mathbf{a}^*) - \frac{1}{2} \tilde{H}_{ba}(\Sigma^*)_{ab} + \frac{1}{2} \tilde{g}_d T_{cba}(\Sigma^*)_{abcd} \\ & + 3k + 3d_C, \end{aligned} \quad (3)$$

$$\begin{aligned} \text{PPIC}_{\mu,P} \approx & \hat{\chi}^2(\mathbf{a}^*) + 2k + d_C + Nd_C \log \left(1 + \frac{1}{N} \right) \\ & - 2 \sum_{i=1}^N \log \left[1 + \frac{1}{2} \left(\frac{1}{4} (g_i)_b (g_i)_a - \frac{1}{2} (H_i)_{ba} \right) \right. \\ & \left. \times (\Sigma^*)_{ab} + \frac{1}{4} (g_i)_d T_{cba}(\Sigma^*)_{abcd} \right], \end{aligned} \quad (4)$$

where k is the number of model parameters, d_C is the number of cut data points, and the other symbols (defined in Sec. IV) represent various derivatives of χ^2 functions. For use in model averaging, all of these ICs should also include a model prior probability term $-2 \log \text{pr}(M_\mu)$ when it is nonconstant, see Sec. II A.

The above formulas for BPIC and PPIC are approximate, based on expansion of integrals in the inverse sample size $1/N$, as discussed in Sec. IV. For the BPIC and PPIC, we recommend the use of these formulas combined with an “optimal truncation” prescription, explained in Sec. IV F, based on the theory of superasymptotics. In the numerical tests we have performed, optimal truncation improves the agreement of these formulas with direct numerical evaluation of the associated integral formulas. Optimal truncation has the additional practical benefit of ensuring that the sum appearing in the PPIC only includes logarithms with positive argument, so that the formula is always well defined.

In order to understand the practical performance of these ICs, we carry out numerical tests on both synthetic data and on real lattice QCD data. As we will see, based on both theoretical considerations and numerical performance, the PPIC is generally the most attractive information criterion for Bayesian model averaging. The BAIC, although its performance in terms of uncertainty is somewhat worse in certain tests, is by far the simplest IC and often gives indistinguishable results from the more complex and expensive to calculate PPIC. Based on these results, we recommend the PPIC as the primary information criteria for Bayesian model averaging in all cases, with BAIC as a backup option in use cases where computing the PPIC is impractical. The BPIC is more aggressive in penalizing model complexity, particularly in the context of data subset selection, which can lead to statistically significant biases at

finite sample size as seen in our numerical examples in Sec. VI. As a result, we do not generally recommend the use of BPIC in practice.

II. MODEL AVERAGING AND BIAS CORRECTION

In this section, we review some preliminary material necessary to understand and motivate the content of subsequent sections, including basic concepts of model averaging and statistical bias. We then examine the relationship between bias in information criteria and bias in model-averaged statistical estimates, finding that the use of asymptotically unbiased ICs is key.

A. Bayesian model-averaging procedure

Bayesian model averaging is a tool that allows for quantitative treatment of uncertainty due to model choice, in situations where many candidate models can plausibly describe a given dataset. Such problems occur commonly in lattice field theory. Even in situations where physical arguments strongly motivate the use of a single theory to describe the data, often the theory is an effective field theory that must be truncated, and uncertainty in the order of truncation is equivalent to model selection uncertainty. For a detailed discussion of Bayesian model averaging with derivations of the basic formulas appearing here, see [26].

Suppose we are interested in determining expectation values of functions of some model parameters \mathbf{a} , marginalized over a set of models $\{M\}$ from a set of data $\{y\}$. The key idea behind Bayesian model averaging is that we can obtain these expectation values as a weighted average over models,

$$\langle f(\mathbf{a}) \rangle = \sum_{\mu} f(\mathbf{a}_{\mu}^*) \text{pr}(M_{\mu}|\{y\}), \quad (5)$$

$$\sigma_{f(\mathbf{a})}^2 = \langle f(\mathbf{a})^2 \rangle - \langle f(\mathbf{a}) \rangle^2 \quad (6)$$

$$= \sum_{\mu} \sigma_{f(\mathbf{a}_{\mu}^*)}^2 \text{pr}(M_{\mu}|\{y\}) + \sum_{\mu} f(\mathbf{a}_{\mu}^*)^2 \text{pr}(M_{\mu}|\{y\}) - \left(\sum_{\mu} f(\mathbf{a}_{\mu}^*) \text{pr}(M_{\mu}|\{y\}) \right)^2, \quad (7)$$

where \mathbf{a}_{μ}^* denotes the best-fit parameters for the model M_{μ} , and the probabilities $\text{pr}(M_{\mu}|\{y\})$ (the ‘‘model weights’’) represent the probability of each model given the data. The quantity $\sigma_{f(\mathbf{a})}^2$ is the estimated variance of the expectation value $\langle f(\mathbf{a}) \rangle$, and includes contributions both from statistical error within each model (the first term) as well as a ‘‘systematic error’’ contribution due to variation of the individual model estimates (second and third terms); see [26] for further discussion. The central problem in computing expectation values is thus to determine the model weights $\text{pr}(M_{\mu}|\{y\})$. From Bayes’ theorem,

$$\text{pr}(M_{\mu}|\{y\}) = \frac{\text{pr}(\{y\}|M_{\mu})\text{pr}(M_{\mu})}{\text{pr}(\{y\})}, \quad (8)$$

where $\text{pr}(M_{\mu})$ is the model prior probability. We will only consider cases where the data $\{y\}$ is fixed for all candidate models M_{μ} , so $\text{pr}(\{y\})$ will henceforth be omitted when irrelevant. The sum of model weights over the space of all models is normalized to 1,

$$\sum_{\mu} \text{pr}(M_{\mu}|\{y\}) = 1. \quad (9)$$

At this point, we remark on the connection of model weights to the common idea of an information criterion. Most ICs are defined explicitly in terms of a likelihood function $\text{pr}(\{y\}|M_{\mu})$, so that a generic information criterion is

$$\text{IC}_{\mu} \equiv -2 \log \text{pr}(\{y\}|M_{\mu}). \quad (10)$$

By Bayes’ theorem, we may define a similar concept of information criterion for use in model averaging simply by including the model prior probabilities,

$$\text{IC}_{\text{MA},\mu} \equiv -2 \log \text{pr}(M_{\mu}|\{y\}) = -2 \log \text{pr}(M_{\mu}) + \text{IC}_{\mu}, \quad (11)$$

where the subscript ‘‘MA’’ denotes model averaging. We will generally work with the former version of the ICs in the text below, to avoid repeatedly writing the factor $-2 \log \text{pr}(M_{\mu})$ that is shared between all of them.

We note in passing that any constant terms (i.e., identical for all models considered) in the definition of an information criterion can be safely ignored, since they will cancel when the normalization condition Eq. (9) is applied. This applies to the factor $-2 \log \text{pr}(M_{\mu})$ in the case of a flat model prior, i.e., if equal prior probability is assigned to all models M_{μ} then this term becomes a constant and drops out. We also exploit this observation to define an equivalent formula for the unnormalized model weight,

$$\text{pr}(M_{\mu}|\{y\}) \propto \exp[-(\text{IC}_{\text{MA},\mu} - \min_{\nu} \text{IC}_{\text{MA},\nu})/2]. \quad (12)$$

This form of the model weight formula is less prone to numerical instability when working at fixed floating-point precision, and is used in practice in our numerical implementations.

B. Bias correction and model averaging

There are a staggering number of information criteria present in the statistics literature. To motivate a specific subset of ICs to study, we first introduce the concept of bias for statistical estimators. Roughly speaking, bias measures the difference between an estimator and the true population value that the estimator is intended to reflect. There are

many possible sources of bias in any statistical study; we will focus here on estimator bias, arising specifically from the choice of sample estimator and not from other systematic effects.

Suppose that $\{y\}$ is a random sample of size N drawn independently from an unknown true distribution with probability density function $\text{pr}_{M_T}(z)$ (i.e., $\{y\}$ are iid samples). Consider a sample estimator $X(\{y\})$ for some property ξ of the true underlying probability distribution $\text{pr}_{M_T}(z)$ from which N independent data samples are drawn. The bias of $X(\{y\})$ is defined as [27]

$$b_y[X(\{y\})] \equiv E_y[X(\{y\}) - \xi] = E_y[X(\{y\})] - \xi, \quad (13)$$

where E_y denotes expectation with respect to the population distribution. An unbiased estimator satisfies $b_y[X(\{y\})] = 0$. The quantity

$$b_z[X(z)] = \lim_{N \rightarrow \infty} b_y[X(\{y\})] \quad (14)$$

is known as the asymptotic bias of the sequence of estimators $\{X(\{y\})\}_{N \in \mathbb{N}}$.

Obviously, it is ideal if one can find a sample estimator that is unbiased at finite N . However, it is not always practical to calculate (and hence correct for) the bias of a given estimator *a priori*. Instead, one can settle for removal of only the asymptotic bias. In the context of lattice simulations, where lattice “data” are generated through a Monte Carlo process, the sample size N tends to be quite large and it can always (in principle) be extended in order to approach the $N \rightarrow \infty$ limit. For this reason, we insist on asymptotic unbiasedness as a primary quality of interest in lattice applications; this guarantees at least that any estimator bias will vanish in the large- N limit. For lattice applications where the goal is typically inference of some physical parameters in a well-motivated theoretical model, this requirement ensures that parameter estimates will converge to the correct answers as $N \rightarrow \infty$.

It is important to emphasize that this goal (removal of asymptotic bias) is not universal across all fields of research. For example, in machine learning the model space is much less well understood, and the primary goal is generally out-of-sample prediction rather than parameter inference. As a result, machine learning applications are often better served by joint optimization of bias and variance; for an accessible review of this so-called “bias-variance trade-off,” see [28]. As will be demonstrated in Sec. VI, the use of model averaging itself represents a form of bias-variance trade-off: inference with a single fixed model will typically have lower variance than a model average but at the risk of asymptotic bias if the model is wrong.

It is important to place the idea of bias properly in the present context of model selection. Suppose that within the space of models $\{M_\mu\}$, there is one model M_T that

corresponds to the true distribution $\text{pr}_{M_T}(z)$ (assuming that any model parameters are set to their correct asymptotic values \mathbf{a}_T^*). Assuming that M_T is in the space of candidate models $\{M_\mu\}$,¹ asymptotically we should find that

$$\lim_{N \rightarrow \infty} \text{pr}(M_\mu | \{y\}) = \text{pr}(M_\mu | z) = \begin{cases} 1, & \mu = T, \\ 0, & \mu \neq T. \end{cases} \quad (15)$$

As discussed above, our primary goal is to remove asymptotic bias from model-averaged estimates. If we assume that the model parameter estimation procedure is consistent² (this is true for, say, least-squares regression), then the asymptotic bias of Eq. (5) is bounded by

$$|b_z[\langle f(\mathbf{a}) \rangle]| \leq \sum_{\mu} |f(\mathbf{a}_\mu^*)| |b_z[\text{pr}(M_\mu | z)]|, \quad (16)$$

with probability 1. For a derivation of this bound and the formal definition of consistency, see Appendix C. Therefore, we can eliminate the asymptotic bias from model-averaged results by using an asymptotically unbiased estimator of the model weights, i.e., $b_z[\text{pr}(M_\mu | z)] = 0$.

It is worth noting in passing that the effect of bias on model-averaged results can be somewhat subtle. As discussed briefly in [26], if several models give near-identical estimates $\langle X \rangle_M$ for some expectation value $\langle X \rangle$, then even the use of a biased model weight estimator will not lead to any significant bias in the estimate for $\langle X \rangle$ itself. Nevertheless, we will insist that all of our model weight estimators be asymptotically unbiased.

III. KULLBACK-LEIBLER DIVERGENCE AND INFORMATION CRITERIA

The problem of estimating model probabilities can be reformulated in terms of the KL divergence, which measures the deviation of a candidate distribution from an underlying true distribution. The KL divergence can be seen as a starting point for the standard methods of model fitting and model weight estimation. Framing the problem of model averaging by beginning with the KL divergence will lead us naturally to the construction of alternative model weight estimators.

In [26], a specific formula for model weight was derived using basic manipulations of probability formulas. (For a

¹The assumption that there is only one model M_T in the space of candidates is for simplicity. For example, say the true distribution is nested within two candidate models $M_{T,1}$ and $M_{T,2}$. In this case, $\lim_{N \rightarrow \infty} (\text{pr}(M_{T,1} | \{y\}) + \text{pr}(M_{T,2} | \{y\})) = \text{pr}(M_{T,1} | z) + \text{pr}(M_{T,2} | z) = 1$, $\lim_{N \rightarrow \infty} \mathbf{a}_{T,1}^* = \lim_{N \rightarrow \infty} \mathbf{a}_{T,2}^* = \mathbf{a}_T^*$, and the model-averaged results Eq. (5) and (7) will be the same as if there were only one true model in the space of candidates.

²Informally, consistency here means that the parameter estimates converge in probability to their true, asymptotic values; see Appendix C for a formal definition.

detailed discussion of that paper’s results and how they can be connected to the present analysis, see Appendix A.) However, if we view the central problem as estimation of probability distributions over the data $\{y\}$, then the model weight formula of [26] is not unique; alternative methods of dealing with the model parameters can be used to give alternative estimators for the model weights. To understand this concept, we step back to understand how the problems of model selection and model fitting can be fundamentally viewed in terms of the KL divergence. This approach, and many of the specific information criteria that we will consider as a result, follows closely the work of Zhou [12,13].

Suppose that $\{y\}$ is a random sample of size N drawn independently from an unknown true distribution with probability density function $\text{pr}_{M_T}(z)$. The basic goal of data modeling is to approximate pr_{M_T} as closely as possible with a model distribution pr_{M_μ} . We may evaluate the “closeness” of a given model distribution to the true probability density with the KL divergence [29],

$$\text{KL}(M_\mu) \equiv \int dz [\text{pr}_{M_T}(z) \log \text{pr}_{M_T}(z) - \text{pr}_{M_T}(z) \log \text{pr}_{M_\mu}(z)] \quad (17)$$

$$= \int dF_{M_T}(z) [\log \text{pr}_{M_T}(z) - \log \text{pr}_{M_\mu}(z)] \quad (18)$$

$$= E_z[\log \text{pr}_{M_T}(z)] - E_z[\log \text{pr}_{M_\mu}(z)], \quad (19)$$

where $F_{M_T}(z)$ is the cumulative distribution function for future observation z drawn from $\text{pr}_{M_T}(z)$, and $E_z[\dots]$ denotes an expectation with respect to the true distribution. The KL divergence, which is also known as the relative entropy, measures the information loss in the estimation of $\text{pr}_{M_T}(z)$ with the model distribution $\text{pr}_{M_\mu}(z)$. The KL divergence is positive semidefinite and vanishes if and only if pr_{M_T} is equivalent in the sense of distributions to pr_{M_μ} . Because the first term in the divergence depends only on the unknown true distribution and not on the candidate model, minimizing the KL divergence with respect to the model is equivalent to maximizing the quantity $E_z[\log \text{pr}_{M_\mu}(z)]$.

The KL divergence can be used as the starting point for a number of standard methods related to modeling data. For example, consider the usual case in which a parameter-dependent version of the model probability density is $\text{pr}_{M_\mu}(z) = \text{pr}(z|\mathbf{a}, M_\mu)$, i.e., our model probability distributions depend on additional parameters \mathbf{a} . Determination of the best-fit parameters \mathbf{a}^* for a given model M_μ can be viewed as an optimization problem over $\text{pr}(z|\mathbf{a}, M_\mu)$ such that $E_z[\log \text{pr}(z|\mathbf{a}, M_\mu)]$ is maximized, so that $\text{KL}(M_\mu)$ is minimized. In practice, the true distribution is inaccessible and estimators using a finite sample $\{y\}$ must be used instead. A common practice is to estimate

$E_z[\log \text{pr}(z|\mathbf{a}, M_\mu)]$ by the standardized out-of-sample log likelihood function:

$$\begin{aligned} E_z[\log \text{pr}(z|\mathbf{a}, M_\mu)] &\simeq \frac{1}{N} \sum_i \log \text{pr}(y_i|\mathbf{a}, M_\mu) \\ &= \frac{1}{N} \log \text{pr}(\{y\}|\mathbf{a}, M_\mu), \end{aligned} \quad (20)$$

where as introduced in Sec. I A, the symbol \simeq indicates that the right-hand side is an unbiased sample estimator of the quantity on the left. Directly maximizing this likelihood function gives the quantity $\mathbf{a}_{\text{MLE}}^*$, known as the maximum likelihood estimator (MLE). The MLE is commonly used in the frequentist literature. On the other hand, in Bayesian modeling the distribution of the parameters is inferred directly by applying Bayes’s theorem to obtain the posterior (i.e., the likelihood weighted by the prior):

$$\log \text{pr}(\mathbf{a}|\{y\}, M_\mu) \propto \frac{1}{N} \sum_i \log [\text{pr}(y_i|\mathbf{a}, M_\mu) \text{pr}(\mathbf{a}|M_\mu)^{1/N}], \quad (21)$$

where the $1/N$ exponent on the prior distribution ensures that this summed version is equivalent to the conventional posterior estimate defined with respect to the full dataset, $(1/N) \log [\text{pr}(\{y\}|\mathbf{a}, M_\mu) \text{pr}(\mathbf{a}|M_\mu)]$. Maximizing the posterior probability over \mathbf{a} gives the posterior mode (PM), \mathbf{a}_{PM}^* .

For the case of model averaging or model selection, rather than a single model distribution, we would like to compare a set of models $\{M_\mu\}$, identifying $\text{pr}_{M_\mu}(z) = \text{pr}(z|M_\mu)$ for each model M_μ in the set. The model weights can then be related directly to the probability density in the KL divergence using Bayes’s theorem,

$$\text{pr}(M_\mu|z) \propto \text{pr}(z|M_\mu). \quad (22)$$

Note that there is no explicit reference to the model parameters \mathbf{a} here. This observation is crucial to a more general treatment of model weights and model averaging. To restate this important idea in other words: in the context of the KL divergence, the model weights are determined by each model’s predicted distribution over the data $\text{pr}(z|M_\mu)$. Since pr_{M_T} is clearly independent of the parameters \mathbf{a} , whatever expression represents the candidate model in the KL divergence must also be independent of \mathbf{a} . From this perspective, we are completely free to specify a prescription for dealing with the model parameters $\{\mathbf{a}\}$. We may view each possible prescription as a variation on the standard definition of the KL divergence. These variations in turn may be used to directly define new information criteria.

In the discussion of bias for information criteria to follow, it will be necessary to consider two matrices defined from the log-likelihood: the Fisher information matrix I_z

and the negative Hessian matrix J_z , which are defined for a given model as

$$(I_z)_{ab}(\mathbf{a}) \equiv E_z \left[\frac{\partial \log L(z; \mathbf{a})}{\partial \mathbf{a}_a} \frac{\partial \log L(z; \mathbf{a})}{\partial \mathbf{a}_b} \right], \quad (23)$$

$$(J_z)_{ab}(\mathbf{a}) \equiv -E_z \left[\frac{\partial^2 \log L(z; \mathbf{a})}{\partial \mathbf{a}_a \partial \mathbf{a}_b} \right], \quad (24)$$

where $L(z; \mathbf{a})$ denotes the asymptotic likelihood function—the left-hand side of either Eq. (20) or Eq. (21), depending on whether it is being evaluated in a frequentist or Bayesian context, respectively (in the latter case, this is the posterior probability function.) Given a finite sample of size N , unbiased estimators for these two matrices are given by

$$(I_N)_{ab}(\mathbf{a}) \equiv \frac{1}{N-1} \sum_{i=1}^N \left[\frac{\partial \log L(x_i; \mathbf{a})}{\partial \mathbf{a}_a} \frac{\partial \log L(x_i; \mathbf{a})}{\partial \mathbf{a}_b} \right], \quad (25)$$

$$(J_N)_{ab}(\mathbf{a}) \equiv -\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \log L(x_i; \mathbf{a})}{\partial \mathbf{a}_a \partial \mathbf{a}_b}, \quad (26)$$

where now $L(x_i; \mathbf{a})$ corresponds to the right-hand side of either Eq. (20) (frequentist) or Eq. (21) (Bayesian). We note in passing that more general definitions exist for the negative Hessian and the Fisher information matrices [30].

Finally, we will frequently assume that the parameter prior information does not grow too rapidly with increasing N , i.e.,

$$\lim_{N \rightarrow \infty} N^{-1} \log \text{pr}(\mathbf{a} | M_\mu) = 0. \quad (27)$$

So long as this condition holds (e.g., the prior information does not depend on N), the PM and MLE become identical as $N \rightarrow \infty$ and the influence of the prior term vanishes relative to the likelihood. This assumption is common in the statistical analysis literature (see [12,13,31,32] among others) and holds in the typical case where priors are independent of the data. In making this assumption, the distinction between the Bayesian and frequentist cases vanishes in the large- N limit. For a more detailed discussion on this assumption, see Appendix 2 of [32].

A. Plug-in KL divergence

We turn now to the problem of how to deal with model parameters in the estimation of $E_z[\log \text{pr}_M(z)]$ in Eq. (19). A simple approach to dealing with model parameter dependence is to determine a “best-fit” value \mathbf{a}^* and plug in this estimate to construct a predictive density function $\text{pr}(z | \mathbf{a}^*, M_\mu)$, which no longer depends on \mathbf{a} . This leads to the plug-in KL divergence:

$$\text{KL}_{\text{plug-in}}(M_\mu) \equiv E_z[\log \text{pr}_{M_T}(z)] - E_z[\log \text{pr}(z | \mathbf{a}^*, M_\mu)]. \quad (28)$$

The exact definition of this estimator depends on the choice of best-fit estimator \mathbf{a}^* .

In the frequentist literature, it is common to use the maximum likelihood estimator $\mathbf{a}_{\text{MLE}}^*$ as the plug-in estimator. In this case, $N^{-1} \log \text{pr}(\{y\} | \mathbf{a}_{\text{MLE}}^*, M_\mu)$ is an asymptotically biased estimator of $E_z[\log \text{pr}(z | \mathbf{a}_{\text{MLE}}^*, M_\mu)]$, as discussed in [30,33–35]. Given a finite sample, we may construct an estimator b_N for the asymptotic bias b_z , which was done in [33]:

$$b_N^{\text{plug-in}} = \frac{1}{N} \text{tr}[J_N^{-1}(\mathbf{a}_{\text{MLE}}^*) I_N(\mathbf{a}_{\text{MLE}}^*)], \quad (29)$$

where I_N and J_N are the sample estimates of the (frequentist) log-likelihood Fisher information and negative Hessian matrices, as defined in Eq. (25) and (26) above. Subtracting $b_N^{\text{plug-in}}$ gives us an asymptotically unbiased estimator:

$$\begin{aligned} E_z[\log \text{pr}(z | \mathbf{a}_{\text{MLE}}^*, M_\mu)] &\simeq \frac{1}{N} \sum_i \log \text{pr}(y_i | \mathbf{a}_{\text{MLE}}^*, M_\mu) \\ &\quad - \frac{1}{N} \text{tr}[J_N^{-1}(\mathbf{a}_{\text{MLE}}^*) I_N(\mathbf{a}_{\text{MLE}}^*)]. \end{aligned} \quad (30)$$

Multiplying by a conventional factor of $-2N$ [see Eq. (10) and Eq. (20)] then gives the Takeuchi information criterion (TIC):

$$\begin{aligned} \text{TIC}_\mu &= -2 \log \text{pr}(\{y\} | \mathbf{a}_{\text{MLE}}^*, M_\mu) \\ &\quad + 2 \text{tr}[J_N^{-1}(\mathbf{a}_{\text{MLE}}^*) I_N(\mathbf{a}_{\text{MLE}}^*)]. \end{aligned} \quad (31)$$

We emphasize that this, and other information criteria to be introduced, may be viewed as formulas for the model weight $\text{pr}(M_\mu | \{y\})$ by way of Eq. (11). To be explicit, the model-averaging version of the TIC is

$$\begin{aligned} \text{TIC}_{\text{MA},\mu} &= -2 \log \text{pr}(M_\mu) - 2 \log \text{pr}(\{y\} | \mathbf{a}_{\text{MLE}}^*, M_\mu) \\ &\quad + 2 \text{tr}[J_N^{-1}(\mathbf{a}_{\text{MLE}}^*) I_N(\mathbf{a}_{\text{MLE}}^*)], \end{aligned} \quad (32)$$

with an implied (unnormalized) model weight of $\text{pr}(M_\mu | \{y\}) = \exp(-\text{TIC}_{\text{MA},\mu}/2)$. Models that minimize the TIC will be favored as they minimize the KL divergence; this will be true for all of the information criteria discussed.

If we assume further that the true distribution belongs to the family of candidate distributions, then we may make the replacement $\text{tr}[J_N^{-1}(\mathbf{a}_{\text{MLE}}^*) I_N(\mathbf{a}_{\text{MLE}}^*)] \rightarrow k$, where k is the number of parameters (i.e., the dimension of the parameter vector \mathbf{a}). This replacement follows from the equivalence of the asymptotic Fisher matrix $I(\mathbf{a})$ and the Hessian matrix $J(\mathbf{a})$, as proven in [26,36,37] among others, so that the trace is over the $k \times k$ identity matrix. With this replacement, the TIC reduces to the Akaike information criterion (AIC) [9–11]:

$$\text{AIC}_\mu = -2 \log \text{pr}(\{y\} | \mathbf{a}_{\text{MLE}}^*, M_\mu) + 2k. \quad (33)$$

We emphasize that the AIC and TIC are frequentist information criteria and thus make no reference to the prior distribution. If we are interested in Bayesian applications, we must modify the derivation above to reflect this. It is shown in [12] that plug-in usage of the posterior mode and removal of asymptotic bias leads to the Bayesian TIC (BTIC):

$$\text{BTIC}_\mu = -2 \log \text{pr}(\{y\} | \mathbf{a}_{\text{PM}}^*, M_\mu) + 2 \text{tr}[J_N^{-1}(\mathbf{a}_{\text{PM}}^*) I_N(\mathbf{a}_{\text{PM}}^*)], \quad (34)$$

where \mathbf{a}_{PM}^* is now the posterior mode. In Eq. (34), the log-likelihood Fisher information and negative Hessian matrices are defined using the Bayesian form of Eq. (25) and (26); henceforth, we will always use the Bayesian form of I_N and J_N unless otherwise stated.

With the further assumption that the candidate models contain the true distribution, we may again make the replacement $\text{tr}[J_N^{-1}(\mathbf{a}_{\text{PM}}^*) I_N(\mathbf{a}_{\text{PM}}^*)] \rightarrow k$, recovering a direct Bayesian analog of the AIC, which we dub the Bayesian AIC:

$$\text{BAIC}_\mu = -2 \log \text{pr}(\{y\} | \mathbf{a}_{\text{PM}}^*, M_\mu) + 2k. \quad (35)$$

As far as we know, the abbreviation ‘‘BAIC’’ is so far unused in the statistics literature. The BAIC is not to be confused with ‘‘a Bayesian information criterion’’ (ABIC) (often referred to as ‘‘Akaike’s Bayesian information criterion’’ [38]), which can be derived from the KL divergence by marginalizing over the parameter space [39,40], or with Schwarz’s ‘‘Bayesian information criterion’’ (BIC) [41], which also has connections to the marginalized KL divergence. See Sec. III A 1 and Appendix A for further discussion of the marginalized KL divergence and associated information criteria.

Although the BTIC and BAIC are appropriate for use in Bayesian inference, we note that the use of a plug-in estimator implies the existence of a fixed underlying set of model parameters, which is more inline with the frequentist approach to inference. A more natural Bayesian approach would consider model probability distributions rather than fixed values; this will be the case for the subsequent information criteria.

Unless otherwise stated, we denote the posterior mode \mathbf{a}_{PM}^* as \mathbf{a}^* omitting the subscript from here forward.

1. Digression: Marginalized KL divergence

Another AIC-like information criterion for Bayesian model averaging is proposed in [26]. This information criterion is derived from the marginalized KL divergence:

$$\text{KL}_{\text{marg}}(M_\mu) \equiv E_z[\log \text{pr}_{M_T}(z)] - E_z \left[\log \int d\mathbf{a} \text{pr}(z | \mathbf{a}, M_\mu) \text{pr}(\mathbf{a} | M_\mu) \right] \quad (36)$$

$$= E_z[\log \text{pr}_{M_T}(z)] - E_z[\log E_{\mathbf{a}}[\text{pr}(z | \mathbf{a}, M_\mu)]]. \quad (37)$$

The expectation value over the parameters with respect to the prior probability distribution is

$$E_{\mathbf{a}}[\dots] \equiv \int d\mathbf{a} \text{pr}(\mathbf{a} | M_\mu) (\dots), \quad (38)$$

where we have assumed that the prior distribution $\text{pr}(\mathbf{a} | M_\mu)$ is normalized.³ Written in this form, it is apparent that the marginalized KL divergence has a strong dependence on the prior distribution. This is manifestly evident by comparison to Eq. (19), where the use of KL_{marg} is equivalent to the identification

$$\text{pr}_{M_\mu}(z) \sim E_{\mathbf{a}}[\text{pr}(z | \mathbf{a}, M_\mu)], \quad (39)$$

which makes no reference at all to the data sample $\{y\}$, only to the prior parameter distribution.

Comparing two models using KL_{marg} is equivalent to evaluating which model (together with its *prior* parameter distribution) is more effective in describing the observed data. This may be desirable in specific contexts, but attempting to use the marginalized KL divergence in more typical cases where the posterior parameter values of interest can lead to counterintuitive effects such as the Jeffreys-Lindley paradox in which the results under certain choices of prior become fully independent of the data (see Appendix A for further discussion).

By approximating the integral in Eq. (36) to leading order in large N and appealing to the use of a cross-validation method to set the priors [26], one can obtain from KL_{marg} what we will refer to as ‘‘ABIC_{CV},’’ a variation of Akaike’s Bayesian information criterion:

$$\text{ABIC}_{\text{CV},\mu} = -2 \log [\text{pr}(\{y\} | \mathbf{a}^*, M_\mu) \text{pr}(\mathbf{a}^* | M_\mu)] + 2k. \quad (40)$$

The ABIC_{CV}, which is just called ‘‘AIC’’ in [26], is identical to the BAIC except for the use of the posterior rather than the likelihood (for emphasis, both use the posterior mode \mathbf{a}_{PM}^* for the plug-in estimator \mathbf{a}^*). While the ABIC_{CV} formula is asymptotically equivalent to the BAIC when Eq. (27) holds (e.g., the prior is N independent), the use of cross-validation requires the priors to be adjusted as more data is accumulated, giving a prior that depends too strongly on N . The full ABIC (without the use of cross-validation) has not been shown to be asymptotically

³In the case of improper priors, the integral in Eq. (36) would be the same but would not be interpreted as an ‘‘expectation value.’’

unbiased, and in fact appears to differ by $O(\log N)$ terms from the (asymptotically unbiased) BAIC at large N . Due to concerns regarding its asymptotic bias, we do not study the marginalized KL divergence further here. Appendix A contains some further discussion of the marginalized KL divergence and the connection to the ABIC_{CV} from [26].

B. Posterior averaged KL divergence

Though adaptations like the BTIC and BAIC exist, the plug-in prescription is an inherently frequentist approach as it considers the underlying model parameters fixed. In Bayesian inference, parameter estimates are given as probability distributions. In light of this distinction, it is natural to consider averaging over the posterior distribution to measure deviations from model truth. This prescription gives the posterior averaged KL divergence:

$$\text{KL}_{\text{post-avg}}(M_\mu) \equiv E_z[\log \text{pr}_{M_\tau}(z)] - E_z[E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_\mu)]], \quad (41)$$

where the expectation value over parameters with respect to the posterior distribution is

$$E_{\mathbf{a}|\{y\}}[\dots] \equiv \frac{\int d\mathbf{a} \text{pr}(\mathbf{a}|\{y\}, M_\mu)(\dots)}{\int d\mathbf{a} \text{pr}(\mathbf{a}|\{y\}, M_\mu)} = \frac{\int d\mathbf{a} \text{pr}(\{y\}|\mathbf{a}, M_\mu)\text{pr}(\mathbf{a}|M_\mu)(\dots)}{\int d\mathbf{a} \text{pr}(\{y\}|\mathbf{a}, M_\mu)\text{pr}(\mathbf{a}|M_\mu)}. \quad (42)$$

With a trivial rewriting of Eq. (41) as

$$\text{KL}_{\text{post-avg}}(M_\mu) \equiv E_z[\log \text{pr}_{M_\tau}(z)] - E_z[\log \exp(E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_\mu)])], \quad (43)$$

we identify $\exp(E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_\mu)])$ as the relevant predictive distribution estimating $\text{pr}_{M_\mu}(z)$; there is no common name associated with this distribution. This rearrangement shows that in the sense of predictive distributions, the posterior averaged KL divergence is somewhat less natural compared to the posterior predictive KL divergence defined in Sec. III B below. We note in passing that unlike the predictive distributions associated with the plug-in or posterior predictive KL divergences, this predictive distribution is not obviously properly normalized. This does not have any obvious impact on the derivations to follow, but it may be interesting to explore the normalization of the predictive distribution in future work.

As above, to convert this to a useful information criterion we must approximate the second term in $\text{KL}_{\text{post-avg}}(M_\mu)$ at finite sample size. One way to do so is to replace the expectation over z by using a sum over the sample data, which in turn will require a bias correction term similar to

the BTIC. This approach, which was proposed by Zhou [12,13], gives the posterior averaging information criterion (PAIC):

$$\text{PAIC}_\mu = -2E_{\mathbf{a}|\{y\}}[\log \text{pr}(\{y\}|\mathbf{a}, M_\mu)] + 2\text{tr}[J_N^{-1}(\mathbf{a}^*)I_N(\mathbf{a}^*)]. \quad (44)$$

Evaluation of the PAIC requires carrying out a full integration of the posterior-weighted likelihood over the parameter space to evaluate $E_{\mathbf{a}|\{y\}}$, which may be difficult or impractical. Historically, alternative ways of estimating $E_z[E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_\mu)]]$ appeared in the literature well before the PAIC. This was first attempted in [31] where the deviance information criterion (DIC) was proposed:

$$\text{DIC} = -2\log \text{pr}(\{y\}|E_{\mathbf{a}|\{y\}}[\mathbf{a}], M_\mu) + 2p_D, \quad (45)$$

where

$$p_D = -2E_{\mathbf{a}|\{y\}}[\log \text{pr}(\{y\}|\mathbf{a}, M_\mu)] + 2\log \text{pr}(\{y\}|E_{\mathbf{a}|\{y\}}[\mathbf{a}], M_\mu). \quad (46)$$

This DIC is defined by analogy to the BAIC where the posterior mean $E_{\mathbf{a}|\{y\}}[\mathbf{a}]$ is an alternative parameter plug-in to the posterior mode \mathbf{a}^* , and p_D is interpreted as an effective number of parameters. $E_z[E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_\mu)]]$ arises implicitly in the DIC through p_D .

Note that like the BAIC, the DIC is defined to estimate $\text{KL}_{\text{plug-in}}$ rather than $\text{KL}_{\text{post-avg}}$. It is only in correcting for asymptotic bias that we see estimates of $E_z[E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_\mu)]]$ appear. This is what inspired studies of $\text{KL}_{\text{post-avg}}$ and supports the idea that all of the variants of KL divergence discussed here are equivalent in some sense.

The DIC has since been criticized for its heuristic derivation and tendency to overfit observed data as it underpenalizes overly complex models [42], and thus will not be discussed further here (although a more detailed exploration of the DIC compared to the other ICs defined here could be an interesting direction for future study). A more rigorous alternative to the DIC was studied in [32] where the Bayesian predictive information criterion was introduced:

$$\text{BPIC}_\mu = -2\log(\text{pr}(\{y\}|\mathbf{a}^*, M_\mu)\text{pr}(\mathbf{a}^*|M_\mu)) + 2E_{\mathbf{a}|\{y\}}[\log \text{pr}(\mathbf{a}|M_\mu)] + 2\text{tr}[J_N^{-1}(\mathbf{a}^*)I_N(\mathbf{a}^*)] + k. \quad (47)$$

The BPIC has also been studied in the context of Bayesian model averaging [43].

Explicitly, the BPIC trades the integration over the full posterior distribution for an integration over the prior distribution $\text{pr}(\mathbf{a}|M_\mu)$. This is accomplished by including a plug-in estimator in the asymptotic bias correction, which

cancels the integral of $\log \text{pr}(\{y\}|\mathbf{a}, M_\mu)$ from $\text{KL}_{\text{post-avg}}$. In other words, the BPIC is defined as

$$\text{BPIC}_\mu = -2E_{\mathbf{a}|\{y\}}[\log \text{pr}(\{y\}|\mathbf{a}, M_\mu)] + 2Nb_N^{\text{BPIC}}, \quad (48)$$

where the asymptotic bias b_z^{BPIC} is estimated by

$$\begin{aligned} Nb_z^{\text{BPIC}} &\simeq Nb_N^{\text{BPIC}} = -\log(\text{pr}(\{y\}|\mathbf{a}^*, M_\mu)\text{pr}(\mathbf{a}^*|M_\mu)) \\ &\quad + E_{\mathbf{a}|\{y\}}[\log(\text{pr}(\{y\}|\mathbf{a}, M_\mu)\text{pr}(\mathbf{a}|M_\mu))] \\ &\quad + \text{tr}[J_N^{-1}(\mathbf{a}^*)I_N(\mathbf{a}^*)] + \frac{1}{2}k, \end{aligned} \quad (49)$$

in contrast to the PAIC, which is defined as

$$\text{PAIC}_\mu = -2E_{\mathbf{a}|\{y\}}[\log \text{pr}(\{y\}|\mathbf{a}, M_\mu)] + 2Nb_N^{\text{PAIC}}, \quad (50)$$

$$Nb_z^{\text{PAIC}} \simeq Nb_N^{\text{PAIC}} = \text{tr}[J_N^{-1}(\mathbf{a}^*)I_N(\mathbf{a}^*)]. \quad (51)$$

For emphasis, the PAIC and the BPIC are both estimators of Eq. (41) and differ only by subleading terms in their bias corrections, a difference that vanishes in the $N \rightarrow \infty$ limit.

The BPIC is easier to evaluate in many situations; we will find below that for the least-squares case, when using approximate expressions for the integrals, the BPIC is much more accurate than the PAIC for smaller sample sizes. On the other hand, the PAIC does have certain advantages. Specifically, by using a plug-in estimator in its asymptotic bias correction, the BPIC loses estimation efficiency compared to the PAIC when the posterior is asymmetric or when there is nonzero correlation between parameters; furthermore, the BPIC is not well defined when the prior distribution is degenerate. For more detail on these cases, see [12,13].

As above, under the usual assumption of correct model specification, we may replace the trace in the BPIC and PAIC bias correction terms with the number of parameters k ; we do not give these variations separate names. As a brief aside (assuming correct model specification for simplicity), we see from Eq. (47) that the BPIC includes a $3k$ term in contrast to the BAIC's $2k$ term; as shown in Appendix D, evaluating the posterior average in Eq. (44) gives rise to an additional k totaling $3k$ for the PAIC as well. The BPIC and PAIC are still asymptotically equivalent to the BAIC, where throughout the paper we use the term ‘‘asymptotically equivalent’’ when referring to information criteria to mean equivalence in the context of model choice; see Appendix B for further discussion. For the sake of concreteness, we hold off on further discussion of the $3k$ term until Sec. IV D.

C. Posterior predictive KL divergence

As a final variation on construction of the KL divergence, we may observe that the way in which the posterior average was constructed in Eq. (41) is not unique. Specifically, the second expectation value can be moved

inside the logarithm,⁴ defining the posterior predictive KL divergence:

$$\begin{aligned} \text{KL}_{\text{post-pred}}(M_\mu) &\equiv E_z[\log \text{pr}_{M_T}(z)] \\ &\quad - E_z[\log E_{\mathbf{a}|\{y\}}[\text{pr}(z|\mathbf{a}, M_\mu)]]. \end{aligned} \quad (52)$$

The name ‘‘posterior predictive’’ follows from the observation that we may rewrite

$$\begin{aligned} E_{\mathbf{a}|\{y\}}[\text{pr}(z|\mathbf{a}, M_\mu)] \\ \propto \int d\mathbf{a} \text{pr}(z|\mathbf{a}, M_\mu)\text{pr}(\{y\}|\mathbf{a}, M_\mu)\text{pr}(\mathbf{a}|M_\mu) \end{aligned} \quad (53)$$

$$\propto \int d\mathbf{a} \text{pr}(z|\mathbf{a}, M_\mu)\text{pr}(\mathbf{a}|\{y\}, M_\mu) \quad (54)$$

$$\equiv \text{pr}(z|\{y\}, M_\mu), \quad (55)$$

which is the predictive distribution for future observation z obtained by averaging the model parameters over the posterior distribution. Though we will see that the commutation of the expectation and the log will add some computational complexity in practice, the use of the posterior predictive distribution as an estimator of $\text{pr}_{M_T}(z)$ makes $\text{KL}_{\text{post-pred}}$ somewhat more natural in Bayesian inference than $\text{KL}_{\text{post-avg}}$ [cf. Eq. (43)]. A less heuristic motivation for $\text{KL}_{\text{post-pred}}$ over $\text{KL}_{\text{post-avg}}$ follows from Jensen's inequality [44,45]. Specifically, we have that for a general expectation operator $E[\dots]$ and random variable X

$$E[\log X] \leq \log E[X]. \quad (56)$$

Therefore,

$$\text{KL}_{\text{post-pred}} \leq \text{KL}_{\text{post-avg}}. \quad (57)$$

Since the KL divergence is positive semidefinite, this in turn implies that $\text{KL}_{\text{post-pred}}$ will be closer to zero. In other words, minimizing the posterior averaged KL divergence with respect to the set of models and parameter values can never do better than minimizing the posterior predictive KL divergence. It is meaningful to compare these two KL divergences by framing them both in terms of the nonparametric KL divergence Eq. (19). The inequality

⁴We note in passing that a similar rearrangement may be done to the marginalized KL divergence, defining $\text{KL}'_{\text{marg}}(M_\mu) \equiv E_z[\log \text{pr}_{M_T}(z)] - E_z[E_{\mathbf{a}}[\log \text{pr}(z|\mathbf{a}, M_\mu)]]$. As far as we know the resulting IC from this definition has not been studied in the literature, but it is not obvious that it has any advantage compared to the other ICs discussed so far, and it is likely to suffer from the same difficulties as the marginalized KL divergence. Based on the argument using Jensen's inequality given in the text below, this IC would also perform worse than KL_{marg} by transposing the logarithm in this way.

above then implies that in terms of closeness to the true distribution $\text{pr}_T(z)$, using the posterior predictive $\log E_{\mathbf{a}|\{y\}}[\text{pr}(z|\mathbf{a}, M_\mu)]$ as our choice for the nonparametric $\text{pr}_{M_\mu}(z)$ will never underperform the posterior average $E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_\mu)]$ for a given choice of model M_μ .

One bias-corrected information criterion corresponding to $\text{KL}_{\text{post-pred}}$ is the posterior predictive information criterion:

$$\text{PPIC}_\mu = -2 \sum_{i=1}^N \log \text{pr}(y_i|\{y\}, M_\mu) + 2\text{tr}[J_N^{-1}(\mathbf{a}^*)I_N(\mathbf{a}^*)]. \quad (58)$$

The PPIC was proposed in [12] as an *ad hoc* information criterion based on certain formulations of Bayes's factors. However, the first term of the PPIC appears earlier in the literature; the first instance we are aware of is [30]. Other information criteria which include the same first term include the predictive information criterion [38,40] and the Watanabe-Akaike information criteria (WAIC) [46]. These information criteria differ from the PPIC in their bias corrections: the PIC lacks a simple general definition for the bias term, and the WAIC includes additional posterior averages, resulting in a higher complexity. We will focus on the PPIC here as its bias correction is of the same form as the other information criteria discussed. We show in Appendix B that the PPIC is asymptotically equivalent to the BAIC.

Although the modification of the KL divergence to obtain the PPIC rather than the PAIC seems relatively minor, we will find in practice that the PPIC is uniquely sensitive to information encoded in the individual fluctuations within the sample $\{y\}$, and as such can be particularly effective for certain problems.

We note in passing that some information criteria in the literature can be derived using a combination of the various KL divergence formulations discussed here (e.g., the WAIC can be written as $2\text{PAIC} - \text{PPIC} - 2\text{tr}[J_N^{-1}(\mathbf{a}^*)I_N(\mathbf{a}^*)]$). It is unclear to us whether doing so has any theoretical or practical motivations, hence we ignore these alternatives and present only the more natural information criteria defined above.

IV. SPECIALIZATION TO LEAST-SQUARES REGRESSION

In this section, we specialize our discussion of Bayesian model averaging, the KL divergence, and information criteria to least-squares regression, which is of primary interest in the context of lattice simulations. We start with a brief overview of least-squares fitting and the relevant notation. The BAIC is discussed as a reformulation of the AIC-like information criterion proposed in [26]. We then discuss an asymptotic integral approximation known as Laplace's method that will be needed in the subsequent sections. Next, we return to some of the aforementioned

information criteria (BPIC, PAIC, and PPIC) and give approximations for each in the case of least-squares fitting. Lastly, we discuss improvements to the information criteria approximations.

A. Least-squares fitting

The discussion thus far has been completely general with regards to the probability distributions appearing in the KL divergence, information criteria, and model-averaging formulas. We now specialize our discussion to the case of least-squares regression of a model M_μ with parameters \mathbf{a} to a set of data $\{y\}$. The likelihood function is

$$\text{pr}(\{y\}|\mathbf{a}, M_\mu) = \prod_{i=1}^N \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \exp\left[-\frac{1}{2}\chi_i^2\right], \quad (59)$$

where

$$\chi_i^2 \equiv (y_i - f_\mu(\mathbf{a}))^T \Sigma^{-1} (y_i - f_\mu(\mathbf{a})) \quad (60)$$

is the standard chi-squared goodness of fit statistic, which involves the data sample y_i , the model function $f_\mu(\mathbf{a})$ corresponding to the model M_μ , and the covariance matrix between the individual samples $\Sigma = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})(y_i - \bar{y})^T$; we assume the samples are drawn independently from some underlying distribution. The dimension of a single observation vector y_i is denoted by d , and the number of independent observations drawn from the true distribution is N .

As for the prior distribution, a common choice is to use a multivariate Gaussian [47,48],

$$\text{pr}(\mathbf{a}|M_\mu) = \frac{1}{(2\pi)^{k/2}(\det \tilde{\Sigma})^{1/2}} \times \exp\left[-\frac{1}{2}(\mathbf{a} - \tilde{\mathbf{a}})^T \tilde{\Sigma}^{-1} (\mathbf{a} - \tilde{\mathbf{a}})\right], \quad (61)$$

where k is the number of fit parameters in model M_μ , $\tilde{\Sigma}$ is the prior covariance matrix, and $\tilde{\mathbf{a}}$ is the prior central value. We define the ‘‘prior chi-squared statistic’’

$$\tilde{\chi}^2 \equiv -2 \log \text{pr}(\mathbf{a}|M_\mu) \quad (62)$$

for later use. In the multivariate Gaussian case, $\tilde{\chi}^2 \equiv (\mathbf{a} - \tilde{\mathbf{a}})^T \tilde{\Sigma}^{-1} (\mathbf{a} - \tilde{\mathbf{a}}) + (\text{const})$, but the approximate formulas derived below apply in general. Unless otherwise stated, we will assume that Eq. (27) holds, i.e., that the prior information grows sufficiently slowly with the sample size.

Since we are only considering the case of a fixed data set, the overall normalization of the likelihood function $\text{pr}(\{y\}|\mathbf{a}, M_\mu)$ will be the same for all models and can be ignored.⁵ On the other hand, in the presence of models with

⁵The problem of data subset selection is treated as a model variation problem, so that this normalization factor remains irrelevant; see Sec. V. See also further discussion of this issue in [49].

varying numbers of parameters the normalization of prior distribution $\text{pr}(\mathbf{a}|M_\mu)$ may not be omitted.

The best-fit point \mathbf{a}^* is the posterior mode, which maximizes the posterior or, equivalently, minimizes the negative log posterior:

$$\begin{aligned} & -2 \log [\text{pr}(\{y\}|\mathbf{a}, M_\mu) \text{pr}(\mathbf{a}|M_\mu)] - (N-1)d \\ & = \hat{\chi}^2(\mathbf{a}) + \tilde{\chi}^2(\mathbf{a}) \equiv \chi_{\text{aug}}^2(\mathbf{a}), \end{aligned} \quad (63)$$

where

$$\begin{aligned} \hat{\chi}^2(\mathbf{a}) & \equiv \sum_{i=1}^N \chi_i^2(\mathbf{a}) - (N-1)d \\ & = (\bar{y} - f_\mu(\mathbf{a}))^T \hat{\Sigma}^{-1} (\bar{y} - f_\mu(\mathbf{a})), \end{aligned} \quad (64)$$

and $\hat{\Sigma} \equiv \Sigma/N$ is the standard-error covariance matrix. χ_{aug}^2 is the so-called augmented chi-squared function [47]. The $(N-1)d$ term appears when converting between the use of the sample-based $\sum_i \chi_i^2$ and the mean-based $\hat{\chi}^2$, which subtracts it by convention. For data with a constant dimension over all models, the $(N-1)d$ term is constant and thus can be ignored. However, it will play an important role in Sec. V where we consider model averaging over different data subsets, i.e., variable d .

B. BAIC

In the context of least-squares regression, the BAIC takes the form

$$\text{BAIC}_\mu \equiv \hat{\chi}^2(\mathbf{a}^*) + 2k. \quad (65)$$

We reiterate here that unless noted otherwise, throughout this work the plug-in estimator \mathbf{a}^* is the posterior mode \mathbf{a}_{PM}^* . For common applications in lattice simulations with weakly informative priors (so that $\tilde{\chi}^2$ is negligible compared to $\hat{\chi}^2$), the BAIC is nearly identical to the ABIC_{CV} :

$$\text{ABIC}_{\text{CV},\mu} \equiv \chi_{\text{aug}}^2(\mathbf{a}^*) + 2k. \quad (66)$$

This is presented as simply the ‘‘AIC’’ in [26]. While the choice between BAIC and ABIC_{CV} should be inconsequential for most lattice applications when the priors are relatively uninformative, we omit further analysis of the ABIC_{CV} due to its lack of solid theoretical foundation (see Appendix A).

C. Laplace’s method

For evaluation of the subsequent information criteria, we will need integrals of the form

$$\mathcal{I}[\psi] = \int d\mathbf{a} \exp \left[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a}) \right] \psi(\mathbf{a}). \quad (67)$$

In the case of nonlinear least squares, this expression cannot be computed analytically in general. One option is numerical evaluation of the integrals, but this can be relatively expensive as part of a fitting analysis and provides an additional source of numerical instability to deal with. Our focus instead will be on the use of a closed-form approximation known in the asymptotics literature as Laplace’s method. Specifically, we will write a next-to-leading-order (NLO) perturbation expansion in the inverse sample size N^{-1} for Eq. (67), which becomes increasingly accurate as $N \rightarrow \infty$. We have implemented integrals of the form in Eq. (67) numerically and verified the accuracy of our approximation.

The details of this approximation are summarized in Appendix E. The main result is

$$\begin{aligned} \mathcal{I}[\psi] & \approx (2\pi)^{k/2} |\Sigma^*|^{1/2} \exp \left[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a}^*) \right] \\ & \times \left(\psi(\mathbf{a}^*) + \frac{1}{2} H_{ba}(\Sigma^*)_{ab} \right. \\ & - \frac{1}{2} g_d T_{cba}(\Sigma_2^*)_{abcd} - \frac{1}{2} \psi(\mathbf{a}^*) F_{dcba}(\Sigma_2^*)_{abcd} \\ & \left. + \frac{1}{8} \psi(\mathbf{a}^*) T_{fed} T_{cba}(\Sigma_3^*)_{abcdef} \right), \end{aligned} \quad (68)$$

where the inverse parameter covariance matrix is

$$(\Sigma^{*-1})_{ab} = \frac{1}{2} \frac{\partial^2 \chi_{\text{aug}}^2}{\partial a_a \partial a_b} \Big|_{\mathbf{a}=\mathbf{a}^*}, \quad (69)$$

the higher-order contractions of the covariance matrix are

$$\begin{aligned} (\Sigma_2^*)_{abcd} & \equiv 3(\Sigma^*)_{ab}(\Sigma^*)_{cd}, \\ (\Sigma_3^*)_{abcdef} & \equiv 9(\Sigma^*)_{ab}(\Sigma^*)_{cd}(\Sigma^*)_{ef} + 6(\Sigma^*)_{ad}(\Sigma^*)_{be}(\Sigma^*)_{cf}, \end{aligned} \quad (70)$$

and the remaining tensors are given by

$$T_{abc} \equiv \frac{1}{6} \frac{\partial^3 \chi_{\text{aug}}^2}{\partial a_a \partial a_b \partial a_c} \Big|_{\mathbf{a}=\mathbf{a}^*}, \quad F_{abcd} \equiv \frac{1}{24} \frac{\partial^4 \chi_{\text{aug}}^2}{\partial a_a \partial a_b \partial a_c \partial a_d} \Big|_{\mathbf{a}=\mathbf{a}^*}, \quad (71)$$

$$g_a \equiv \frac{\partial \psi}{\partial a_a} \Big|_{\mathbf{a}=\mathbf{a}^*}, \quad H_{ab} \equiv \frac{\partial^2 \psi}{\partial a_a \partial a_b} \Big|_{\mathbf{a}=\mathbf{a}^*}. \quad (72)$$

Note the use of Einstein summation notation for the tensor contractions. This result is in agreement with a special case of a more general integral computed in [50].

In the cases of interest, the integral Eq. (67) will appear with the following normalization:

$$\frac{\mathcal{I}[\psi]}{\mathcal{I}[1]} = \frac{\int d\mathbf{a} \exp \left[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a}) \right] \psi(\mathbf{a})}{\int d\mathbf{a} \exp \left[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a}) \right]}. \quad (73)$$

As shown in Appendix E, normalized integrals of this form can be approximated by first applying Laplace's method Eq. (68) to both the numerator and denominator followed by a geometric expansion. Keeping terms to NLO gives

$$\frac{\mathcal{I}[\psi]}{\mathcal{I}[1]} \approx \psi(\mathbf{a}^*) + \frac{1}{2} H_{ba}(\Sigma^*)_{ab} - \frac{1}{2} g_d T_{cba}(\Sigma^*)_{abcd}. \quad (74)$$

This geometric expansion maintains the same order of accuracy and is used in the probability literature [51].

For the case of linear least squares, the χ^2_{aug} function can be written as a quadratic form in the fit parameter vector \mathbf{a} in which case the tensors T and F are identically zero. Furthermore, if ψ is quadratic in \mathbf{a} so that its higher derivatives vanish, then the ‘‘approximations’’ in Eq. (68) and (74) are in fact exact. This will be the case for linear fit models and the BPIC, but not for the PPIC in which ψ has exponential form. For linear fit models, the PPIC integrals are Gaussian and can be computed exactly, but the expression is unwieldy and since this only works for linear models, we do not pursue it further here.

We emphasize that the rationale for this approximation is based on expansion in the inverse sample size N^{-1} . To verify that the order of approximation is consistent, it is useful to note the N dependence of these tensors: $\Sigma^* = O(N^{-1})$, $\Sigma_2^* = O(N^{-2})$, $\Sigma_3^* = O(N^{-3})$, $T, F = O(N)$, and $g, H = O(\psi(\mathbf{a}^*))$ as $N \rightarrow \infty$. Thus, the approximation in Eq. (68) is accurate to $O(N^{-1}\psi(\mathbf{a}^*))$. We will consider cases where $\psi(\mathbf{a}^*) = O(1)$ and $\psi(\mathbf{a}^*) = O(N)$.

D. BPIC

In Sec. III B, we introduced two other information criteria, the BPIC and PAIC, based on the posterior averaged KL divergence in Eq. (41). Here we specialize our discussion of the BPIC and PAIC to the case of least-squares regression (see Sec. IV A) with correct model specification (so that we may replace $\text{tr}[J^{-1}(\mathbf{a}^*)I(\mathbf{a}^*)] \rightarrow k$). Using the NLO Laplace approximation discussed in Sec. IV C, we give a computationally efficient approximation of the BPIC. We will not pursue the PAIC further in the body of the text due to the lower order of accuracy of the NLO Laplace approximation in this case (see discussion below); the relevant formulas for the PAIC are summarized in Appendix D.

First, we consider the BPIC. In the cases of interest, Eq. (47) reduces to (up to constant terms)

$$\text{BPIC}_\mu = \chi^2_{\text{aug}}(\mathbf{a}^*) - E_{\mathbf{a}|\{y\}}[\tilde{\chi}^2(\mathbf{a})] + 3k, \quad (75)$$

where

$$E_{\mathbf{a}|\{y\}}[\dots] = \frac{\int d\mathbf{a} \exp[-\frac{1}{2}\chi^2_{\text{aug}}(\mathbf{a})](\dots)}{\int d\mathbf{a} \exp[-\frac{1}{2}\chi^2_{\text{aug}}(\mathbf{a})]}. \quad (76)$$

Using the NLO Laplace approximation with the geometric expansion simplification given in Eq. (74), we obtain

$$E_{\mathbf{a}|\{y\}}[\tilde{\chi}^2(\mathbf{a})] \approx \tilde{\chi}^2(\mathbf{a}^*) + \frac{1}{2} \tilde{H}_{ba}(\Sigma^*)_{ab} - \frac{1}{2} \tilde{g}_d T_{cba}(\Sigma^*)_{abcd} \quad (77)$$

where

$$\tilde{g}_a \equiv \left. \frac{\partial \tilde{\chi}^2}{\partial a_a} \right|_{\mathbf{a}=\mathbf{a}^*}, \quad \tilde{H}_{ab} \equiv \left. \frac{\partial^2 \tilde{\chi}^2}{\partial a_a \partial a_b} \right|_{\mathbf{a}=\mathbf{a}^*}. \quad (78)$$

Substituting Eq. (77) into Eq. (75) gives

$$\text{BPIC}_\mu \approx \hat{\chi}^2(\mathbf{a}^*) - \frac{1}{2} \tilde{H}_{ba}(\Sigma^*)_{ab} + \frac{1}{2} \tilde{g}_d T_{cba}(\Sigma^*)_{abcd} + 3k. \quad (79)$$

An interesting feature of the BPIC is the last term in Eq. (47), which gives the $3k$ term in Eq. (79) as opposed to the $2k$ term in the BAIC. The additional k in Eq. (47) comes from the posterior averaging prescription (this is seen explicitly for the PAIC as shown in Appendix D). As a result, the BPIC tends to favor more parsimonious models than the BAIC. While this may seem like an advantage of the BPIC, we emphasize that additional parsimony comes at the cost of larger KL divergence as discussed in Sec. III C. Despite this difference, the BPIC remains asymptotically equivalent to the BAIC in the limit of infinite sample size, as shown in Appendix B.

It is worth discussing a limit in which the other $O(1)$ terms in the BPIC cancel the additional k [in this case, equality holds for Eq. (57)]. Specifically, consider the case of infinitesimal prior widths, i.e., the prior information is infinitely constraining. In this case, Σ^* goes to $\frac{1}{2}\tilde{H}$, and the trace term in Eq. (79) cancels the additional k exactly (of course, there will be additional asymptotic bias from the $\hat{\chi}^2$ unless the prior center value is the true model parameter value). The third term containing \tilde{g} goes to zero, since in this limit $\mathbf{a}^* \rightarrow \tilde{\mathbf{a}}$. In the case of finite prior widths, results will depend on the observed data and the additional k persists. This behavior demonstrates that the additional dependence on the observed information over the prior information manifests itself by favoring parsimonious models more than if there were no posterior averaging, as is the case for the BAIC.

As an aside, consider the PAIC. From Eq. (44), once again assuming correct model specification the PAIC is given by

$$\text{PAIC}_\mu = E_{\mathbf{a}|\{y\}}[\tilde{\chi}^2(\mathbf{a})] + 2k, \quad (80)$$

where $E_{\mathbf{a}|\{y\}}[\dots]$ is defined in Eq. (76). We could proceed by attempting to approximate the expectation as we did in above for the BPIC. However, a complication arises from the fact that $\hat{\chi}^2$ is itself $O(N)$. This means that working to the same order in inverse sample size, $O(N^{-1})$, would require a next-to-next-to-leading order (NNLO) evaluation

of the integral. This endeavor would require a notable increase in mathematical complexity. As discussed in Sec. III B, the PAIC only differs from BPIC in subleading terms in the bias correction. Since the PAIC will be a lower order or more complicated version of the BPIC in practice with similar theoretical motivations, it suffices for our present purposes to omit further discussion of the PAIC. The corresponding results for the PAIC are given in Appendix D; these may be used in cases where the BPIC is not well defined, as discussed in Sec. III B; see also the discussion in [12,13].

E. PPIC

Finally, we turn to the PPIC, defined in Eq. (58). The PPIC involves the posterior predictive distribution, which can be rearranged to the form

$$\text{pr}(y_i|\{y\}, M_\mu) = \int d\mathbf{a} \text{pr}(y_i|\mathbf{a}, M_\mu) \text{pr}(\mathbf{a}|\{y\}, M_\mu) \quad (81)$$

$$= \frac{\int d\mathbf{a} \text{pr}(y_i|\mathbf{a}, M_\mu) \text{pr}(\{y\}|\mathbf{a}, M_\mu) \text{pr}(\mathbf{a}|M_\mu)}{\int d\mathbf{a} \text{pr}(\{y\}|\mathbf{a}, M_\mu) \text{pr}(\mathbf{a}|M_\mu)}. \quad (82)$$

Note that this contains both the combined likelihood for the entire dataset $\text{pr}(\{y\}|\mathbf{a}, M_\mu)$ as well as the likelihood function for the i th single data sample $\text{pr}(y_i|\mathbf{a}, M_\mu)$. In the case of least-squares regression with correct model specification, the PPIC then becomes

$$\text{PPIC}_\mu = -2 \sum_{i=1}^N \log \frac{\int d\mathbf{a} \exp[-\frac{1}{2}\chi_{\text{aug}}^2(\mathbf{a})] \exp[-\frac{1}{2}\chi_i^2(\mathbf{a})]}{\int d\mathbf{a} \exp[-\frac{1}{2}\chi_{\text{aug}}^2(\mathbf{a})]} + 2k. \quad (83)$$

Here, we must compute N integrals, one for each observation. Using Eq. (74), we obtain

$$\text{PPIC}_\mu \approx \hat{\chi}^2(\mathbf{a}^*) + 2k - 2 \sum_{i=1}^N \log \left[1 + \frac{1}{2} \left(\frac{1}{4} (g_i)_b (g_i)_a - \frac{1}{2} (H_i)_{ba} \right) (\Sigma^*)_{ab} + \frac{1}{4} (g_i)_d T_{cba} (\Sigma^*)_{abcd} \right], \quad (84)$$

where

$$(g_i)_a \equiv \left. \frac{\partial \chi_i^2}{\partial a_a} \right|_{\mathbf{a}=\mathbf{a}^*}, \quad (H_i)_{ab} \equiv \left. \frac{\partial^2 \chi_i^2}{\partial a_a \partial a_b} \right|_{\mathbf{a}=\mathbf{a}^*}. \quad (85)$$

We see from Eq. (84) that through the final log term, the PPIC relies on information from each individual observation rather than solely on averaged and prior statistics like the other information criteria discussed. As we will see in Sec. VI B, this sensitivity to sample fluctuations gives the PPIC the ability to parse out models with poor parameter estimates, a very attractive quality in model averaging.

F. Supersymptotics and optimal truncation

In the preceding subsections, approximate expressions for the BPIC and PPIC were obtained using the NLO Laplace approximation. In the limit of large sample size N , this approximation should become increasingly accurate and the size of the subleading terms in N^{-1} should become negligible. However, in practice these information criteria will be computed and used at fixed, finite sample size. For a given value of N , it is possible for the coefficients of our expansion to be such that the subleading terms are larger than the leading terms. For example, there is no reason to suspect the gradients appearing in the NLO Laplace approximations should remain small in cases where the candidate model is unable to accurately represent the data. While such poor models will likely be rejected based on their χ_{aug}^2 values alone, numerical issues can arise in model averaging when subleading terms are dominant. In particular, this effect can cause logarithms with negative real arguments to arise in the PPIC.

As discussed previously, the Laplace approximation discussed in Sec. IV C and derived in Appendix E is an NLO perturbation expansion. This type of fixed-order expansion is known in the asymptotics literature as a Poincaré expansion [52–56]. Another type of expansion is the “supersymptotic” expansion. First proposed by Sir George Gabriel Stokes for a similar integral approximation problem [57], supersymptotics rely on the fact that an asymptotic series need not converge to give an accurate approximation with finitely many terms. Ignoring the case of singular perturbation expansions for simplicity, this means as additional terms are added to the asymptotic series, a formally divergent regular perturbation expansion has a “convergent” part where terms decrease in magnitude algebraically in the perturbation parameter and a “divergent” part which typically grows with additional terms, causing the series to diverge. A supersymptotic expansion is one that is “optimally truncated” after the term of minimum modulus [53–56].

While supersymptotics have been applied to an array of problems (see [58] and references therein), it is often used for the method of steepest descent [53]. Since Laplace’s method is a special case of the method of steepest descent, the use of supersymptotics is well suited for our purposes. In principle, this approach can achieve $O(\exp(-N))$ accurate integral approximations. (Since this error is exponentially small rather than algebraically small, supersymptotics is sometimes referred to as “asymptotics beyond all orders” [59].) While we cannot guarantee this level of accuracy due to other sources of error in the derivation of each criteria (e.g., subleading terms in the bias corrections, see [12,13,32] for details), it does suggest the power of optimal truncation. Furthermore, supersymptotic expansions are known to work well even when the perturbation parameter (N^{-1} here) approaches $O(1)$ [55], which will be the case for small sample sizes. Possible

issues could arise with optimal truncation due to the high-dimensionality of the integrals considered, but this is unlikely assuming that the extremum in χ_{aug}^2 is a simple global minimum [56].

To benefit from these ideas, we propose using the Poincaré expansions previously developed, unless the second term in the NLO Laplace approximation is larger than the first. If the second term is larger, then

superasymptotics suggest that optimal truncation should leave only the leading order term. Note that this prescription does not apply for the case of linear least squares for the BPIC, in which case the NLO “approximations” are exact.

Under the use of optimal truncation, a combined approximate formula for the BPIC becomes (suppressing the tensor indices for compactness)

$$\text{BPIC}_\mu \approx \begin{cases} \hat{\chi}^2(\mathbf{a}^*) - \frac{1}{2}\tilde{H}(\Sigma^*) + \frac{1}{2}\tilde{g}T(\Sigma_2^*) + 3k, & \left| \frac{1}{2}\tilde{H}\Sigma^* - \frac{1}{2}\tilde{g}T\Sigma_2^* \right| < \hat{\chi}^2(\mathbf{a}^*), \\ \hat{\chi}^2(\mathbf{a}^*) + 3k, & \text{otherwise.} \end{cases} \quad (86)$$

In the case of the PPIC, there are N separate integrals for each of the data samples. We can consider optimal truncation case-by-case for each individual integration:

$$\frac{\int d\mathbf{a} \exp[-\frac{1}{2}\chi_{\text{aug}}^2(\mathbf{a}^*)] \exp[-\frac{1}{2}\chi_i^2(\mathbf{a})]}{\int d\mathbf{a} \exp[-\frac{1}{2}\chi_{\text{aug}}^2(\mathbf{a}^*)]} = \begin{cases} \exp[-\frac{1}{2}\chi_i^2(\mathbf{a}^*)](1 + \text{SL}_i), & |\text{SL}_i| < 1, \\ \exp[-\frac{1}{2}\chi_i^2(\mathbf{a}^*)], & \text{otherwise,} \end{cases} \quad (87)$$

where SL_i denotes the subleading terms associated with the i th data sample,

$$\text{SL}_i = \frac{1}{2} \left(\frac{1}{4}(g_i)_b(g_i)_a - \frac{1}{2}(H_i)_{ba} \right) (\Sigma^*)_{ab} + \frac{1}{4}(g_i)_d T_{cba} (\Sigma_2^*)_{abcd}. \quad (88)$$

The form of the PPIC under optimal truncation is therefore

$$\text{PPIC}_\mu \approx \hat{\chi}^2(\mathbf{a}^*) - 2 \sum_{\{i:|\text{SL}_i|<1\}} \log(1 + \text{SL}_i) + 2k. \quad (89)$$

Adopting this prescription thus eliminates the possibility of negative arguments within the logarithms.

There is also a wealth of literature on hyperasymptotics where the divergent part of the asymptotic series is used to obtain orders of accuracy superior to even those of superasymptotics [53–56,60,61]. Since other sources of error would diminish the efficacy of such a procedure here, we advocate for the use of the superasymptotic schema described above rather than develop a hyperasymptotic one. This also maintains relative simplicity in the implementation of the information criteria.

We emphasize that the integral expansions we have carried out above are done primarily for ease of calculation and to reveal useful details about the structure of the various information criteria. The only obstacles, in principle, to direct evaluation of the integral versions of each IC are the lack of a general analytic solution and the computational cost associated with accurate numerical evaluation. However, we find in practice in our numerical tests that the NLO Laplace approximations with optimal truncation yield essentially identical results to direct integration with much lower computational cost. We will discuss this comparison further in Sec. VI B.

V. DATA SUBSET SELECTION

As part of a lattice field theory analysis (or in Bayesian model-averaging applications more broadly), it is often desirable to additionally select a subset of the data beyond which the model is not applied, i.e., selecting d_C dimensions of the samples y_i to be ignored and fitting models only to the other $d_K \equiv d - d_C$ dimensions. The subscript “C” refers to the “cut” portion of the data and the subscript “K” refers to the “kept” portion of the data. A simple and common example of such a procedure in the context of lattice field theory is fitting a two-point correlation function $C(t)$ for the ground-state energy. The full model describing $C(t)$ involves an exponential decay series

$$C(t) = \sum_{m=0}^{\infty} A_m e^{-E_m t}, \quad (90)$$

where $\{E_m\}$ increases monotonically with m . If only the first few states are of interest (as is often the case), it is sufficient to apply the model to times with $t \geq t_{\min}$ for some t_{\min} after which the more rapidly decaying modes become negligible. Choosing t_{\min} has (historically) often been done

manually, although outside of model averaging there have been a variety of methods used for determination of t_{\min} and/or estimation of associated systematic errors, see e.g. [25,62,63].

Though the problem described is one of data subset selection, it can be reformulated as one of model selection. The key to doing so is to define a joint model that describes the full data set. First, choose a subset of the data to which the model M_μ is fit. Next, fit the remaining data to a “perfect” model M_{perf} with zero degrees of freedom (with the use of the approximate formulas we give below, M_{perf} need not be constructed in practice.) An example of such a perfect model is a polynomial of degree $d_C - 1$; in this example, fitting the data for model parameters is equivalent to finding the polynomial interpolant of the data where the differences between the model and sample means vanish identically.

To give a more explicit construction, we first define P as the partition of each observation vector into $y_i = (y_{C,i} \ y_{K,i})^T$, where $y_{K,i} \in \mathbb{R}^{d_K}$ is to be modeled by M_μ and $y_{C,i} \in \mathbb{R}^{d_C}$ is to be modeled by $M_{\text{perf},P}$. We can similarly divide up the inverse sample standard-error covariance matrix as

$$\hat{\Sigma}^{-1} = \begin{pmatrix} (\hat{\Sigma}^{-1})_C & (\hat{\Sigma}^{-1})_O \\ (\hat{\Sigma}^{-1})_O^T & (\hat{\Sigma}^{-1})_K \end{pmatrix}, \quad (91)$$

where the subscript “O” stands for “off-block-diagonal.”

We then define the corresponding partitioned model $\phi_{M,P}(\mathbf{a})$ as

$$(y_i - \phi_{M,P}(\mathbf{a}))_x = \begin{cases} (y_i - \mathbf{a}_C)_x, & (y_i)_x \in y_{C,i}, \\ (y_i - f_M(\mathbf{a}_K))_x, & (y_i)_x \in y_{K,i}. \end{cases} \quad (92)$$

We note for later use that the cut part of the best-fit parameter \mathbf{a}_C^* is simply the mean of the cut data, i.e., $\mathbf{a}_C^* = \bar{y}_C$. Even though \mathbf{a}_C^* are known *a priori*, we cannot take the cut parameter priors to be too constraining as this would violate Eq. (27) and thus not guarantee asymptotic unbiasedness of the information criteria. Therefore, we will take the cut parameter priors to be infinitely diffuse, i.e., $(\hat{\Sigma}_C)^{-1} \rightarrow 0$, which is the limit where predictions rely solely on the data.

Based on these definitions, we can define a partition of the chi-squared function

$$\hat{\chi}^2(\mathbf{a}) = (\bar{y} - \phi_{M,P}(\mathbf{a}))^T \hat{\Sigma}^{-1} (\bar{y} - \phi_{M,P}(\mathbf{a})) \quad (93)$$

$$= \begin{pmatrix} \bar{y}_C - \mathbf{a}_C \\ \bar{y}_K - f_M(\mathbf{a}_K) \end{pmatrix}^T \begin{pmatrix} (\hat{\Sigma}^{-1})_C & (\hat{\Sigma}^{-1})_O \\ (\hat{\Sigma}^{-1})_O^T & (\hat{\Sigma}^{-1})_K \end{pmatrix} \begin{pmatrix} \bar{y}_C - \mathbf{a}_C \\ \bar{y}_K - f_M(\mathbf{a}_K) \end{pmatrix} \quad (94)$$

$$\equiv \hat{\chi}_C^2(\mathbf{a}_C) + \hat{\chi}_K^2(\mathbf{a}_K) + 2\hat{\chi}_O^2(\mathbf{a}_C, \mathbf{a}_K), \quad (95)$$

with analogous definitions for partitions of $\tilde{\chi}^2$ and χ_{aug}^2 .

The ABIC_{CV} (see Appendix A) is derived under this construction for the case of least-squares regression in [26]:

$$\text{ABIC}_{\text{CV},\mu,P} = \chi_{\text{aug},K}^2(\mathbf{a}^*) + 2k + 2d_C, \quad (96)$$

where $\chi_{\text{aug}}^2(\mathbf{a}^*)$ is evaluated only for M_μ , as the contribution from $M_{\text{perf},P}$ vanishes. Note that this result holds even without taking the infinitely diffuse cut prior limit, and without any assumptions on the structure of the correlations between the $y_{C,i}$ and $y_{K,i}$ partitions. The derivation for the BAIC is similar, giving

$$\text{BAIC}_{\mu,P} = \hat{\chi}_K^2(\mathbf{a}^*) + 2k + 2d_C, \quad (97)$$

which should also hold for any cut prior width satisfying Eq. (27). We will rederive this result below.

A subtle point which appears here is the distinction between the sub-blocks of the inverse covariance matrix, e.g., $(\hat{\Sigma}^{-1})_K$, and the inverse of a sub-block, e.g., $(\hat{\Sigma}_K)^{-1}$. The former quantity contains indirect contributions from the cut portion of the data. If we use the kept data exclusively to compute $\text{BAIC}_{\mu,P}$ above, then this is equivalent to making the approximation

$$(\hat{\Sigma}^{-1})_K \approx (\hat{\Sigma}_K)^{-1}, \quad (98)$$

which is commonly used in the lattice community [26]. This approximation can avoid numerical instabilities that may occur when inverting the full $\hat{\Sigma}^{-1}$, particularly when $d_K \ll d$. In fact, this approximation is better than it may seem. Even if $\hat{\Sigma}$ is estimated unbiasedly, simply inverting to find $\hat{\Sigma}^{-1}$ will introduce some finite- N bias (that vanishes asymptotically). The corrected estimator is [64,65]

$$\hat{\Sigma}_{\text{BC}}^{-1} = \frac{N - d - 2}{N - 1} \hat{\Sigma}^{-1} = \frac{N - d_C - d_K - 2}{N - 1} \hat{\Sigma}^{-1}, \quad (99)$$

where the subscript “BC” denotes the bias corrected inverse. The analogous expression for $(\hat{\Sigma}_K)_{\text{BC}}^{-1}$ is

$$(\hat{\Sigma}_K)_{\text{BC}}^{-1} = \frac{N - d_K - 2}{N - 1} (\hat{\Sigma}_K)^{-1}. \quad (100)$$

So, when d_C is large [i.e., when Eq. (98) would be a poor approximation], $(\hat{\Sigma}_K)^{-1}$ will in fact give a less biased result than $(\hat{\Sigma}^{-1})_K$ at finite N .

The distinction between $(\hat{\Sigma}^{-1})_K$ and $(\hat{\Sigma}_K)^{-1}$ will become negligible in the case of weak long-range correlations, that is to say, when the off-block-diagonal elements of the sample covariance $\hat{\Sigma}_O$ are small (in the sense of induced operator norm) relative to the elements of $\hat{\Sigma}_C$ and $\hat{\Sigma}_K$. For the BPIC and PPIC, unlike the BAIC, there will be additional contributions to the “perfect model” IC when the long-range correlations are not negligible. In order to

have tractable approximate formulas for these criteria, we will assume weak long-range correlations so that the correlation matrix is approximately block diagonal, $\Sigma^{-1} \approx \text{diag}(\Sigma_K^{-1}, \Sigma_C^{-1})$. For data where this assumption is badly violated, we advocate the use of the BAIC for subset selection, or one may explicitly construct a piecewise model including the perfect model and perform joint fits to the data as a whole.⁶

One way to derive the perfect model formulas is to explicitly plug in the partitions for $\hat{\chi}^2$, $\tilde{\chi}^2$, and χ_i^2 to the definitions of the PPIC and BPIC and integrate over the perfect-model parameters \mathbf{a}_C . This derivation is shown in Appendix F. Here, we take an alternative and simpler approach, which is to work with the KL divergences directly using a particularly simple choice of the perfect model.

In the following derivations, we use the assumption of negligible long-range correlations described above to separate each definition of the KL divergence as $\text{KL} = \text{KL}_K + \text{KL}_C$; we are able to do this decomposition by Theorem 3.1 from [29] (see discussion in [49]). Explicit calculation of the second term in KL_C as defined by equations Eq. (28), Eq. (41), and Eq. (52) will give us exact results for the associated ICs for the perfect model. These can then be combined with the formulas for the ICs derived above on the kept portion of the data.

For the remainder of this section, unless otherwise noted we focus on a single data set of size d_C and ignore the kept data. We assume a specific perfect model M_{perf} construction of the form $f(x) = \mathbf{a}$ and $\mathbf{a}^* = \bar{y}$, i.e. a model which is defined piecewise for each value in the vector \bar{y} . The number of model parameters is manifestly $k = d_C$.

Given a data sample $\{y\}$ of size N , the predicted least-squares likelihood function for a single future observation z is

$$\text{pr}(z|\mathbf{a}, M_{\text{perf}}) = \frac{1}{(2\pi)^{d_C/2}(\det \Sigma)^{1/2}} \times \exp\left[-\frac{1}{2}(z - \mathbf{a})^T \Sigma^{-1}(z - \mathbf{a})\right], \quad (101)$$

where Σ is the sample covariance matrix. Note that technically, this means the likelihood function should be written as is $\text{pr}(z|\mathbf{a}, M_{\text{perf}}, \{y\})$ since Σ depends on $\{y\}$, although in the large- N limit $\Sigma \rightarrow \Sigma_T$, the true covariance matrix.

Dropping constant factors from the normalization (they will not be constant as d_C is varied, but they will combine

⁶Direct estimation of $\hat{\chi}_0^2$ and its derivatives may be challenging; strongly correlated covariance matrices can have large condition numbers, especially at small sample sizes. Careful treatment of the covariance matrix (e.g., regularization via singular value decomposition) before inversion is essential in this case.

with similar normalization factors from the kept data to become overall constants), then, we have

$$E_z[\log \text{pr}(z|\mathbf{a}^*, M_{\text{perf}})] = -\frac{1}{2}E_z[(z - \bar{y})^T \Sigma^{-1}(z - \bar{y})]. \quad (102)$$

In order to simplify further, suppose that the true model is represented by a vector μ_T , and data y are generated from a multivariate Gaussian random process with true covariance Σ_T . (In general, we could work with a sample estimator of this probability instead so that the central limit theorem applies, leading to the same form.) Then the ‘‘true model’’ probability distribution is

$$\text{pr}_T(z) = \frac{1}{(2\pi)^{d_C/2}(\det \Sigma_T)^{1/2}} \times \exp\left[-\frac{1}{2}(z - \mu_T)^T \Sigma_T^{-1}(z - \mu_T)\right]. \quad (103)$$

Putting these together and using the formula Eq. (E22) derived in Appendix E, we have

$$E_z[\log \text{pr}(z|\mathbf{a}^*, M_{\text{perf}})] = \int dz \text{pr}_T(z) \left(-\frac{1}{2}(z - \bar{y})^T \Sigma^{-1}(z - \bar{y})\right) \quad (104)$$

$$= -\frac{1}{2}\text{tr}[\Sigma^{-1}\Sigma_T] - \frac{1}{2}(\mu_T - \bar{y})^T \Sigma^{-1}(\mu_T - \bar{y}). \quad (105)$$

Considering the second term first, based on a result by White [26,37] (or for this particular model, simply invoking the central limit theorem), the difference $\sqrt{N}(\mu_T - \bar{y})$ is normally distributed as $N \rightarrow \infty$, with mean zero and covariance $C = J^{-1}JJ^{-1}$. Due to the simple structure of this perfect model, we have $J = I = \Sigma_T$, giving the result

$$E_z[\log \text{pr}(z|\mathbf{a}^*, M_{\text{perf}})] = -\frac{1}{2}\text{tr}[\Sigma^{-1}\Sigma_T] - \frac{1}{2N}\text{tr}[\Sigma^{-1}\Sigma_T] \rightarrow -\frac{d_C}{2} - \frac{d_C}{2N}, \quad (106)$$

where we have simplified using the fact that $\Sigma \rightarrow \Sigma_T$, i.e., it is a consistent estimator of the true covariance. In terms of information criteria, this translates to

$$-2NE_z[\log \text{pr}(z|\mathbf{a}^*, M_{\text{perf}})] \simeq \text{BAIC}_{\text{perf}} = \hat{\chi}^2 + 2k + (N - 1)d_C = (N + 1)d_C, \quad (107)$$

where $\hat{\chi}^2 = 0$ identically, the number of parameters $k = d_C$, and we are being careful to keep the overall factor of $(N - 1)d_C$ that appears in the definition of $\hat{\chi}^2$ in terms of sample means, see the discussion around Eq. (63).

Moving on to the second definition of the KL divergence, we have to simplify the posterior average. Using Eq. (42), we have

$$\begin{aligned}
& E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_{\text{perf}})] \\
&= -\frac{1}{2Z} \int d\mathbf{a} \exp\left[-\frac{1}{2}\chi_{\text{aug}}^2(\mathbf{a})\right] \\
&\quad \times (z - \mathbf{a})^T \Sigma^{-1} (z - \mathbf{a}) \quad (108)
\end{aligned}$$

$$\begin{aligned}
&= -\frac{1}{2Z} \int d\mathbf{a} \exp[(\bar{y} - \mathbf{a})^T (\Sigma^*)^{-1} (\bar{y} - \mathbf{a})] \\
&\quad \times (z - \mathbf{a})^T \Sigma^{-1} (z - \mathbf{a}), \quad (109)
\end{aligned}$$

where due to the infinitely diffuse priors $\Sigma^* = \hat{\Sigma} = \Sigma/N$. The normalizing factor Z is

$$Z \equiv \int d\mathbf{a} \exp\left[-\frac{1}{2}\chi_{\text{aug}}^2(\mathbf{a})\right]. \quad (110)$$

Using Eq. (E22) once again, we find the result

$$\begin{aligned}
E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_{\text{perf}})] &= -\frac{1}{2}((z - \bar{y})^T \Sigma^{-1} (z - \bar{y}) \\
&\quad + \text{tr}[\Sigma^{-1} \Sigma^*]). \quad (111)
\end{aligned}$$

The first term is precisely the plug-in log likelihood, while the second term reduces to a constant, $\text{tr}[\Sigma^{-1} \Sigma^*] = \frac{1}{N} \text{tr}[\Sigma^{-1} \Sigma] = d_C/N$. Thus, we find

$$\begin{aligned}
& E_z[E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a}, M_{\text{perf}})]] \\
&= E_z[\log \text{pr}(z|\mathbf{a}^*, M_{\text{perf}})] - \frac{d_C}{2N}, \quad (112)
\end{aligned}$$

which in terms of information criteria translates to

$$\text{BPIC}_{\text{perf}} = \text{BAIC}_{\text{perf}} + d_C. \quad (113)$$

Although we do not focus on the PAIC (see Appendix D), since the PAIC and BPIC both estimate the same posterior averaged KL divergence, this also implies that $\text{PAIC}_{\text{perf}} = \text{BAIC}_{\text{perf}} + d_C$.

Finally, we consider the posterior predictive KL divergence, for which the posterior average and the log are transposed. To evaluate this, we first need the posterior average without the log,

$$E_{\mathbf{a}|\{y\}}[\text{pr}(z|\mathbf{a}, M_{\text{perf}})] = \frac{1}{Z} \int d\mathbf{a} \exp\left[-\frac{1}{2}(\bar{y} - \mathbf{a})^T (\Sigma^*)^{-1} (\bar{y} - \mathbf{a}) - \frac{1}{2}(z - \mathbf{a})^T \Sigma^{-1} (z - \mathbf{a})\right]. \quad (114)$$

Applying Eq. (E27), the result of the integration is

$$E_{\mathbf{a}|\{y\}}[\text{pr}(z|\mathbf{a}, M_{\text{perf}})] = \left(\frac{N}{N+1}\right)^{d_C/2} \exp\left[-\frac{1}{2} \frac{1}{N+1} (z - \bar{y})^T (\Sigma^*)^{-1} (z - \bar{y})\right]. \quad (115)$$

Taking the log and then the z -expectation, we find the result:

$$E_z[\log E_{\mathbf{a}|\{y\}}[\text{pr}(z|\mathbf{a}, M_{\text{perf}})]] = \frac{d_C}{2} \log \frac{N}{N+1} - \frac{1}{2} \frac{N}{N+1} E_z[(z - \bar{y})^T \Sigma^{-1} (z - \bar{y})] \quad (116)$$

$$= \frac{N}{N+1} E_z[\log \text{pr}(z|\mathbf{a}^*, M_{\text{perf}})] - \frac{d_C}{2} \log\left(1 + \frac{1}{N}\right), \quad (117)$$

or in terms of information criteria once more, and dropping the N -dependent constant term since it will cancel out in any model averages,

$$\text{PPIC}_{\text{perf}} = \frac{N}{N+1} \text{BAIC}_{\text{perf}} + Nd_C \log\left(1 + \frac{1}{N}\right) \quad (118)$$

$$= \frac{N}{N+1} (N+1)d_C + Nd_C \log\left(1 + \frac{1}{N}\right) \quad (119)$$

$$\approx (N+1)d_C - \frac{d_C}{2N} = \text{BAIC}_{\text{perf}} - \frac{d_C}{2N}. \quad (120)$$

When used in data subset selection, there is an additional factor of $(N-1)d_K$ that arises from the definition of

chi-squared over the kept part of the data. This combines with $(N-1)d_C$ to give an overall shift of $(N-1)d$, which is constant and may be dropped. Doing so, we find our final results for the contribution of the cut portion of the data to each information criterion:

$$\Delta_P \text{BAIC} = 2d_C, \quad (121)$$

$$\Delta_P \text{BPIC} = 3d_C, \quad (122)$$

$$\Delta_P \text{PPIC} = d_C + Nd_C \log\left(1 + \frac{1}{N}\right) \approx 2d_C - \frac{d_C}{2N}, \quad (123)$$

where Δ_P denotes the change in the overall model-averaging formulas due to the cut data in data subset selection.

Putting this together with the formulas for the kept data, we have our final results for the three ICs in the presence of data subset selection:

$$\text{BAIC}_{\mu,P} = \text{BAIC}_{\mu} + \Delta_P \text{BAIC} = \hat{\chi}^2(\mathbf{a}^*) + 2k + 2d_C, \quad (124)$$

$$\begin{aligned} \text{BPIC}_{\mu,P} = \text{BPIC}_{\mu} + \Delta_P \text{BPIC} \approx & \hat{\chi}^2(\mathbf{a}^*) - \frac{1}{2} \tilde{H}_{ba}(\Sigma^*)_{ab} \\ & + \frac{1}{2} \tilde{g}_d T_{cba}(\Sigma_2^*)_{abcd} + 3k + 3d_C, \end{aligned} \quad (125)$$

$$\begin{aligned} \text{PPIC}_{\mu,P} = \text{PPIC}_{\mu} + \Delta_P \text{PPIC} \\ \approx \hat{\chi}^2(\mathbf{a}^*) + 2k + d_C + Nd_C \log \left(1 + \frac{1}{N} \right) \\ - 2 \sum_{i=1}^N \log \left[1 + \frac{1}{2} \left(\frac{1}{4} (g_i)_b (g_i)_a - \frac{1}{2} (H_i)_{ba} \right) (\Sigma^*)_{ab} \right. \\ \left. + \frac{1}{4} (g_i)_d T_{cba}(\Sigma_2^*)_{abcd} \right], \end{aligned} \quad (127)$$

where $\hat{\chi}^2$ and all other quantities are evaluated only for the kept data. In cases of optimal truncation, Eq. (86) and (89) should be used for the BPIC and PPIC, respectively, with the addition of $\Delta_P \text{BPIC}$ and $\Delta_P \text{PPIC}$, respectively. We remind the reader that for use with model averaging, the factor $-2 \log \text{pr}(M_{\mu})$ should be added to all ICs as in Eq. (11), although this factor may be ignored completely if $\text{pr}(M_{\mu})$ is flat (independent of μ).

An alternative derivation for these formulas starting from the level of information criteria rather than KL divergence is provided in Appendix F. In addition to the relative simplicity of the KL divergence approach taken here, we are able to obtain exact results for KL_C , whereas starting from the ICs neglects higher-order bias corrections (see discussion in Appendix F).

As a final remark on the data subset selection procedure outlined above, it is worth discussing the full bias correction, i.e. the case in which $\text{tr}[J_N^{-1}(\mathbf{a}^*) I_N(\mathbf{a}^*)]$ is used rather than replacing the trace by the number of parameters. While the d_C contributions computed above are exact for the perfect model, in general the replacement of $\text{tr}[(J_N^{-1}(\mathbf{a}^*))_{\mathbf{K}} (I_N(\mathbf{a}^*))_{\mathbf{K}}] \rightarrow k$ on the kept data may not hold (as discussed in Sec. III A.) In particular, long-range correlations will correct all information criteria through contributions to this trace, as $\hat{\Sigma}_O^{-1}$ appears in the analytical expressions for $(J_N^{-1})_{\mathbf{K}}$ and $(I_N)_{\mathbf{K}}$. This can lead to numerical instabilities that will be more significant than the bias correction introduced by the full trace [35,66,67]. In such cases, the simplified bias corrections should be used even when the true model is not in the family of candidate models. In general, bias can be reduced by expanding the space of candidate models, ideally to include the true

model. In future work, it would be interesting to explore the use of more robust methods for estimation of these matrices, such as shrinkage [20,68–71].

VI. NUMERICAL TESTS

In this section, we give several numerical examples of model averaging with the various information criteria derived above and comparing their performance to fixed-model parameter estimation procedures.⁷ While tests with the ABIC_{CV} were conducted, we omit any numerical results due to its similar performance to the BAIC in the following examples. All Bayesian least squares fits were performed using the LSQFIT package in Python [47,72], which uses the Gaussian random variable data type from GVAR [73].

A. Example 1: Polynomial models

Consider a simple toy problem where the “true model” is a quadratic polynomial:

$$f_{\text{T}}(x) = 1.80 - 0.53 \left(\frac{x}{16} \right) + 0.31 \left(\frac{x}{16} \right)^2. \quad (128)$$

A set of N samples are generated on $x \in \{1, 2, \dots, 15\}$ using f_{T} at each point multiplied by uncorrelated noise $1 + \eta(x)$, where $\eta(x)$ is drawn from a Gaussian with mean $\bar{\eta} = 0.0$ and variance $\sigma_{\eta}^2 = 1.0$. To be explicit, the mock data are drawn from $y(x) = (1 + \eta(x)) f_{\text{T}}(x)$.

Our space of candidate models are polynomials labeled by their degree $\mu \in \{0, 1, \dots, 5\}$:

$$f_{\mu}(x) = \sum_{m=0}^{\mu} a_m \left(\frac{x}{16} \right)^m. \quad (129)$$

We take uniform model priors $\text{pr}(M_{\mu}) = 1/6$ corresponding to minimal prior information on the functional form of the true model (except that it can be approximated by a polynomial). We consider the case of moderately unconstrained parameter priors of the Gaussian form given in Eq. (61) with mean zero and width 10.

We use the previously developed model-averaging procedures to determine the parameter estimate and error for a_0 . Since the model functions are linear in the parameters, we use the following forms of the information criteria to determine the model weights:

$$\text{BAIC}_{\mu} = -2 \log \text{pr}(M_{\mu}) + \hat{\chi}^2(\mathbf{a}^*) + 2k, \quad (130)$$

$$\text{BPIC}_{\mu} = -2 \log \text{pr}(M_{\mu}) + \hat{\chi}^2(\mathbf{a}^*) - \frac{1}{2} \tilde{H}_{ba}(\Sigma^*)_{ab} + 3k, \quad (131)$$

⁷The code used to generate the examples in Secs. VI A and VI B is publicly available at https://github.com/jwsitison/improved_model_avg_paper.

TABLE I. Individual best-fit results with information criteria values and corresponding model weights for $N = 160$.

	$\mu = 0$	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$	$\mu = 5$	$\langle a_0 \rangle$
a_0	1.587(32)	1.803(67)	1.89(11)	2.01(16)	1.98(17)	1.94(18)	
a_1		-0.41(11)	-0.88(50)	-2.2(1.3)	-1.6(1.5)	-1.0(1.8)	
a_2			0.44(46)	3.6(3.0)	0.4(5.0)	-1.0(5.5)	
a_3				-2.1(2.0)	3.4(7.1)	-3.0(3.7)	
a_4					-3.0(3.7)	1.5(7.7)	
a_5						-3.1(4.7)	
$\hat{\chi}^2$	28.85	15.17	14.23	12.88	12.23	11.79	
Q-value	0.02	0.44	0.50	0.59	0.64	0.67	
BAIC	30.85	19.17	20.23	20.88	22.22	23.79	
$\text{pr}(M_\mu \{y\})_{\text{BAIC}}$	0.00	0.43	0.25	0.18	0.09	0.04	1.89(14)
BPIC	31.85	21.17	23.23	24.73	26.30	28.13	
$\text{pr}(M_\mu \{y\})_{\text{BPIC}}$	0.00	0.61	0.22	0.10	0.05	0.02	1.85(12)
PPIC	30.85	19.18	20.24	20.89	22.23	23.80	
$\text{pr}(M_\mu \{y\})_{\text{PPIC}}$	0.00	0.43	0.25	0.18	0.09	0.04	1.88(14)

$$\begin{aligned}
\text{PPIC}_\mu &\approx -2 \log \text{pr}(M_\mu) + \hat{\chi}^2(\mathbf{a}^*) + 2k \\
&- 2 \sum_{i=1}^N \log \left[1 + \frac{1}{2} \left(\frac{1}{4} (g_i)_b (g_i)_a \right. \right. \\
&\quad \left. \left. - \frac{1}{2} (H_i)_{ba} \right) (\Sigma^*)_{ab} \right]. \quad (132)
\end{aligned}$$

Due to the linearity of the model function in this example, the NLO Poincaré expansion formula is exact for the BPIC; the supersymptotic schema discussed in Sec. IV F are applied only for the PPIC, although in practice truncation does not occur in this test. Furthermore, the BPIC and PPIC are simplified for a linear model function since the tensor T is zero, see Sec. IV C. In this example the data set is held fixed, so we drop all terms depending on d_C .

The model-averaged results are summarized in Table I and shown in Fig. 1; we also report the Q -value of the fit (a Bayesian analog of the p -value, see Appendix B of [74]), which gives a measure of the fit quality. The model-averaged results are consistent with model truth but with a larger uncertainty than the individual fit to the correct model with $\mu = 2$. The larger error with model averaging is an inherent feature, reflective of a bias-variance trade-off; in the face of model uncertainty, model averaging hedges against the possibility of biased results due to selection of the wrong model, at the cost of increased error with a given data sample. See the further discussion in Sec. VII. In the top panel of Fig. 1, the advantage of model averaging over model selection is evident as the model probabilities happen to favor the $\mu = 1$ linear model, which is in fact incorrect. As the sample size N increases, this model will eventually be ruled out and the model weight will peak at the true model $\mu = 2$ as seen in the bottom panel of Fig. 1.

Note that in this case, the models are “nested” in the sense that any $\mu > 2$ can capture the true model by setting

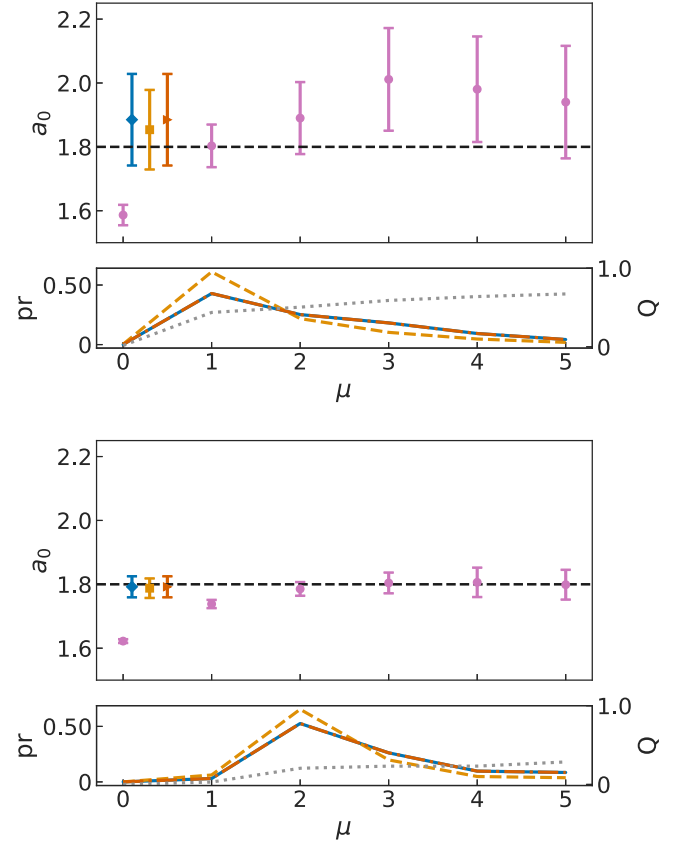


FIG. 1. Fit results at $N = 160$ (top panel) and $N = 5120$ (bottom panel) for $f_\mu(x)$ (purple solid circles) and model-averaged results with the BAIC (blue solid diamonds), BPIC and PAIC (yellow solid squares) and PPIC (red solid triangles) compared to the known value $a_0 = 1.80$ (black dashed line). The lower inset shows the standard Q -value (gray dotted curve) and the model weights $\text{pr}(M_\mu|\{y\})$ from the BAIC (blue solid curve), BPIC and PAIC (yellow dashed curve), and PPIC (red dash-dotted curve).

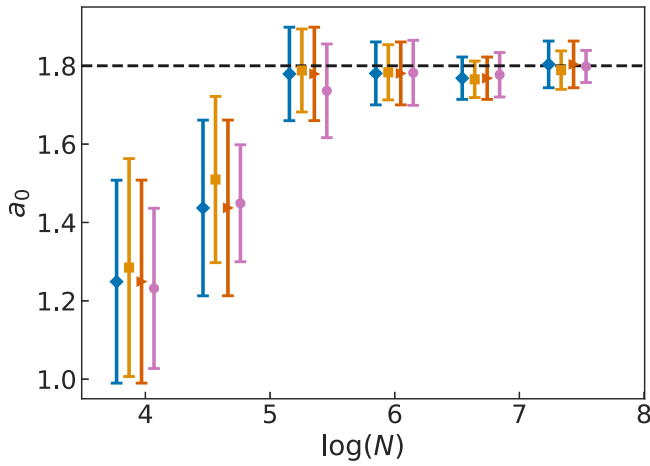


FIG. 2. N -dependent scaling of the various estimates of the intercept a_0 . The true value (black dashed line) is $a_0 = 1.80$. The model-averaged results using the BAIC (blue solid diamonds), the BPIC and PAIC (yellow solid squares), and PPIC (red solid triangles) are consistent with both model truth and the fit results of the correct quadratic model (purple solid circles).

higher-order a_m to zero. This means that even as $N \rightarrow \infty$, the model probability $\text{pr}(M_\mu|\{y\})$ for $\mu > 2$ will never go to zero, although models with higher complexity will be penalized by the $+2k$ bias-correction term so that peak model probability will be at $\mu = 2$.

Observe that the BPIC has less uncertainty than the BAIC or PPIC. This is because the BPIC favors simpler models even more so than the BAIC or PPIC as a result of the additional k that comes from the posterior averaging of the KL divergence when parameter priors are sufficiently diffuse, as discussed in Sec. IV D. The additional parsimony implied by the BPIC comes with a larger KL divergence for the model distribution. While this might cause concern that the BPIC may actually underestimate the model uncertainty in this example, in practice this does not appear to be the case (as seen in this example by comparing the BPIC-averaged parameter uncertainty with that for the true $\mu = 2$ model).

We repeat the previous numerical test with several values of $N = 40, 80, 160, 320, 640, 1280$; the final estimates for a_0 are in Fig. 2. The model-averaged results using the

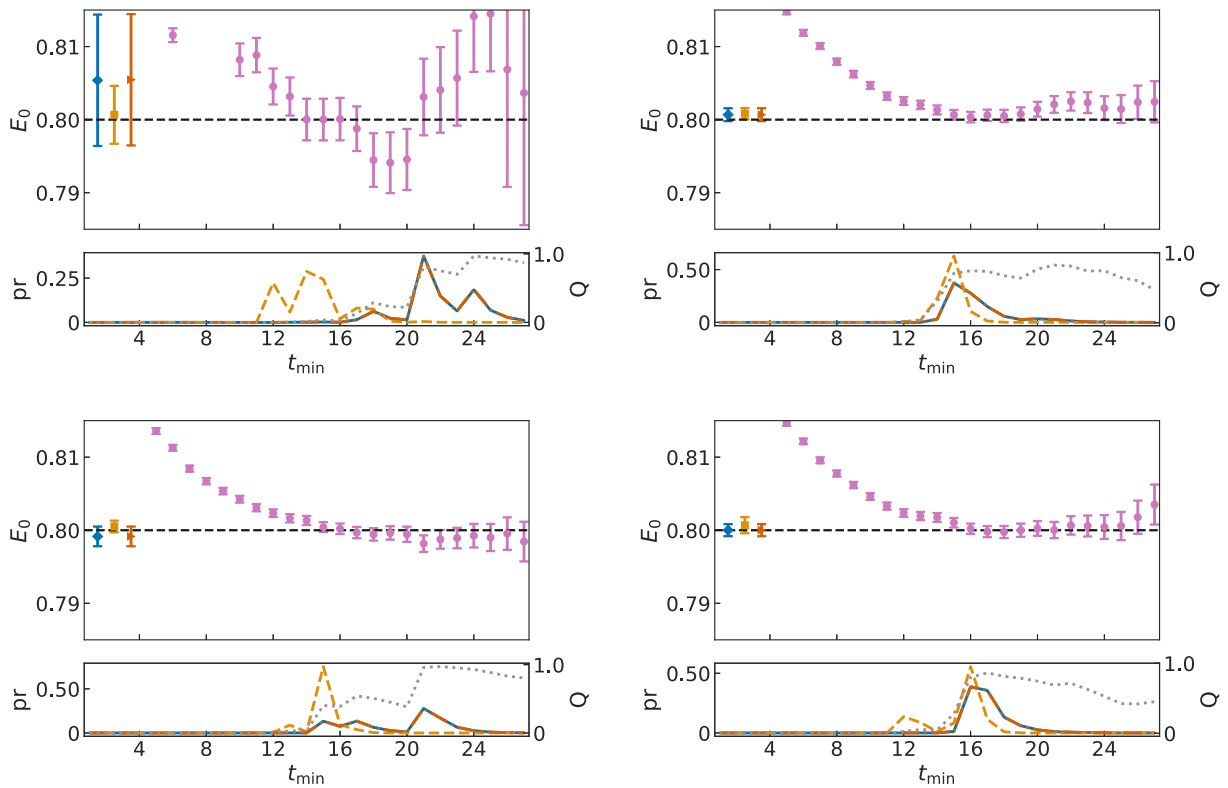


FIG. 3. Fit results for the ground-state energy with true value $E_0 = 0.80$ (black dashed line) for $N = 30$, $\sigma_\eta = 0.3$, and $\sigma_\theta = 0$, for data subset $t \in [t_{\min}, 31]$ (purple solid circles). Model-averaged results with the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles) agree well with model truth and each other in any case. The lower inset shows the standard Q -value (gray dotted line) and model weights $\text{pr}(M_\mu|\{y\})$ corresponding to the BAIC (blue solid curve), BPIC (yellow dashed curve), and PPIC (red dash-dotted curve). The four separate figures represent four random draws of correlated Gaussian fractional noise, but are otherwise identical.

BAIC, BPIC, and PPIC are consistent with model truth in all cases. As in the fixed- N study, the uncertainties in the model-averaged estimates are typically larger than that of the fixed quadratic estimate; this is because model averaging has two sources of error—(i) parameter uncertainty in individual fits and (ii) variance across the individual fit means—whereas using a fixed model only has the former.

B. Example 2: Exponential model and subset selection

To test the data subset selection procedure developed in Sec. V, we consider another toy problem meant to resemble a two-point correlation function (see Sec. V for a brief discussion of two-point correlators). For this example, we will set model truth to a two-state exponential:

$$f_T(t) = 2.0e^{-0.80t} + 10.4e^{-1.16t}. \quad (133)$$

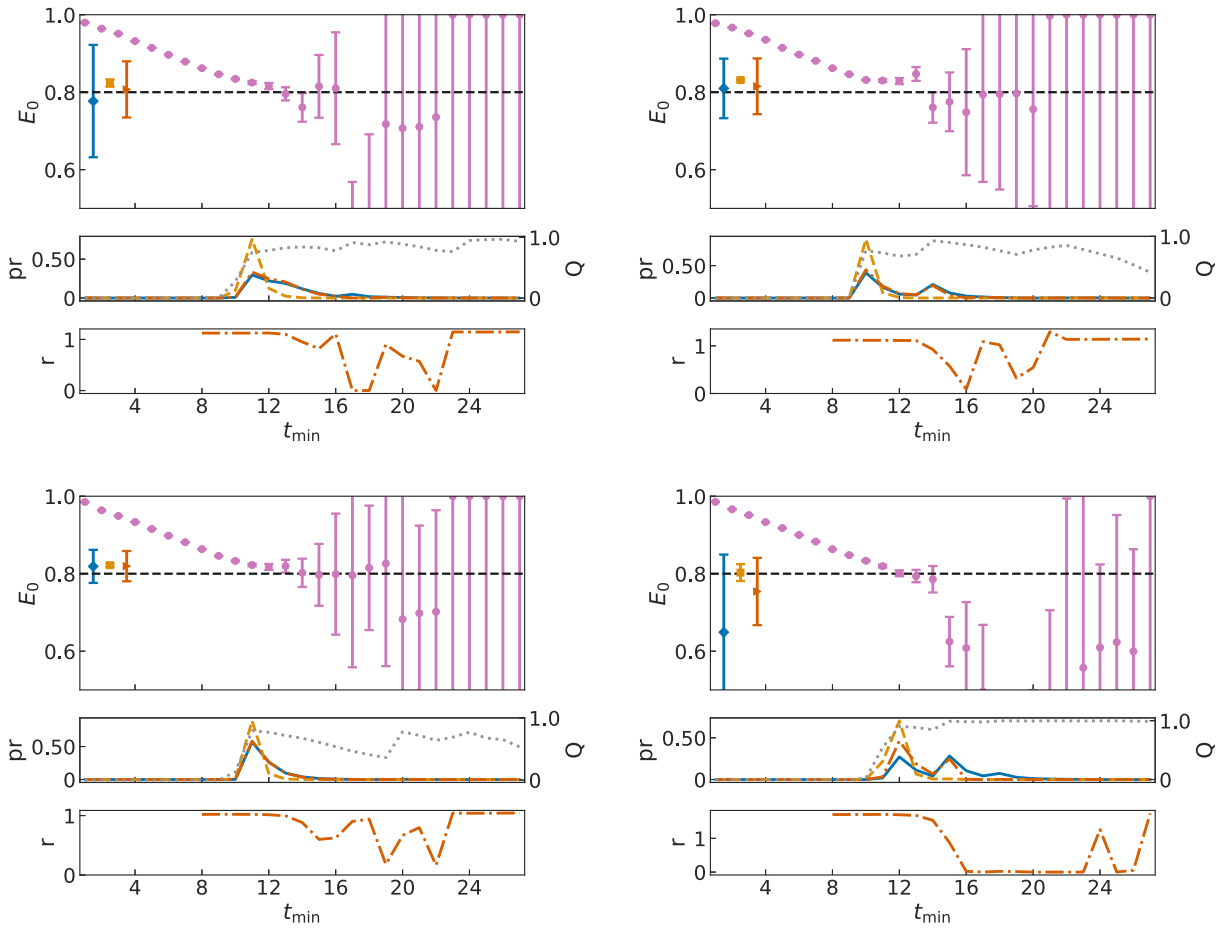


FIG. 4. Fit results for the ground-state energy with true value $E_0 = 0.80$ (black dashed line) for $N = 200$, $\sigma_\eta = 0.003$, and $\sigma_\theta = 10^{-5}$, for data subset $t \in [t_{\min}, 31]$ (purple solid circles). Model-averaged results with the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles) agree well with model truth and each other in any case. The middle inset shows the standard Q -value (gray dotted line) and model weights $\text{pr}(M_\mu|\{y\})$ corresponding to the BAIC (blue solid curve), BPIC (yellow dashed curve), and PPIC (red dash-dotted curve). The lower inset shows the ratios of the PPIC model weights to the BAIC model weights $r \equiv \text{pr}(M_\mu|\{y\})_{\text{PPIC}}/\text{pr}(M_\mu|\{y\})_{\text{BAIC}}$. The four separate figures represent four random draws of correlated Gaussian fractional noise and uncorrelated Gaussian additive noise, but are otherwise identical.

To generate synthetic data, we multiply f_T by correlated noise $1 + \eta(t)$, where $\eta(t)$ is drawn from a Gaussian with mean $\bar{\eta} = 0.0$ and standard deviation $\sigma_\eta \in \{0.3, 0.003\}$, as well as an uncorrelated noise floor $\theta(t)$ drawn from a Gaussian with mean $\bar{\theta} = 0.0$ and standard deviation $\sigma_\theta \in \{0, 10^{-5}\}$, i.e., the synthetic data are generated from $y(t) = (1 + \eta(t))f_T(t) + \theta(t)$. The correlation matrix used to generate $\eta(t)$ takes the form $\rho_{xy} = \rho^{|t_x - t_y|}$ so that ρ equals one on the diagonal and decreases as a power law as the temporal separation between points increases (similar to a real lattice QCD correlation function). We fix the correlation coefficient to $\rho = 0.6$. We generate N mock data samples on $t \in \{1, 2, \dots, 31\}$; the initial time is omitted from the analysis due to the certain excited state contamination at $t = 0$. Other sets of parameters with $\sigma_\theta = 0.0$ were considered in [26] with and without correlation on $\eta(t)$ to no qualitative effect; the same is true for $\sigma_\theta > 0$

except for very large values of ρ where numerical issues lead to unreliable simulations.

We note that in terms of the observed signal-to-noise ratio in the mock data, the case without the noise floor $\sigma_\theta = 0$ has roughly constant signal-to-noise, analogous to a pion two-point correlation function [75]. On the other hand, with $\sigma_\theta > 0$ a threshold in t is introduced at which the signal-to-noise drops exponentially, with the signal being completely overwhelmed at large t where $\sigma_\theta \gg f_T(t)$. This more closely resembles the behavior of something like a nucleon two-point function, with extremely poor signal-to-noise at large time separation reflective of a sign problem [76,77]. These resemblances to real data are, at best, qualitative; we do not claim to use a realistic noise model for this toy example. We will consider the application of our methods to real lattice QCD data in Sec. VIC.

Initially, we consider a single candidate model that consists of a single exponential term:

$$f_1(t) = A_0 e^{-E_0 t}. \quad (134)$$

This model is fit to data $(y_i)_x$ corresponding to $t_x \in [t_{\min}, 31]$ for $t_{\min} \in \{1, 2, \dots, 28\}$. Model-averaged results for the ground-state energy E_0 are obtained using

$$\text{BAIC}_{\mu, d_C} = -2 \log \text{pr}(M_\mu) + \hat{\chi}^2(\mathbf{a}^*) + 2k + 2d_C, \quad (135)$$

$$\begin{aligned} \text{BPIC}_{\mu, d_C} &\approx -2 \log \text{pr}(M_\mu) + \hat{\chi}^2(\mathbf{a}^*) - \frac{1}{2} \tilde{H}_{ba}(\Sigma^*)_{ab} \\ &+ \frac{1}{2} \tilde{g}_d T_{cba}(\Sigma^*)_{abcd} + 3k + 3d_C, \end{aligned} \quad (136)$$

$$\begin{aligned} \text{PPIC}_{\mu, d_C} &\approx -2 \log \text{pr}(M_\mu) + \hat{\chi}^2(\mathbf{a}^*) + 2k + d_C \\ &+ Nd_C \log \left(1 + \frac{1}{N} \right) - 2 \sum_{i=1}^N \log \left[1 \right. \\ &+ \frac{1}{2} \left(\frac{1}{4} (g_i)_b (g_i)_a - \frac{1}{2} (H_i)_{ba} \right) (\Sigma^*)_{ab} \\ &\left. + \frac{1}{4} (g_i)_d T_{cba}(\Sigma^*)_{abcd} \right], \end{aligned} \quad (137)$$

as derived in Sec. V. In this example, $d_C = t_{\min}$ entirely determines model complexity as $k = 2$ is fixed across the space of models. Unlike the polynomial example, the integrals used to obtain the information criteria are not computed exactly as the model function Eq. (134) is non-linear in the model parameters. To improve the accuracy of these results, we implement the superasymptotic schema described in Sec. IV F for the BPIC and PPIC.

The results of four independent trials of the above procedure with $(N, \sigma_\eta, \sigma_\theta) = (30, 0.3, 0)$ and $(N, \sigma_\eta, \sigma_\theta) = (200, 0.003, 10^{-5})$ are shown in Figs. 3 and 4, respectively. Excited-state contamination is clear at low t_{\min} as the second exponential state has influence over the fit results

before it has decayed away. Model-averaged results agree well with model truth for all information criteria considered. Like the polynomial example, the model-averaged results favor parsimonious models; in the present context, parsimony corresponds to fits that cut away as little data as possible without compromising fit quality.

In Fig. 3, where there is no noise floor, the BAIC and PPIC perform identically. The BPIC shows a preference for smaller t_{\min} cuts due to the extra factor of d_C in its subset selection formula. This leads to generally smaller uncertainty but potential for finite-sample bias, which will be more strongly evident below in the presence of additional noise.

Figure 4 represents a more challenging noisy-data case study, as the noise floor $\theta(t)$ is typically larger than the last several data points, which is reflected in the large E_0 error for fits at large t_{\min} . In this case, the PPIC consistently outperforms the BAIC in its estimation of the parameter

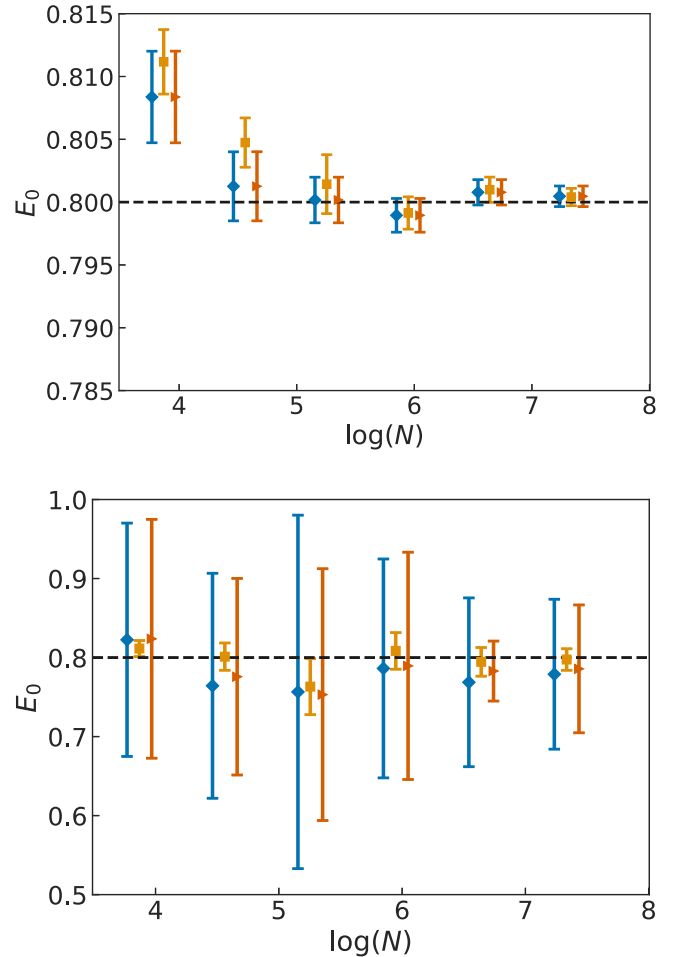


FIG. 5. N -dependent scaling of the various estimates of the ground-state energy E_0 for $\sigma_\eta = 0.3$ and $\sigma_\theta = 0$ (top) and $\sigma_\eta = 0.003$ and $\sigma_\theta = 10^{-5}$ (bottom). The true value (black dashed line) is $E_0 = 0.80$. The model-averaged results using the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles) are consistent with model truth in all cases.

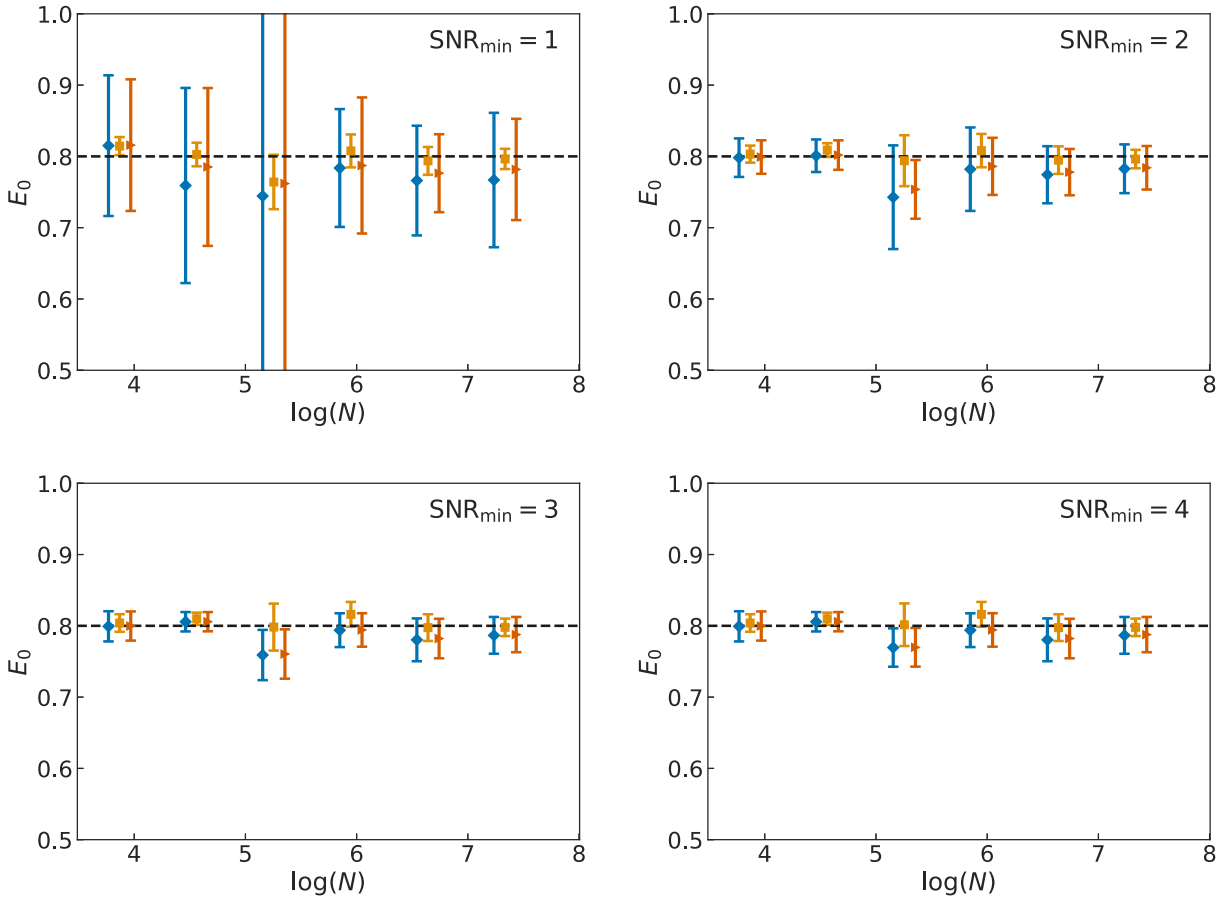


FIG. 6. Model-averaging results versus the number N of data samples included in the analysis. The data is the same as the bottom panel of Fig. 5 and the subset sizes shown are $N = 40, 80, 160, 320, 640, 1280$. The black dashed line shows the value of the model truth. The model-averaged results shown use the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles). Each panel corresponds to a different value of minimum signal-to-noise ratio imposed on the data denoted by $\text{SNR}_{\min} = 1, 2, 3, 4$ in the top right of each panel. For $\text{SNR}_{\min} = 0$, see Fig. 5.

mean and the error of this estimate. This results from the phenomenon discussed in Sec. IV E where, in its use of information from each individual observation, the PPIC is able to penalize models that fail to predict the data and hence give poor parameter estimates. The improved error of the PPIC E_0 estimate is coupled to this effect as poor models tend to give larger parameter uncertainties. The relatively low weight for noisy models at high t_{\min} using PPIC versus BAIC can be seen from the lowest panel in each subfigure, which shows the ratio $r \equiv \text{pr}(M_\mu|\{y\})_{\text{PPIC}}/\text{pr}(M_\mu|\{y\})_{\text{BAIC}}$ of the model weights using PPIC to BAIC. The PPIC also outperforms the BPIC in terms of finite- N bias; while the BPIC often gives much smaller error estimates than the other information criteria, this is due to its overly aggressive penalty for d_C which heavily weights a single fit with small t_{\min} as seen in the figure. As a result, the BPIC often disagrees strongly with the true asymptotic value for E_0 .

In Fig. 5, we repeat the model-averaging test with varying sample size N , again over $N = 40, 80, 160, 320, 640, 1280$,

with $\sigma_\eta = 0.3$ and $\sigma_\theta = 0$ (top panel) and with $\sigma_\eta = 0.003$ and $\sigma_\theta = 10^{-5}$ (bottom panel). All model-averaged results based on the tested information criteria agree well with model truth. As in Fig. 2, model averaging leads to larger uncertainties compared to the result of fixing a single fit at fixed t_{\min} . However, the fixed-model approach does not account for systematic error due to model truncation, as t_{\min} must be adjusted as $N \rightarrow \infty$ for the estimate of E_0 to remain uncontaminated by higher-energy states.

Though model averaging seems to protect the final results from excessively noisy data, the exponential signal-to-noise problem when the noise floor is present may cause concern about the accuracy of Laplace's method in these cases. In general, one should take caution in using data subset averaging procedure when a region of the data is effectively pure noise. (Pure noise is, in some sense, drawn from the incorrect distribution, one centered at zero instead of around the desired signal.) A straightforward approach to mitigating this effect is to impose a minimum signal-to-noise cut on the data before implementing the

model-averaging procedure. Figure 6 shows the results of different minimum signal-to-noise ratios for the same N -scaling test data used in the bottom panel of Fig. 5. Because the fit model decreases monotonically, we cut away the data for t greater than the first time where the minimum signal-to-noise ratio is exceeded.

We emphasize that the signal-to-noise ratio cut in these tests is distinct from the data subset selection procedure outlined in Sec. V. Unlike the choice of a t_{\min} cut, a signal-to-noise cut removes data with minimal information content and therefore the choice of a specific threshold by the analyst will not be a significant source of systematic error. Since a signal-to-noise cut discards the data before any further analysis, no information penalty need be assessed when it is used, i.e., d_C does not include the signal-to-noise cut data.⁸

Even with the minimum signal-to-noise ratio imposed on the data in Fig. 6, the variance on the model-averaged results behaves counterintuitively. Namely, the uncertainty does not decrease monotonically with N . This is a symptom of the systematic error due to model truncation discussed in the context of the fixed-model in Fig. 5. As N increases, small values of t are not well approximated by one-state exponential for the fit model Eq. (134). For this reason, small t_{\min} values contaminate the model-averaged results, inflating parameter uncertainties. Eventually, as N increases, the data will be precise enough that there will be no region in which the one-state model is sufficient to describe the data. Continuing to use model averaging over only one-state fits in this limit would lead to incorrect results due to model misspecification. Instead, the model space should be expanded; a two-state fit to

$$f_2(t) = A_0 e^{-E_0 t} + A_1 e^{-E_1 t} \quad (138)$$

can be performed to account for the excited state contamination. In contrast to correlator fits to real lattice data (cf. Sec. VIC), Eq. (138) is of the same form as Eq. (133) so there should be no further contamination from higher excited states. To improve numerical stability and ensure ordering of energy levels, the excited-state energy is fit in practice using the fit parameter $ldE_1 \equiv \log(E_1 - E_0)$ to replace E_1 . Eq. (138) is fit to construct Figs. 7 and 8 for a minimum signal-to-noise ratio of $\text{SNR}_{\min} = 4$, which is sufficient to stabilize the results as shown in Fig. 6. Figure 7 shows the expected improved accuracy of the fits at small t_{\min} using the two-state model. Here the BPIC shows no bias with respect to the true value of E_0 , since the two-state model is exactly the true model from which the data are drawn.

⁸While this distinction is moot in the examples above (where the resulting penalty would change each IC by only an additive constant), it could be important if the BPIC or PPIC with long-range correlations are used (see Sec. V).

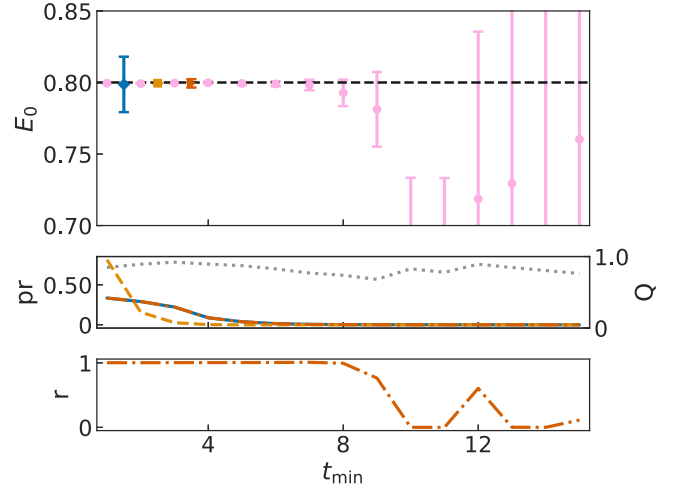


FIG. 7. Two-state fit results for the ground-state energy with true value $E_0 = 0.80$ (black dashed line) for $N = 640$, $\sigma_\eta = 0.003$, and $\sigma_\theta = 10^{-5}$, for data subset $t \in [t_{\min}, 31]$ (pink solid circles); a minimum signal to noise ratio of $\text{SNR}_{\min} = 4$ has been imposed on the data. Model-averaged results with the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles) agree well with model truth and each other in any case. The middle inset shows the standard Q -value (gray dotted line) and model weights $\text{pr}(M_\mu|\{y\})$ corresponding to the BAIC (blue solid curve), BPIC (yellow dashed curve), and PPIC (red dash-dotted curve). The lower inset shows the ratios of the PPIC model weights to the BAIC model weights $r \equiv \text{pr}(M_\mu|\{y\})_{\text{PPIC}} / \text{pr}(M_\mu|\{y\})_{\text{BAIC}}$.

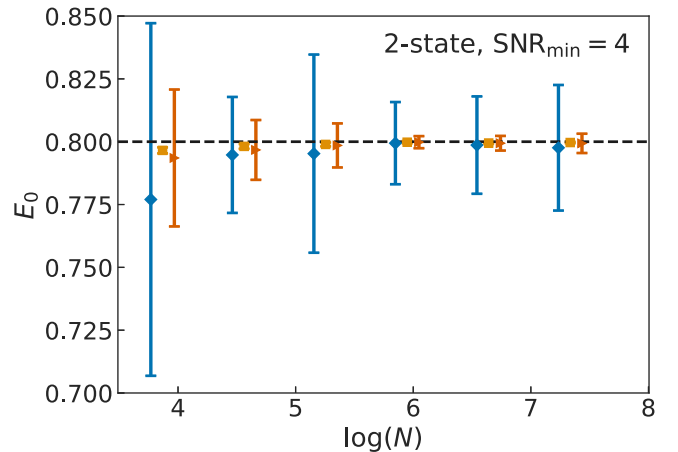


FIG. 8. N -dependent scaling of the various estimates of the ground-state energy E_0 for $\sigma_\eta = 0.003$ and $\sigma_\theta = 10^{-5}$ using a two-state fit; a minimum signal to noise ratio of $\text{SNR}_{\min} = 4$ has been imposed on the data. The true value (black dashed line) is $E_0 = 0.80$. The model-averaged results using the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles) are consistent with model truth.

While Fig. 8 shows improved behavior of the model-averaged uncertainty at large N compared to Fig. 6 (particularly for the PPIC), the uncertainty still does not decrease monotonically as a function of N , as we would

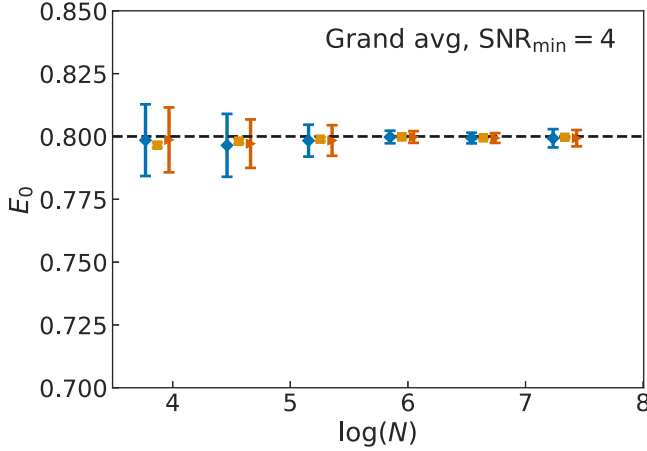


FIG. 9. N -dependent scaling of the various estimates of the ground-state energy E_0 for $\sigma_\eta = 0.003$ and $\sigma_\theta = 10^{-5}$ averaging over both data subsets and the one- and two-state models; a minimum signal to noise ratio of $\text{SNR}_{\min} = 4$ has been imposed on the data. The true value (black dashed line) is $E_0 = 0.80$. The model-averaged results using the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles) are consistent with model truth.

expect for, e.g., a parameter estimate from a single model. This behavior may be concerning at face value since it indicates situations where adding more data does not result in reduction of error. This behavior cannot persist indefinitely; since the error on individual sample estimates of E_0 scales as $1/\sqrt{N}$, and the model average is obtained from these estimates, the model-averaged uncertainty must decrease as $1/\sqrt{N}$ asymptotically. The enhanced fluctuations in the model-averaged uncertainty versus sample size compared to single-model results are an inherent trade-off for robustness against incorrect model specification; see the discussion of bias-variance trade-off in Sec. VII.

Figure 9 shows the results of a “grand average” procedure, i.e., a model average including both one-state and two-state fits in addition to data subset selection. If we attempt to compare these results to both Figs. 6 and 8, we see that the grand average results actually give more precise estimates than either individual set of fits. This shows concretely that expansion of the model space does not always result in increased uncertainty; in this case, the relative weight of individual fits with larger errors seems to be reduced in the grand average.

Briefly summarizing what we have found in our numerical results above, our main finding is that the PPIC shows the best overall performance in all tests. Its precision is generally the same as or better than the precision of the BAIC, but without any signs of statistically significant bias versus the true value of E_0 . The PPIC is especially robust in cases where very noisy model estimates are part of the average; when signal-to-noise cuts are used, the PPIC’s advantage tends to decrease relative to BAIC as the latter is

improved more significantly, although there are counter-examples, e.g., Fig. 8. The BAIC is the simplest criterion and also shows good performance with respect to the absence of statistically significant finite- N bias. On the other hand, the BPIC is overly aggressive in penalizing data cuts, resulting in the highest precision in many cases but together with a significant bias, which would be unacceptable in many applications in lattice field theory.

As a final remark on this example, we note that accuracy of the numerical approximations developed in Secs. IV C and IV F have been corroborated in several test cases using the VEGAS algorithm, an importance-sampling-based Monte Carlo integration scheme [78–80]. We do not show additional tables or figures with VEGAS evaluation of the full integrals, as in all cases tested these results are essentially indistinguishable from the approximate formulas.

C. Example 3: Lattice QCD correlation functions

To further test our methodology on a more realistic example, we apply it to a real lattice QCD dataset, specifically a nucleon two-point correlation function. The data consists of measurements on 615 configurations from the JLab/W&M/MIT/LANL ensemble a091m170 (see [81] for details). On each configuration, correlators were measured on an even grid of 512 sources, projected to zero momentum, and averaged over sources. Gauge-invariant Gaussian smearing to radius 4.5 was applied at both the source and sink. We find no evidence of residual thermalization or autocorrelation effects, so we take these samples to be independent. All numerical values below are provided in implicit lattice units, i.e., $a = 1$.

We carry out fits using two different model functions: a simple one-state model Eq. (134), and a two-state model Eq. (138). In practice, the parameter ldE_1 defined in Sec. VI B is fit in lieu of E_1 for the same reasons discussed in Sec. VI B. The fits are done with ground-state priors $A_0 = 3(3000) \times 10^{-8}$ and $E_0 = 0.4(4)$; these were chosen primarily to ensure fit convergence. For two-state fits, the priors are $A_1 = 3(10000) \times 10^{-8}$ and $ldE_1 = -0.5(1.0)$. The choice of parameter priors does not affect the results qualitatively.

Individual fit results versus t_{\min} , as well as the corresponding model-averaging results, are shown in Figs. 10 and 11. Here we initially impose no signal-to-noise cut, although we would certainly advocate for doing so in a serious analysis of this data and will explore the effect of such a cut below. Qualitatively, we see that the results are very similar to those obtained for the synthetic two-state exponential example above: overall the model-averaged results for the PPIC and BAIC are very similar, but the PPIC is slightly more precise. The BPIC results tend to be much more precise, but suffer from potential bias, giving estimates at smaller N which significantly disagree with the full dataset estimates.

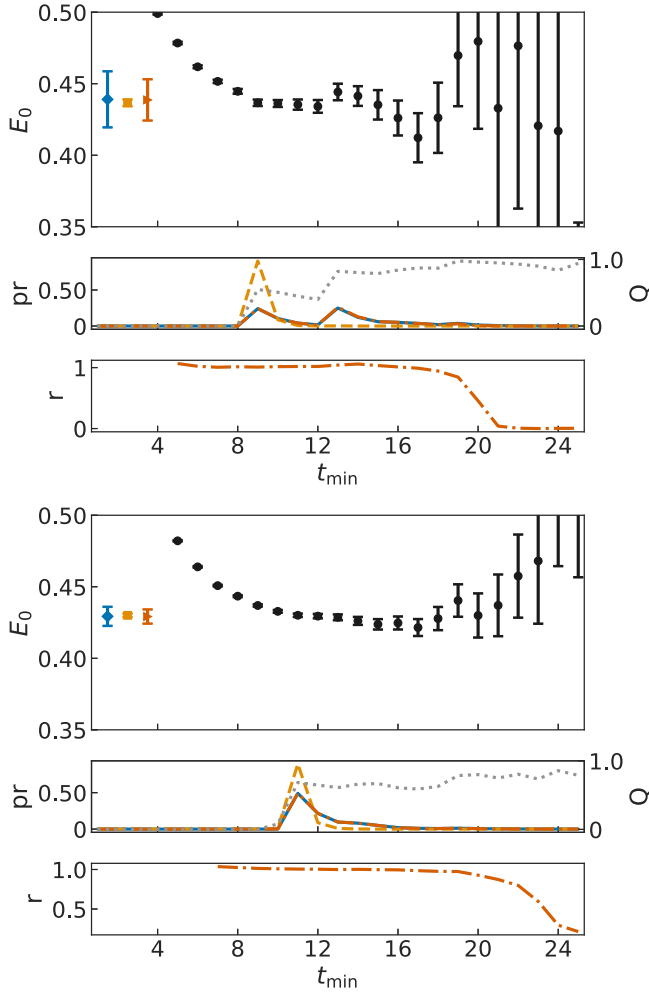


FIG. 10. Fit and model-averaging results for the lattice QCD nucleon data using the one-state model for $N = 40$ (top panel) and $N = 615$ (bottom panel). Individual fit results to data subsets $t \in [t_{\min}, 30]$ are shown as black solid circles. The model-averaged results shown use the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles). The lower inset shows the standard Q -value (gray dotted line) and model weights $\text{pr}(M_\mu|\{y\})$ corresponding to the BAIC (blue solid curve), BPIC (yellow dashed curve), and PPIC (red dash-dotted curve). The lower inset shows the ratios of the PPIC model weights to the BAIC model weights $r \equiv \text{pr}(M_\mu|\{y\})_{\text{PPIC}}/\text{pr}(M_\mu|\{y\})_{\text{BAIC}}$.

Within Fig. 10, we also show the difference in results between the full dataset $N = 615$ and a much smaller subsample of $N = 40$. Within the smaller dataset, increased model uncertainty is apparent in the plot of $\text{pr}(M_\mu|\{y\})$, with multiple local peaks appearing for the BAIC and PPIC. This uncertainty is reflected in relatively large model-averaged uncertainty relative to any of the individual fits in the “plateau” region for these ICs. With the full $N = 615$ dataset, the model probability becomes sharply peaked at the lowest value of t_{\min} for which the one-state model is a good description of the data, decaying exponentially as t_{\min} increases; this is precisely the behavior we expect at

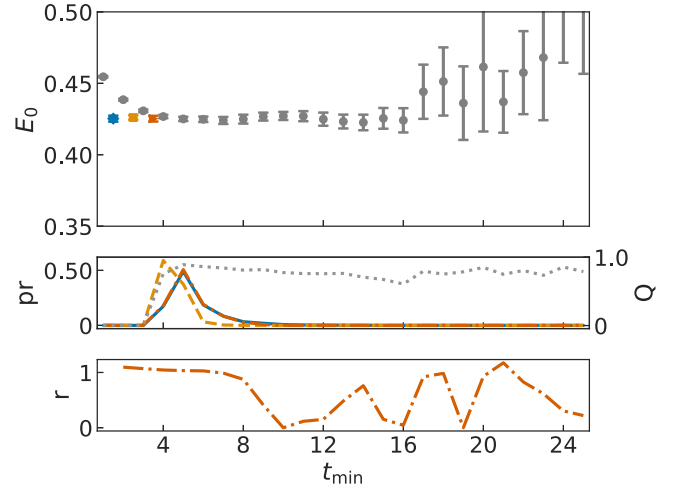


FIG. 11. Fit and model-averaging results for the lattice QCD nucleon data ($N = 615$) using the two-state model. Individual fit results to data subsets $t \in [t_{\min}, 30]$ are shown as gray solid circles. Other colors, symbols, and the lower subpanel are all defined as in Fig. 10.

relatively large N . Note that despite their close agreement for most fits, the PPIC once again results in a smaller uncertainty than the BAIC due to the rejection of noisy fit results at the largest t_{\min} , as visible from the ratio r of PPIC to BAIC model weights which goes to zero at large t_{\min} .

We further explore dependence on sample size using the nucleon data by cutting down to the first N out of 615 measurements and repeating the model-averaging analysis, for N taking on the values $\{40, 80, 160, 320, 615\}$. The results of this procedure are shown in Fig. 12, imposing various levels of minimum signal-to-noise cut and for both one-state and two-state fits; the grand average over both models is shown in Fig. 13. Most of the qualitative conclusions are very similar to those drawn from the synthetic data example; imposition of a signal-to-noise cut generally improves both the total uncertainty at fixed N and the scaling of errors as a function of N . For the one-state fits (left column), we can see a clear saturation of the error as N increases, with no decrease in uncertainty from $N = 320$ to $N = 615$. This effect can be explained by the effect discussed in the previous subsection and clearly visible in Fig. 10, namely that as N increases and data errors decrease, the “plateau” region in which the one-state model adequately describes the data shrinks. The saturation of model-averaged uncertainty is thus an indicator that our model space is incomplete. Indeed, we see that going instead to two-state fits (either exclusively or in the grand average) allows smaller error estimates to be obtained from the full $N = 615$ dataset.

An important difference between this real-world nucleon data and the controlled example shown above is that for the nucleon data, the “true model” is not accessible—in principle, it is a sum over an infinite number of excited

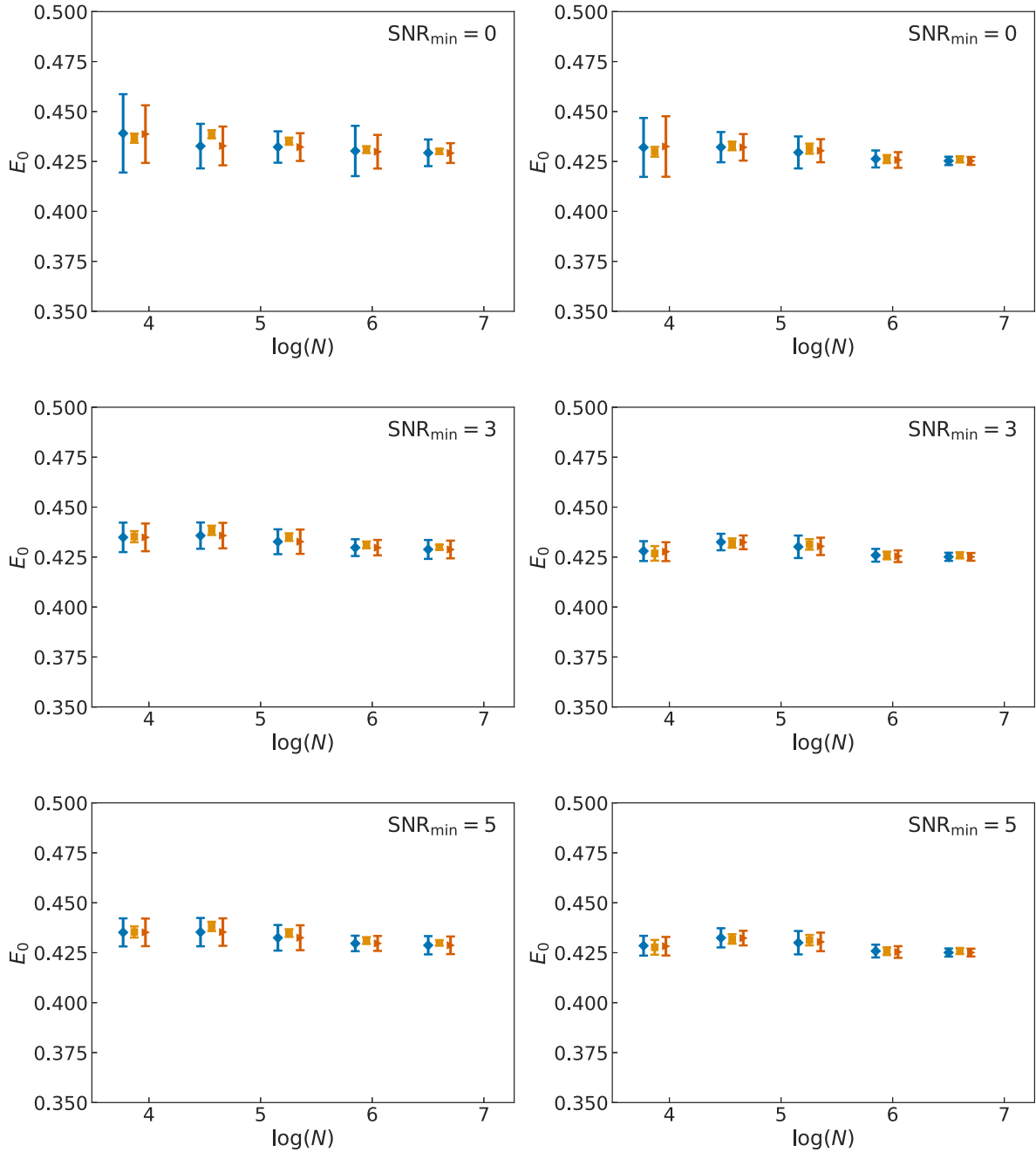


FIG. 12. For the lattice QCD nucleon data, model-averaging results versus the number N of data samples included in the analysis. The data subset sizes shown are $N = 40, 80, 160, 320, 615$. The model-averaged results shown use the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles). The left column shows results obtained with the one-state fit model; the right column shows results using the two-state fit model. Each row corresponds to a different value of minimum signal-to-noise cut imposed on the data: from the top row, the values used are $\text{SNR}_{\min} = 0, 3, 5$.

states. This means that our true model is never contained in the model space we consider, strictly speaking, no matter how many excited states we include. However, at any given N and t_{\min} the difference between the true model and a truncated model will be exponentially small as long as enough states are included. For a given data sample, the Bayesian model-averaging philosophy suggests to include all possible numbers of states and perform a “grand

average” to determine the relative weights. In practice, one should typically truncate the number of states once the number required to give stable descriptions of the data and saturation of error estimates, e.g. as done in [47] without model averaging.

Overall, the performance of the three information criteria tested against this real-world data mirrors what we saw in the toy-model case of Sec. VI B. The BAIC and PPIC are

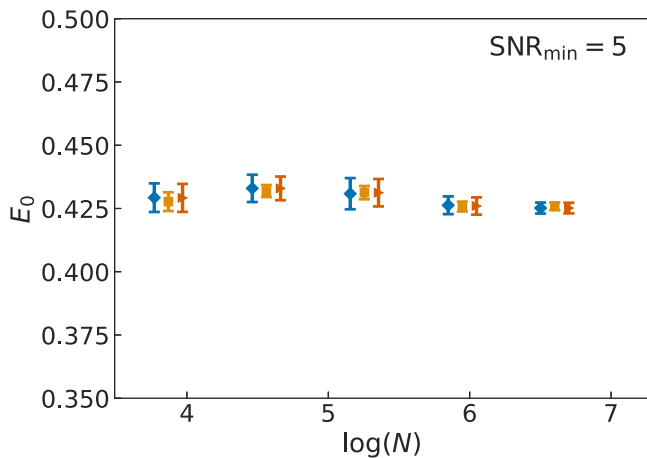


FIG. 13. For the lattice QCD nucleon data, combined results averaged across both data subsets and the single and double exponential model for $\text{SNR}_{\min} = 5$ versus the number N of data samples included in the analysis. The data subset sizes shown are $N = 40, 80, 160, 320, 615$. The model-averaged results shown use the BAIC (blue solid diamonds), BPIC (yellow solid squares), and PPIC (red solid triangles).

consistent with one another, with the PPIC often having smaller uncertainties particularly in the presence of noisy estimates for E_0 . The BPIC shows the smallest uncertainty but often also shows a consistent offset relative to the other ICs, and based on our other numerical results we would be concerned that its results are biased significantly.

VII. CONCLUSION

We have adapted several information criteria from the model selection literature for use in Bayesian model averaging. The information criteria described give asymptotically unbiased estimates of the Kullback-Leibler divergence, which is sufficient to remove asymptotic bias from model-averaged results when the regression procedure gives consistent model parameter estimates. By connecting these ideas to the KL divergence, we are able to present a very general and rigorous statistical theory. We also provide specialized discussion of least squares, which is a consistent regression procedure. In the case of least squares, we derived numerically efficient and accurate asymptotic and superasymptotic approximations of the information criteria using Laplace's method and optimal truncation, respectively; these approximations are in fact exact for linear fit functions. For each information criteria, we extend the model-averaging framework to data subset averaging.

The information criteria studied are the BAIC, BPIC, and PPIC, all of which are asymptotically unbiased with $o(N^{-1})$ finite- N bias, where N is the data sample size [12,13,32]. The approximate formulas provided for the BPIC and PPIC are at least $O(N^{-1})$ accurate; higher orders of accuracy (potentially exponential) may be achieved in cases of optimal truncation. We have chosen not to study

the PAIC in detail in the body due to the lower order of accuracy [$O(1)$] of the integral approximation required and its theoretical similarity to the BPIC; the relevant PAIC formulas are given in Appendix D.

Each of the ICs have various strengths and weaknesses. The simplest is the BAIC, which only requires the number of parameters (and the number of the excluded data points) and evaluation of the likelihood function (i.e., $\hat{\chi}^2$) at the posterior mode \mathbf{a}_{PM}^* [i.e., the (Bayesian) best-fit point]. The other information criteria studied can be thought of as finite sample size corrections to the BAIC as they all give equivalent model probabilities in the large N limit. On the other hand, the use of a plug-in estimator in the BAIC is closer to frequentist than to Bayesian statistical practice as it assumes that there is a true value as opposed to a true distribution. A more natural treatment of parametric models in Bayesian inference should integrate over an estimate of this true distribution. One approach is to average the log likelihood over the posterior giving rise to the BPIC. However, due to Jensen's inequality, the BPIC is unable to give a smaller KL divergence than the PPIC where the log is taken after doing the posterior averaging. This seemingly minor change gives PPIC the ability to sense to individual fluctuations in the data leading to better performance in practice, particularly in situations where signal-to-noise is poor. As seen in our numerical tests, this allows the PPIC to in some cases outperform both the BAIC, giving smaller error estimates, and the BPIC, giving greatly reduced finite-sample bias. For this reason, we advocate for the use of the PPIC in applications of Bayesian model averaging. We note in passing that there are instances in which the individual data are not accessible and only the average statistics are known; in these cases, the PPIC cannot be computed and one of the other ICs must be used.

In the context of data subset selection, the BPIC in particular leads to an especially aggressive penalty for data cuts, which may result in significant finite-sample bias as seen in the numerical results of Secs. VI B and VI C. Based on these results, we recommend against the use of BPIC in particular for problems where data subset selection is an important part of the model variation problem. Both PPIC and BAIC should be used for subset variation, with preference for the PPIC when possible.

While the BAIC, BPIC, and PPIC perform differently in practice, this is merely an artifact of finite sample sizes. Being asymptotically unbiased estimators of the KL divergence, all three ICs will become equivalent in the limit of infinite data. For this reason, it is logical to consider the BPIC and PPIC as finite- N corrections to the BAIC. To better understand the behavior at finite N , it would be interesting in future work to consider higher-order corrections to the ICs; however, any corrections beyond $o(N^{-1})$ should be accompanied by a more detailed study of the finite- N bias corrections. Higher order bias correction

similar to that of the so-called “corrected AIC” (abbreviated as AIC_C in the literature) could be a fruitful direction for future improvements [38,82–85]. Studies of nonasymptotic bias corrections are rare in the model selection literature but could be theoretically interesting, albeit practically challenging.

It should be noted that in all of the above cases, we make use of the approximate result $\text{tr}[J^{-1}I] \rightarrow k$. This is an asymptotic result that only holds if the true distribution belongs to the model space, as discussed in Sec. III A. Use of the more general formula requires estimation of the I and J matrices, which can be numerically unstable particularly at smaller values of N . An interesting direction for future work could be to explore more reliable numerical methods for estimation of this trace, such as shrinkage [20,68–71].

In our numerical studies of scaling of model averaging with sample size N , we sometimes encounter situations where increasing N results in larger model-average uncertainty. This effect is sharply counterintuitive, since most familiar statistical error estimators (e.g., the standard error on a parameter obtained from a single model) decrease monotonically with N . We interpret this effect as a manifestation of bias-variance trade-off (see [28] for a general discussion of this phenomenon, and our own discussion in Sec. II B.) Model averaging typically results in increased uncertainty compared to the use of a single, fixed model. This is a feature, as the increased error reflects systematic error due to model uncertainty that is neglected in the fixed-model case. Removing this potential source of bias with model-averaging results in increased error. This trade-off, combined with the discreteness of the model space, sometimes leads to complicated and counterintuitive behavior of uncertainty versus sample size. Of course, it remains true that all of the individual model parameter estimates that go into the model average do have errors which decrease monotonically with sample size N ; therefore, any large enough increase in N must always tend to reduce the model-averaged error as well.

It would be very interesting to explore the use of resampling techniques such as bootstrap estimation with model averaging. For example, bootstrap methods could be used to directly estimate IC bias at finite sample size, possibly leading to improved model-averaging performance at finite N . Improvements to bias estimation could give better control over the bias-variance trade-off. It may also be interesting to explore whether direct bootstrap estimation of the KL divergence itself, as in [86] for example, might lead to useful insights or practical improvements for model averaging.

While we have provided insight into the underlying statistical theory at play (e.g., the KL divergence), another objective has been to give results that are useful in practical applications. While our results are very general and can be applied to a wide array of fields, we are primarily motivated by applications to lattice field theory. Bayesian model averaging is well-suited for lattice applications due to the

physical motivations and relevant functional forms of lattice models. Furthermore, the practical need for model and data truncation in lattice analyses fits naturally into the Bayesian model-averaging framework.

As a brief aside, we should mention that some believe that the discrete optimization problem of model-averaging (or model selection) is not a proper usage of Bayesian inference. For instance, [87] advocates for continuous model expansion, i.e., forming a larger model that includes the successful candidate models as special cases. This procedure poses obvious practical limitations as it continues *ad infinitum* and could hide interesting physical insight within a large and complicated model. Despite philosophical qualms with model comparison, [87] agrees that model averaging is still a useful technique given a finite amount of available information. Furthermore, Bayesian model averaging is a natural procedure when the model space is truly discrete as is the case for data subset selection in the analysis of lattice simulation data. That being said, it could be interesting to incorporate the idea of continuity into lattice analyses such as with Bayesian mixture models [88].

We conclude with some practical disclaimers. A notable advantage of model averaging over model selection is the removal of subjectivity from the analysis (e.g., in the analyst’s choice of data subset to fit to a two-point correlator). To this end, we recommend the use of uniform parameter priors $\text{pr}(M_\mu)$ rather than, say, weighting parsimonious models more heavily as this is built into the ICs. On the other hand, the factor $\text{pr}(M_\mu)$ should not be ignored in cases where there is good reason to include it. For example, in cases where there is a strong theoretical reason *not* to include a particular model, i.e., $\text{pr}(M_\mu) \approx 0$, this should be reflected in the model priors rather than relying on the data entirely. The model prior can also be used to adjust for situations where a uniform prior would result in bias; for example, if a family of highly similar models are included in the model space, a uniform prior may overweight this class of models solely due to the number of variations in the family.

We emphasize finally that model averaging is not an alternative to a statistically correct treatment of the data. While we have seen the PPIC to be somewhat robust against these effects, including incorrect results (e.g., failed fits, excessively small signal-to-noise ratios, autocorrelation effects) in the average can invalidate statistical estimates.

ACKNOWLEDGMENTS

We give special thanks to the MIT lattice group for allowing us to make use of their nucleon two-point function data, generated using computing resources provided by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; the Oak Ridge Leadership Computing Facility at the Oak Ridge

National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725; the USQCD Collaboration, which are funded by the Office of Science of the U.S. Department of Energy; and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin. We thank Daniel Hackett, George Fleming, and Will Jay for helpful conversations. This work was supported in part by the U.S. Department of Energy (DOE), Office of Science, Office of High Energy Physics, under Award Number DE-SC0010005.

APPENDIX A: MARGINALIZED INFORMATION CRITERIA

In this appendix, we connect this work to [26] and discuss the ABIC_{CV} in more detail. As discussed in Sec. III, there are many ways to analyze parametric models using the KL divergence. While the simplest is the plug-in estimator introduced in Sec. III A, a natural alternative is to marginalize over the parameter space giving the marginalized KL divergence KL_{marg} defined in Eq. (36). The corresponding information criteria is the ABIC [39,40], defined as

$$\text{ABIC}_{\mu} = -2 \log \text{pr}(M_{\mu}) - 2 \log E_{\mathbf{a}}[\text{pr}(\{y\}|\mathbf{a}, M_{\mu})] + 2k, \quad (\text{A1})$$

where we have written the marginalization with $E_{\mathbf{a}}[\dots]$, the expectation with respect to the prior distribution as defined in Eq. (38).

In [26], the ABIC is applied to the cases of least-squares regression (see Sec. IV for notation). Computing the integral to leading order via Laplace's method gives

$$\text{ABIC}_{\mu} \approx -2 \log \text{pr}(M_{\mu}) + \chi_{\text{aug}}^2(\mathbf{a}^*) + 2k - 2 \log \frac{\det \tilde{\Sigma}}{\det \Sigma^*}, \quad (\text{A2})$$

where $\tilde{\Sigma}$ and Σ^* are the prior and best-fit covariance matrices, respectively. To argue that the determinant terms are subleading in the limit of large sample size, [26] consider the case of cross-validation, where the prior probabilities scale with the size of the data sample. In this cases, the ABIC reduces to the “ ABIC_{CV} ” (referred to as simply the AIC in [26]) defined in Eq. (66) for least squares or more generally in Eq. (40).

This issue with this derivation is that the determinant terms can be significant in some cases, such as when priors are held fixed as is the case for the examples in Sec. VI. In fact, a similar derivation [38] that retains the N dependence of $\det \tilde{\Sigma} / \det \Sigma^*$ leads to Schwarz's BIC [41]:

$$\text{BIC}_{\mu} = -2 \log \text{pr}(M_{\mu}) - 2 \log \text{pr}(\{y\}|\mathbf{a}^*, M_{\mu}) + \log(N)k, \quad (\text{A3})$$

which can be thought of as an alternative simplification to the ABIC_{CV} . The BIC will clearly behave much differently from the BAIC for large N .

Though we do not show this rigorously here, the ABIC does not share the asymptotic unbiasedness property of the information criteria studied in the body of the text. Therefore, it can give asymptotically biased model averaging results when some models in the space of candidate models poorly reflect the data. This is a result of the outsized dependence of the ABIC on the prior as opposed to the data. The BAIC, BPIC, PAIC, and PPIC all depend directly on the data either through the plug-in estimator \mathbf{a}^* or through the posterior average $E_{\mathbf{a}|\{y\}}[\dots]$. In contrast, the ABIC uses an expectation over the prior. As a result, the ABIC may behave counterintuitively even in the limit of infinite data. This behavior is related to the so-called “Jeffreys-Lindley paradox” in which a Bayesian analysis under certain choices of prior distribution can give incorrect results compared to, say, the analogous frequentist analysis. For further discussion, see [89–91].

APPENDIX B: ASYMPTOTIC EQUIVALENCE OF INFORMATION CRITERIA

In this appendix, we compare the asymptotic forms of the different KL divergences defined in Sec. III. In doing so, we will establish that each form of the KL divergence is asymptotically equivalent in the sense of model choice (e.g., model selection, model averaging). This is weaker than asserting that the different forms of the KL divergence approach each other numerically and only requires convergence of the leading-order terms in respective asymptotic expansions in the inverse sample size N^{-1} .

Reproducing only the model-dependent terms here in a convenient form, and omitting the expectation value $E_z[\dots]$ and the model choice M_{μ} (we want to compare them for a fixed model), we have

$$\text{KL}_{\text{plug-in}} \supset \log \text{pr}(z|\mathbf{a}^*), \quad (\text{B1})$$

$$\text{KL}_{\text{post-avg}} \supset E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a})], \quad (\text{B2})$$

$$\text{KL}_{\text{post-pred}} \supset \log E_{\mathbf{a}|\{y\}}[\text{pr}(z|\mathbf{a})], \quad (\text{B3})$$

where

$$\begin{aligned} E_{\mathbf{a}|\{y\}}[\dots] &\equiv \frac{\int d\mathbf{a} \text{pr}(\mathbf{a}|\{y\})(\dots)}{\int d\mathbf{a} \text{pr}(\mathbf{a}|\{y\})} \\ &= \frac{\int d\mathbf{a} \text{pr}(\{y\}|\mathbf{a})\text{pr}(\mathbf{a})(\dots)}{\int d\mathbf{a} \text{pr}(\{y\}|\mathbf{a})\text{pr}(\mathbf{a})}, \end{aligned} \quad (\text{B4})$$

and $\text{pr}(\mathbf{a})$ is the prior probability distribution for the given model. Throughout this appendix, we will assume that the prior information grows sufficiently slowly with N

compared to the data; specifically, we assume that Eq. (27) holds (e.g., fixed priors without data dependence).

To proceed, note that for any continuous f

$$\lim_{N \rightarrow \infty} \frac{\int d\mathbf{a} \, \text{pr}(\{y\}|\mathbf{a}) \text{pr}(\mathbf{a}) f(\mathbf{a})}{(2\pi)^{k/2} |J_N^{-1}(a^*)|^{1/2} \text{pr}(\mathbf{a}^*) f(\mathbf{a}^*)} = 1, \quad (\text{B5})$$

which follows from convergence of the leading-order Laplace approximation. This is a standard result proven in many texts (e.g., [50]). A more detailed discussion of the Laplace approximation is given in Appendix E. Therefore,

$$E_{\mathbf{a}|\{y\}}[\log \text{pr}(z|\mathbf{a})] = \frac{\int d\mathbf{a} \, \text{pr}(\{y\}|\mathbf{a}) \text{pr}(\mathbf{a}) \log \text{pr}(z|\mathbf{a})}{\int d\mathbf{a} \, \text{pr}(\{y\}|\mathbf{a}) \text{pr}(\mathbf{a})} \quad (\text{B6})$$

$$\rightarrow \frac{(2\pi)^{k/2} |J_N^{-1}(a^*)|^{1/2} \text{pr}(\mathbf{a}^*) \log \text{pr}(z|\mathbf{a}^*)}{(2\pi)^{k/2} |J_N^{-1}(a^*)|^{1/2} \text{pr}(\mathbf{a}^*)} = \log \text{pr}(z|\mathbf{a}^*), \quad (\text{B7})$$

$$\log E_{\mathbf{a}|\{y\}}[\text{pr}(z|\mathbf{a})] = \frac{\int d\mathbf{a} \, \text{pr}(\{y\}|\mathbf{a}) \text{pr}(\mathbf{a}) \text{pr}(z|\mathbf{a})}{\int d\mathbf{a} \, \text{pr}(\{y\}|\mathbf{a}) \text{pr}(\mathbf{a})} \quad (\text{B8})$$

$$\rightarrow \log \frac{(2\pi)^{k/2} |J_N^{-1}(a^*)|^{1/2} \text{pr}(\mathbf{a}^*) \text{pr}(z|\mathbf{a}^*)}{(2\pi)^{k/2} |J_N^{-1}(a^*)|^{1/2} \text{pr}(\mathbf{a}^*)} = \log \text{pr}(z|\mathbf{a}^*), \quad (\text{B9})$$

in the $N \rightarrow \infty$ limit. This establishes the asymptotic equivalence of $\text{KL}_{\text{post-avg}}$ and $\text{KL}_{\text{post-pred}}$ to $\text{KL}_{\text{plug-in}}$. Since the various ICs we considered follow directly from the KL divergence definitions, this establishes that both the PPIC and the PAIC/BPIC converge asymptotically to the BTIC/BAIC.

Similarly to Sec. IV C, this result is established using Laplace's method. However, we emphasize that this equivalence is much more generally applicable than to just the case of least-squares discussed in Sec. IV C. See [50,92] for the required regularity conditions.

In effect, the above argument indicates

$$\lim_{N \rightarrow \infty} \text{pr}(\mathbf{a}|\{y\}) \propto \delta(\mathbf{a} - \mathbf{a}^*). \quad (\text{B10})$$

To understand this we note that both $\text{KL}_{\text{post-avg}}$ and $\text{KL}_{\text{post-pred}}$ contain the posterior probability

$$\log \text{pr}(\mathbf{a}|\{y\}) \propto \sum_i \log [\text{pr}(y_i|\mathbf{a}) \text{pr}(\mathbf{a})^{1/N}]. \quad (\text{B11})$$

Since we have assumed Eq. (27), the influence of the prior is negligible compared to that of the data as $N \rightarrow \infty$, so that this simply approaches the log likelihood function $\sum_i \log \text{pr}(y_i|\mathbf{a})$. In terms of the parameters, the likelihood function becomes Gaussian asymptotically with a width decreasing proportional to N . As a result, we have the proportionality in Eq. (B10). This connection is not made

rigorous here, but is expected to hold except in pathological cases.

APPENDIX C: A BOUND ON THE ASYMPTOTIC BIAS OF MODEL AVERAGING

Here we derive the bound on the model averaging asymptotic bias given in Eq. (16). For a general discussion of asymptotic bias and the relevant notation, see Sec. II B. First, it will be useful to introduce some new notation for this appendix. Specifically, dependencies on the sample size will be shown explicitly with a subscript N . When the subscript N is absent, this denotes the asymptotic value (e.g., $A \equiv \lim_{N \rightarrow \infty} A_N$ is the asymptotic value for a sequence of sample estimators $\{A_N\}_{N \in \mathbb{N}}$). The one exception is the sample data $\{y\}$, for which the N dependence is clear and the asymptotic value (a random variable drawn from pr_T) is denoted by z . For simplicity, we do not distinguish between b_y (finite-sample bias) and b_z (asymptotic bias) as defined in Eqs. (13) and (14) in this appendix, as making the N -dependence explicit is sufficient.

The bound in Eq. (16) holds (with probability 1) if the parameter estimation procedure is consistent. The sequence of sample estimators $\{X(\{y\})\}_{N \in \mathbb{N}}$ of ξ is consistent if it satisfies [27]

$$\lim_{N \rightarrow \infty} \text{pr}(|X(\{y\}) - \xi| > \epsilon) = 0, \quad (\text{C1})$$

for any $\epsilon > 0$. This form of consistency is also known as weak consistency, in contrast to strong consistency where $\{X(\{y\})\}_{N \in \mathbb{N}}$ satisfies [93]

$$\text{pr}(\lim_{N \rightarrow \infty} X(\{y\}) = \xi) = 1. \quad (\text{C2})$$

Weak consistency is defined using convergence in probability whereas strong consistency is defined using almost sure convergence. Since almost sure convergence implies convergence in probability, strong consistency implies weak consistency. Another related notion is convergence in the sense of distributions, which is implied by convergence in probability; convergence in probability and convergence in the sense of distributions are equivalent if the limiting random variable $X(z)$ is a constant.

As discussed in Sec. II B, our primary goal is to remove asymptotic bias from the model average parameter estimates. For concreteness, consider the bias of a single parameter a_0 given by

$$b_z[\langle a_0 \rangle_N] = E_z[\langle a_0 \rangle_N] - a_{0,T}^*. \quad (\text{C3})$$

By Eq. (5), we have

$$\langle a_0 \rangle_N = a_{0,T,N}^* \text{pr}(M_T|\{y\}) + \sum_{\mu \neq T} a_{0,\mu,N}^* \text{pr}(M_\mu|\{y\}), \quad (\text{C4})$$

where $\text{pr}(M_\mu|\{y\})$ satisfy Eq. (15).

We wish to derive a bound on $b_z[\langle a_0 \rangle_N]$ in terms of the asymptotic bias in the model weights. To that end, observe that

$$|b_z[\langle a_0 \rangle_N]| = |E_z[\langle a_0 \rangle_N] - a_{0,T}^*| \quad (\text{C5})$$

$$= \left| \sum_{\mu} \{E_z[a_{0,\mu,N}^* \text{pr}(M_{\mu}|\{y\})] - a_{0,\mu}^* \text{pr}(M_{\mu}|z)\} \right| \quad (\text{C6})$$

$$= \left| \sum_{\mu} \{E_z[(a_{0,\mu,N}^* - a_{0,\mu}^*) \text{pr}(M_{\mu}|\{y\})] + a_{0,\mu}^* b_z[\text{pr}(M_{\mu}|\{y\})]\} \right| \quad (\text{C7})$$

$$\leq \sum_{\mu} \{|E_z[(a_{0,\mu,N}^* - a_{0,\mu}^*) \text{pr}(M_{\mu}|\{y\})]| + |a_{0,\mu}^*| |b_z[\text{pr}(M_{\mu}|\{y\})]|\}. \quad (\text{C8})$$

Let $\epsilon > 0$ and observe that

$$|E_z[(a_{0,\mu,N}^* - a_{0,\mu}^*) \text{pr}(M_{\mu}|\{y\})]| = \left| \int dF_{M_T}(z) (a_{0,\mu,N}^* - a_{0,\mu}^*) \text{pr}(M_{\mu}|\{y\}) \right| \quad (\text{C9})$$

$$\leq \int dF_{M_T}(z) |a_{0,\mu,N}^* - a_{0,\mu}^*| \text{pr}(M_{\mu}|\{y\}) \quad (\text{C10})$$

$$\leq \int dF_{M_T}(z) |a_{0,\mu,N}^* - a_{0,\mu}^*| \quad (\text{C11})$$

$$= \int_{\Omega_{N,\epsilon}} dF_{M_T}(z) |a_{0,\mu,N}^* - a_{0,\mu}^*| + \int_{\Omega_{N,\epsilon}^c} dF_{M_T}(z) |a_{0,\mu,N}^* - a_{0,\mu}^*| \quad (\text{C12})$$

$$\leq \|a_{0,\mu,N}^* - a_{0,\mu}^*\|_{L^\infty(\Omega_{N,\epsilon})} F_{M_T}(\Omega_{N,\epsilon}) + \|a_{0,\mu,N}^* - a_{0,\mu}^*\|_{L^\infty(\Omega_{N,\epsilon}^c)} F_{M_T}(\Omega_{N,\epsilon}^c), \quad (\text{C13})$$

where $\Omega_{N,\epsilon} = \{z \in \mathbb{R}^d : |a_{0,\mu,N}^* - a_{0,\mu}^*| \leq \epsilon\}$ and $\Omega_{N,\epsilon}^c = \{z \in \mathbb{R}^d : |a_{0,\mu,N}^* - a_{0,\mu}^*| > \epsilon\}$ is its complement. In Eq. (C13), $\|\dots\|_{L^\infty(\Omega)}$ denotes the L^∞ -norm with respect to the true probability measure $F_{M_T}(z)$ [defined in Eq. (17)] on the set Ω , i.e.,

$$\|f\|_{L^\infty(\Omega)} \equiv \inf\{M \in \mathbb{R} : |f(\Omega)| \leq M \text{ almost surely with respect to } F_{M_T}(z)\} \quad (\text{C14})$$

$$= \text{ess sup}_{z \in \Omega} (|f(z)|), \quad (\text{C15})$$

where ‘‘ess sup’’ denotes the essential supremum with respect to $F_{M_T}(z)$. For simplicity, we have also adopted the notation that the measure of a set Ω is denoted by

$$F_{M_T}(\Omega) \equiv \int_{\Omega} dF_{M_T}(z). \quad (\text{C16})$$

To proceed, we assume the parameter estimation procedure is weakly consistent, i.e.,

$$\lim_{N \rightarrow \infty} \text{pr}(|a_{0,\mu,N}^* - a_{0,\mu}^*| > \epsilon) = 0. \quad (\text{C17})$$

This will be true for, say, least-squares regression, which is in fact strongly consistent in some cases [94,95]. It follows from this assumption that, in the large N limit, $\Omega_{N,\epsilon}$ has unit measure and $\Omega_{N,\epsilon}^c$ has measure zero. Therefore,

$$\lim_{N \rightarrow \infty} |E_z[(a_{0,\mu,N}^* - a_{0,\mu}^*) \text{pr}(M_{\mu}|\{y\})]| \leq \lim_{N \rightarrow \infty} \|a_{0,\mu,N}^* - a_{0,\mu}^*\|_{L^\infty(\Omega_{N,\epsilon})} \leq \epsilon. \quad (\text{C18})$$

Taking ϵ arbitrarily small, Eq. (C8) gives the following bound on the asymptotic bias:

$$|b_z[\langle a_0 \rangle]| = \lim_{N \rightarrow \infty} |b_z[\langle a_0 \rangle_N]| \leq \sum_{\mu} |a_{0,\mu}^*| \lim_{N \rightarrow \infty} |b_z[\text{pr}(M_{\mu}|\{y\})]| \quad (\text{C19})$$

$$= \sum_{\mu} |a_{0,\mu}^*| |b_z[\text{pr}(M_{\mu}|z)]|, \quad (\text{C20})$$

which holds with probability 1. Using Eq. (5), a similar argument holds for $\langle f(\mathbf{a}) \rangle$ giving the bound in Eq. (16).

While this derivation seems rather technical, it is related to Hölder's inequality [45]:

$$\left| \int_{\Omega} d\mu f g \right| \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}, \quad (\text{C21})$$

where $(\Omega, \mathcal{S}, \mu)$ is a measure space (\mathcal{S} being some σ -algebra on Ω), $f, g \in L^1(\Omega)$, $f \in L^p(\Omega)$, and $g \in L^q(\Omega)$ with respect to μ . We have assumed that $1 \leq p, q \leq \infty$ satisfy $1/p + 1/q = 1$ (i.e., p and q are Hölder conjugates); “ $1/\infty$ ” is defined as zero in this context. We are able to achieve a somewhat sharper bound on $|E_z[(a_{0,\mu,N}^* - a_{0,\mu}^*) \text{pr}(M_{\mu}|\{y\})]|$ using the fact that $\text{pr}(M_{\mu}|\{y\})$ is bounded between zero and one and assuming consistency (along with partitioning Ω as above).

APPENDIX D: FORMULAS FOR THE LEAST SQUARES PAIC

In this appendix, we discuss approximations to Eq. (80)—the PAIC for least-squares regression—in the various cases of interest and the practical complications that arise.

Similarly to the BPIC in Sec. IV D and the PPIC in Sec. IV E, the PAIC can be approximated by

$$\text{PAIC}_{\mu} \approx \hat{\chi}^2(\mathbf{a}^*) + \frac{1}{2} \hat{H}_{ba}(\Sigma^*)_{ab} - \frac{1}{2} \hat{g}_d T_{cba}(\Sigma^*)_{abcd} + 2k, \quad (\text{D1})$$

where we have assumed that the correct model is in the family of candidates so that $\text{tr}[J^{-1}(\mathbf{a}^*)I(\mathbf{a}^*)] \rightarrow k$ and

$$\hat{g}_a \equiv \left. \frac{\partial \hat{\chi}^2}{\partial a_a} \right|_{\mathbf{a}=\mathbf{a}^*}, \quad \hat{H}_{ab} \equiv \left. \frac{\partial^2 \hat{\chi}^2}{\partial a_a \partial a_b} \right|_{\mathbf{a}=\mathbf{a}^*}. \quad (\text{D2})$$

One might suspect that Eq. (D1) is an NLO asymptotic expansion in the inverse sample size N^{-1} . However, there are some subtleties in the power counting for nominally $O(1)$ terms $\frac{1}{2} \hat{H}_{ba}(\Sigma^*)_{ab}$ and $-\frac{1}{2} \hat{g}_d T_{cba}(\Sigma^*)_{abcd}$. First, note that

$$\frac{1}{2} \text{tr}[\hat{H}\Sigma^*] = \text{tr} \left[\left(\Sigma^{*-1} - \frac{1}{2} \hat{H} \right) \Sigma^* \right] = k - \frac{1}{2} \text{tr}[\hat{H}\Sigma^*], \quad (\text{D3})$$

where the substitution for \hat{H} follows from the definition $\chi_{\text{aug}}^2 = \hat{\chi}^2 + \tilde{\chi}^2$ and from Eq. (69). So, we see that the second term in Eq. (D1) includes an $O(N^{-1})$ contribution $-\frac{1}{2} \tilde{H}_{ba}(\Sigma^*)_{ab}$.

Second, \mathbf{a}^* is chosen so that the gradient of $\chi_{\text{aug}}^2(\mathbf{a})$ vanishes. It follows that $-\hat{g} = \tilde{g}$, and thus

$$-\frac{1}{2} \hat{g}_d T_{cba}(\Sigma^*)_{abcd} = \frac{1}{2} \tilde{g}_d T_{cba}(\Sigma^*)_{abcd}, \quad (\text{D4})$$

giving rise to another $O(N^{-1})$ term.

Therefore,

$$\text{PAIC}_{\mu} = \hat{\chi}^2(\mathbf{a}^*) + 3k + O(N^{-1}). \quad (\text{D5})$$

In applying Laplace's method to Eq. (80)—an integral with $O(N)$ integrand—we have already neglected some $O(N^{-1})$ terms [cf. Eq. (E6)]. So, we again must neglect the $O(N^{-1})$ terms in order to maintain a consistent $O(1)$ asymptotic expansion giving

$$\text{PAIC}_{\mu} \approx \hat{\chi}^2(\mathbf{a}^*) + 3k. \quad (\text{D6})$$

The corresponding superasymptotic expansion is

$$\text{PAIC}_{\mu} \approx \begin{cases} \hat{\chi}^2(\mathbf{a}^*) + 3k, & k < \hat{\chi}^2(\mathbf{a}^*), \\ \hat{\chi}^2(\mathbf{a}^*) + 2k, & \text{otherwise,} \end{cases} \quad (\text{D7})$$

which should be used for nonlinear least squares. For linear least squares, the NLO expansion is exact and Eq. (D1) can be used (with $T = 0$), which is also identical to the form of the BPIC in Eq. (131).

For data subset selection, the derivation is similar to that of the BPIC in Sec. V; the PAIC is modified simply by the addition of a factor of $3d_C$ for the cut portion of the data, i.e.

$$\text{PAIC}_{\mu,P} = \text{PAIC}_{\mu} + 3d_C \quad (\text{D8})$$

with PAIC_{μ} given by the superasymptotic formula in Eq. (D7). For use in model averaging, Eq. (11) applies.

APPENDIX E: LAPLACE'S METHOD AND GAUSSIAN INTEGRAL FORMULAS

Here we summarize the derivation of the NLO Laplace approximation applied to the integrals needed for the

various information criteria appearing in the body of the paper in the case of least squares. A more rigorous treatment of Laplace method can be found in many texts on asymptotics [96,97], and generalizations appear in the asymptotics literature [50].

Here we consider integrals of the form

$$\mathcal{I}[\psi] = \int d\mathbf{a} \exp\left[-\frac{1}{2}\chi_{\text{aug}}^2(\mathbf{a})\right] \psi(\mathbf{a}). \quad (\text{E1})$$

In the limit of large sample size, $\chi_{\text{aug}}^2(\mathbf{a}) = O(N)$. In the $N \rightarrow \infty$ limit, the integrand becomes sharply peaked about the best-fit parameter $\mathbf{a}^* = \text{argmin}_{\mathbf{a}} \chi_{\text{aug}}^2(\mathbf{a})$. Assuming for simplicity that \mathbf{a}^* is on the interior of the parameter space

(as is the case for the examples considered above), the integral becomes localized to $B_\epsilon(\mathbf{a}^*) = \{\mathbf{a} : \|\mathbf{a} - \mathbf{a}^*\| < \epsilon\}$, the ball of radius $\epsilon > 0$ centered at \mathbf{a}^* :

$$\mathcal{I}[\psi] \approx \int_{\mathbf{a} \in B_\epsilon(\mathbf{a}^*)} d\mathbf{a} \exp\left[-\frac{1}{2}\chi_{\text{aug}}^2(\mathbf{a})\right] \psi(\mathbf{a}). \quad (\text{E2})$$

Assuming ϵ is small, we can use truncated Taylor expansions about $\|\mathbf{a} - \mathbf{a}^*\| = 0$. To obtain the leading-order integral expansion, we would Taylor expand χ_{aug}^2 and ψ to $O(\|\mathbf{a} - \mathbf{a}^*\|^2)$ and $O(\|\mathbf{a} - \mathbf{a}^*\|^0)$, respectively (see [26] for example). For the NLO integral expansion, we need to include two more terms in both Taylor expansions (cf. [96,97]):

$$\chi_{\text{aug}}^2(\mathbf{a}) = \chi_{\text{aug}}^2(\mathbf{a}^*) + (\Sigma^{*-1})_{ba} \delta_a \delta_b + T_{cba} \delta_a \delta_b \delta_c + F_{dcba} \delta_a \delta_b \delta_c \delta_d + O(\|\delta\|^5), \quad (\text{E3})$$

$$\psi(\mathbf{a}) = \psi(\mathbf{a}^*) + g_a \delta_a + \frac{1}{2} H_{ba} \delta_a \delta_b + O(\|\delta\|^3), \quad (\text{E4})$$

where $\delta \equiv \mathbf{a} - \mathbf{a}^*$, the inverse parameter covariance matrix Σ^* is defined in Eq. (69), and the remaining tensors T , F , g , H are defined in Eq. (71) and (72). Substituting Eq. (E3) and (E4) into Eq. (E2) gives

$$\begin{aligned} \mathcal{I}[\psi] &\approx \int_{\delta \in B_\epsilon(0)} d\delta \exp\left[-\frac{1}{2}(\chi_{\text{aug}}^2(\mathbf{a}^*) + (\Sigma^{*-1})_{ba} \delta_a \delta_b + T_{cba} \delta_a \delta_b \delta_c + F_{dcba} \delta_a \delta_b \delta_c \delta_d)\right] \\ &\quad \times \left(\psi(\mathbf{a}^*) + g_a \delta_a + \frac{1}{2} H_{ba} \delta_a \delta_b\right) \end{aligned} \quad (\text{E5})$$

$$\begin{aligned} &\approx \int_{\delta \in B_\epsilon(0)} d\delta \exp\left[-\frac{1}{2}(\chi_{\text{aug}}^2(\mathbf{a}^*) + (\Sigma^{*-1})_{ba} \delta_a \delta_b)\right] \\ &\quad \times \left(\psi(\mathbf{a}^*) + \frac{1}{2} H_{ba} \delta_a \delta_b - \frac{1}{2} g_d T_{cba} \delta_a \delta_b \delta_c \delta_d - \frac{1}{2} \psi(\mathbf{a}^*) F_{dcba} \delta_a \delta_b \delta_c \delta_d \right. \\ &\quad \left. + \frac{1}{8} \psi(\mathbf{a}^*) T_{fed} T_{cba} \delta_a \delta_b \delta_c \delta_d \delta_e \delta_f\right), \end{aligned} \quad (\text{E6})$$

where the second approximation is obtained by Taylor expanding the highest order terms of the exponential $\exp[-\frac{1}{2}(T_{cba} \delta_a \delta_b \delta_c + F_{dcba} \delta_a \delta_b \delta_c \delta_d)]$ about $\delta = \mathbf{0}$, neglecting terms that only contribute to NNLO, and omitting odd terms that do not contribute to the integral. Again using the fact that the integral is sharply peaked about $\delta = \mathbf{0}$ in the $N \rightarrow \infty$ limit, expanding the domain of integral to all of \mathbb{R}^k introduces only a small error. After doing so, we are left with

$$\begin{aligned} \mathcal{I}[\psi] &\approx \int_{\mathbb{R}^k} d\delta \exp\left[-\frac{1}{2}(\chi_{\text{aug}}^2(\mathbf{a}^*) + (\Sigma^{*-1})_{ba} \delta_a \delta_b)\right] \\ &\quad \times \left(\psi(\mathbf{a}^*) + \frac{1}{2} H_{ba} \delta_a \delta_b - \frac{1}{2} g_d T_{cba} \delta_a \delta_b \delta_c \delta_d - \frac{1}{2} \psi(\mathbf{a}^*) F_{dcba} \delta_a \delta_b \delta_c \delta_d \right. \\ &\quad \left. + \frac{1}{8} \psi(\mathbf{a}^*) T_{fed} T_{cba} \delta_a \delta_b \delta_c \delta_d \delta_e \delta_f\right). \end{aligned} \quad (\text{E7})$$

After extending the domain, each term of the integral is proportional to one of the following Gaussian integrals:

$$\int_{\mathbb{R}^k} d\delta \exp\left[-\frac{1}{2}(\Sigma^{*-1})_{ba} \delta_a \delta_b\right] = (2\pi)^{k/2} (\det \Sigma^*)^{1/2}, \quad (\text{E8})$$

$$\int_{\mathbb{R}^k} d\delta \exp \left[-\frac{1}{2} (\Sigma^{*-1})_{ba} \delta_a \delta_b \right] \delta_a \delta_b = (2\pi)^{k/2} (\det \Sigma^*)^{1/2} (\Sigma^*)_{ab}, \quad (\text{E9})$$

$$\int_{\mathbb{R}^k} d\delta \exp \left[-\frac{1}{2} (\Sigma^{*-1})_{ba} \delta_a \delta_b \right] \delta_a \delta_b \delta_c \delta_d = (2\pi)^{k/2} (\det \Sigma^*)^{1/2} (\Sigma_2^*)_{abcd}, \quad (\text{E10})$$

$$\int_{\mathbb{R}^k} d\delta \exp \left[-\frac{1}{2} (\Sigma^{*-1})_{ba} \delta_a \delta_b \right] \delta_a \delta_b \delta_c \delta_d \delta_e \delta_f = (2\pi)^{k/2} (\det \Sigma^*)^{1/2} (\Sigma_3^*)_{abcdef}, \quad (\text{E11})$$

where the high-order contractions of the covariance matrix are defined in Eq. (70). Using these integral identities, we obtain Eq. (68):

$$\begin{aligned} \mathcal{I}[\psi] &\approx (2\pi)^{k/2} |\Sigma^*|^{1/2} \exp \left[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a}^*) \right] \\ &\times \left(\psi(\mathbf{a}^*) + \frac{1}{2} H_{ba}(\Sigma^*)_{ab} - \frac{1}{2} g_d T_{cba}(\Sigma_2^*)_{abcd} - \frac{1}{2} \psi(\mathbf{a}^*) F_{dcba}(\Sigma_2^*)_{abcd} \right. \\ &\left. + \frac{1}{8} \psi(\mathbf{a}^*) T_{fed} T_{cba}(\Sigma_3^*)_{abcdef} \right). \end{aligned} \quad (\text{E12})$$

We will also need to consider the case were $\mathcal{I}[\psi]$ is normalized by $\mathcal{I}[1]$:

$$\frac{\mathcal{I}[\psi]}{\mathcal{I}[1]} = \frac{\int d\mathbf{a} \exp \left[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a}) \right] \psi(\mathbf{a})}{\int d\mathbf{a} \exp \left[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a}) \right]}. \quad (\text{E13})$$

To approximate ratios such as this, we first apply the NLO Laplace approximation Eq. (E12) to both the numerator and denominator to obtain

$$\frac{\mathcal{I}[\psi]}{\mathcal{I}[1]} \approx \frac{\psi(\mathbf{a}^*) + \frac{1}{2} H \Sigma^* - \frac{1}{2} g T \Sigma_2^* - \frac{1}{2} \psi(\mathbf{a}^*) F \Sigma_2^* + \frac{1}{8} \psi(\mathbf{a}^*) T T \Sigma_3^*}{1 - \frac{1}{2} F \Sigma_2^* + \frac{1}{8} T T \Sigma_3^*}, \quad (\text{E14})$$

where tensor indices are suppressed for simplicity. By the tensor power counting summarized at the end of Sec. IV C, we can expand the denominator as a geometric series; this expansion will maintain the same order of accuracy and is known in the probability literature [51]. Keeping enough terms to maintain an overall NLO approximation, we obtain

$$\frac{\mathcal{I}[\psi]}{\mathcal{I}[1]} \approx \left(\psi(\mathbf{a}^*) + \frac{1}{2} H \Sigma^* - \frac{1}{2} g T \Sigma_2^* - \frac{1}{2} \psi(\mathbf{a}^*) F \Sigma_2^* + \frac{1}{8} \psi(\mathbf{a}^*) T T \Sigma_3^* \right) \left(1 + \frac{1}{2} F \Sigma_2^* - \frac{1}{8} T T \Sigma_3^* \right) \quad (\text{E15})$$

$$\approx \psi(\mathbf{a}^*) \left(1 + \frac{1}{2} F \Sigma_2^* - \frac{1}{8} T T \Sigma_3^* \right) + \frac{1}{2} H \Sigma^* - \frac{1}{2} g T \Sigma_2^* - \frac{1}{2} \psi(\mathbf{a}^*) F \Sigma_2^* + \frac{1}{8} \psi(\mathbf{a}^*) T T \Sigma_3^* \quad (\text{E16})$$

$$= \psi(\mathbf{a}^*) + \frac{1}{2} H_{ba}(\Sigma^*)_{ab} - \frac{1}{2} g_d T_{cba}(\Sigma_2^*)_{abcd}, \quad (\text{E17})$$

where we have restored the indices in the last line.

Outside the context of the Laplace method, we also make use of some additional Gaussian integrals in the exact treatment of perfect model KL divergences in Sec. V. Consider an integral of the following form:

$$\mathcal{J}_1 \equiv \int_{\mathbb{R}^k} d\mathbf{a} \exp \left[-\frac{1}{2} (\mu_0 - \mathbf{a})^T (\Sigma_0^{-1}) (\mu_0 - \mathbf{a}) \right] (\mu_1 - \mathbf{a})^T \Sigma_1^{-1} (\mu_1 - \mathbf{a}). \quad (\text{E18})$$

Defining the change of variables

$$\delta \equiv \mathbf{a} - \mu_0, \quad (\text{E19})$$

$$\xi \equiv \mu_0 - \mu_1, \quad (\text{E20})$$

we can rewrite the integral as

$$\mathcal{J}_1 = \int_{\mathbb{R}^k} d\delta \exp \left[-\frac{1}{2} (\Sigma_0^{-1})_{ba} \delta_a \delta_b \right] (\Sigma_1^{-1})_{ba} (\delta + \xi)_a (\delta + \xi)_b. \quad (\text{E21})$$

Using the Gaussian integral formulas above to simplify gives the result:

$$\mathcal{J}_1 = (2\pi)^{k/2} (\det \Sigma_0)^{1/2} (\xi^T \Sigma_1^{-1} \xi + \text{tr}[\Sigma_0 \Sigma_1^{-1}]). \quad (\text{E22})$$

We need one more additional integral:

$$\mathcal{J}_2 \equiv \int_{\mathbb{R}^k} d\mathbf{a} \exp \left[-\frac{1}{2} (\mu_0 - \mathbf{a})^T (\Sigma_0^{-1}) (\mu_0 - \mathbf{a}) - \frac{1}{2} (\mu_1 - \mathbf{a})^T (N \Sigma_0)^{-1} (\mu_1 - \mathbf{a}) \right]. \quad (\text{E23})$$

where N is an integer. Applying the same change of variables as above and gathering terms, we have

$$\mathcal{J}_2 = \int_{\mathbb{R}^k} d\delta \exp \left[-\frac{1}{2} \left(\frac{N+1}{N} \Sigma_0^{-1} \right)_{ba} \delta_a \delta_b - \frac{1}{2N} (\Sigma_0^{-1})_{ba} (\xi_a \xi_b + \xi_a \delta_b + \delta_a \xi_b) \right]. \quad (\text{E24})$$

We change variables again to $\delta' = \sqrt{(N+1)/N} \delta$ to absorb the extra factor in the first term. Pulling the δ -independent exponential factor out front, we then have

$$\mathcal{J}_2 = e^{-\xi^T \Sigma_0^{-1} \xi / (2N)} \left(\frac{N}{N+1} \right)^{k/2} \int_{\mathbb{R}^k} d\delta' \exp \left[-\frac{1}{2} (\Sigma_0^{-1})_{ba} \delta'_a \delta'_b - \frac{1}{2} \sqrt{\frac{1}{N(N+1)}} (\Sigma_0^{-1})_{ba} (\xi_a \delta'_b + \delta'_a \xi_b) \right] \quad (\text{E25})$$

$$= e^{-\xi^T \Sigma_0^{-1} \xi / (2N)} \left(\frac{N}{N+1} \right)^{k/2} (2\pi)^{k/2} (\det \Sigma_0)^{1/2} \exp \left[\frac{1}{2N(N+1)} \xi^T \Sigma_0^{-1} \xi \right] \quad (\text{E26})$$

or simplifying,

$$\mathcal{J}_2 = \left(\frac{N}{N+1} \right)^{k/2} (2\pi)^{k/2} (\det \Sigma_0)^{1/2} \exp \left[-\frac{1}{2(N+1)} \xi^T \Sigma_0^{-1} \xi \right]. \quad (\text{E27})$$

APPENDIX F: ALTERNATIVE DERIVATIONS FOR DATA SUBSET SELECTION

In this appendix, we give an alternative derivation for the data subset selection formulas given in Sec. V that uses the partition of data and the least-squares ICs rather than computing the KL divergences directly.

In the case of least-squares regression with correct model specification, the BPIC and PPIC (before the integral approximations) are given by

$$\text{BPIC}_\mu = \chi_{\text{aug}}^2(\mathbf{a}^*) - \frac{\int d\mathbf{a} \exp[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a})] \tilde{\chi}^2(\mathbf{a})}{\int d\mathbf{a} \exp[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a})]} + 3k, \quad (\text{F1})$$

$$\text{PPIC}_\mu = -2 \sum_{i=1}^N \log \frac{\int d\mathbf{a} \exp[-\frac{1}{2} (\chi_{\text{aug}}^2(\mathbf{a}) + \chi_i^2(\mathbf{a}))]}{\int d\mathbf{a} \exp[-\frac{1}{2} \chi_{\text{aug}}^2(\mathbf{a})]} + 2k. \quad (\text{F2})$$

For data selection, $k \rightarrow k + d_C$, to account for the additional d_C parameters for the “perfect” model as before. We now must understand how the chi-squared functions and integrals transform as well.

As described in the main body in Sec. V, we partition the χ^2 functions into kept, cut, and off-block-diagonal pieces:

$$\hat{\chi}^2(\mathbf{a}) = (\bar{y} - \phi_{M,P}(\mathbf{a}))^T \hat{\Sigma}^{-1} (\bar{y} - \phi_{M,P}(\mathbf{a})) \quad (\text{F3})$$

$$= \begin{pmatrix} \bar{y}_C - \mathbf{a}_C \\ \bar{y}_K - f_M(\mathbf{a}_K) \end{pmatrix}^T \begin{pmatrix} (\hat{\Sigma}^{-1})_C & (\hat{\Sigma}^{-1})_O \\ (\hat{\Sigma}^{-1})_O^T & (\hat{\Sigma}^{-1})_K \end{pmatrix} \begin{pmatrix} \bar{y}_C - \mathbf{a}_C \\ \bar{y}_K - f_M(\mathbf{a}_K) \end{pmatrix} \quad (\text{F4})$$

$$\equiv \hat{\chi}_C^2(\mathbf{a}_C) + \hat{\chi}_K^2(\mathbf{a}_K) + 2\hat{\chi}_O^2(\mathbf{a}_C, \mathbf{a}_K), \quad (\text{F5})$$

where

$$\hat{\chi}_C^2(\mathbf{a}_C) \equiv (\bar{y}_C - \mathbf{a}_C)^T (\hat{\Sigma}^{-1})_C (\bar{y}_C - \mathbf{a}_C), \quad (\text{F6})$$

$$\hat{\chi}_K^2(\mathbf{a}_K) \equiv (\bar{y}_K - f_M(\mathbf{a}_K))^T (\hat{\Sigma}^{-1})_K (\bar{y}_K - f_M(\mathbf{a}_K)), \quad (\text{F7})$$

$$\hat{\chi}_O^2(\mathbf{a}_C, \mathbf{a}_K) \equiv (\bar{y}_C - \mathbf{a}_C)^T (\hat{\Sigma}^{-1})_O (\bar{y}_C - f_M(\mathbf{a}_K)) \quad (\text{F8})$$

$$= (\bar{y}_K - f_M(\mathbf{a}_K))^T (\hat{\Sigma}^{-1})_O^T (\bar{y}_C - \mathbf{a}_C). \quad (\text{F9})$$

Similarly, we define

$$\hat{\chi}_{K,i}^2(\mathbf{a}_K) \equiv (y_{K,i} - f_M(\mathbf{a}_K))^T (\Sigma^{-1})_K (y_{K,i} - f_M(\mathbf{a}_K)), \quad (\text{F10})$$

$$\chi_{C,i}^2(\mathbf{a}_C) \equiv (y_{C,i} - \mathbf{a}_C)^T (\Sigma^{-1})_C (y_{C,i} - \mathbf{a}_C), \quad (\text{F11})$$

$$\chi_{O,i}^2(\mathbf{a}_C, \mathbf{a}_K) \equiv (y_{C,i} - \mathbf{a}_C)^T (\Sigma^{-1})_O (y_{K,i} - f_M(\mathbf{a}_K)). \quad (\text{F12})$$

Furthermore,

$$\tilde{\chi}^2(\mathbf{a}) = (\mathbf{a}_C - \bar{y}_C)^T \tilde{\Sigma}^{-1} (\mathbf{a}_C - \bar{y}_C) + (\mathbf{a}_K - \bar{\mathbf{a}})^T \tilde{\Sigma}_K^{-1} (\mathbf{a}_K - \bar{\mathbf{a}}) \equiv \tilde{\chi}_C^2(\mathbf{a}_C) + \tilde{\chi}_K^2(\mathbf{a}_K), \quad (\text{F13})$$

where $\tilde{\Sigma}^{-1} = \text{diag}(\tilde{\Sigma}_C^{-1}, \tilde{\Sigma}_K^{-1})$ by construction. It follows that

$$\chi_{\text{aug}}^2(\mathbf{a}) = \chi_{C,\text{aug}}^2(\mathbf{a}_C) + \chi_{K,\text{aug}}^2(\mathbf{a}_K) + 2\hat{\chi}_O^2(\mathbf{a}_C, \mathbf{a}_K), \quad (\text{F14})$$

where $\chi_{C,\text{aug}}^2$ and $\chi_{K,\text{aug}}^2$ are defined analogously to Eq. (63) with the cut and kept statistics, respectively. We note for later use that $\hat{\chi}_C^2(\bar{y}_C)$, $\tilde{\chi}_C^2(\bar{y}_C)$, $\chi_{C,\text{aug}}^2(\bar{y}_C)$, and $\hat{\chi}_O^2(\bar{y}_C, \mathbf{a}_K)$ vanish identically (for all \mathbf{a}_K).

With these definitions, the information criteria become

$$\text{BPIC}_{\mu,P} = - \frac{\int d\mathbf{a}_C d\mathbf{a}_K \exp[-\frac{1}{2}(\chi_{K,\text{aug}}^2 + \chi_{C,\text{aug}}^2 + 2\hat{\chi}_O^2)] (\tilde{\chi}_K^2 + \tilde{\chi}_C^2)}{\int d\mathbf{a}_C d\mathbf{a}_K \exp[-\frac{1}{2}(\chi_{K,\text{aug}}^2 + \chi_{C,\text{aug}}^2 + 2\hat{\chi}_O^2)]} + \chi_{K,\text{aug}}^2(\mathbf{a}_K^*) + 3(k + d_C), \quad (\text{F15})$$

$$\text{PPIC}_{\mu,P} = -2 \sum_{i=1}^N \log \frac{\int d\mathbf{a}_C d\mathbf{a}_K \exp[-\frac{1}{2}(\chi_{K,\text{aug}}^2 + \chi_{C,\text{aug}}^2 + 2\hat{\chi}_O^2 + \chi_{K,i}^2 + \chi_{C,i}^2 + 2\chi_{i,O}^2)]}{\int d\mathbf{a}_C d\mathbf{a}_K \exp[-\frac{1}{2}(\chi_{K,\text{aug}}^2 + \chi_{C,\text{aug}}^2 + 2\hat{\chi}_O^2)]} + 2(k + d_C), \quad (\text{F16})$$

where we have suppressed the arguments of the χ^2 functions in the integrands for simplicity.

As in the main body, here we assume that the off-block-diagonal elements of the sample covariance $\hat{\Sigma}_O$ are small (in the sense of induced operator norm) relative to the on-block-diagonal elements $\hat{\Sigma}_C$ and $\hat{\Sigma}_K$. It follows from this approximation that $\chi_{i,C}^2, \chi_{i,K}^2 \gg \chi_{i,O}^2$ and $\hat{\chi}_C^2, \hat{\chi}_K^2 \gg \hat{\chi}_O^2$. Therefore,

$$\text{BPIC}_{\mu,P} \approx \chi_{K,\text{aug}}^2(\mathbf{a}_K^*) - \frac{\int d\mathbf{a}_K \exp[-\frac{1}{2}\chi_{K,\text{aug}}^2] \tilde{\chi}_K^2}{\int d\mathbf{a}_K \exp[-\frac{1}{2}\chi_{K,\text{aug}}^2]} + 3k - \frac{\int d\mathbf{a}_C \exp[-\frac{1}{2}\chi_{C,\text{aug}}^2] \tilde{\chi}_C^2}{\int d\mathbf{a}_C \exp[-\frac{1}{2}\chi_{C,\text{aug}}^2]} + 3d_C, \quad (\text{F17})$$

$$\text{PPIC}_{\mu,P} \approx -2 \sum_{i=1}^N \log \frac{\int d\mathbf{a}_K \exp[-\frac{1}{2}(\chi_{K,\text{aug}}^2 + \chi_{K,i}^2)]}{\int d\mathbf{a}_K \exp[-\frac{1}{2}\chi_{K,\text{aug}}^2]} + 2k - 2 \sum_{i=1}^N \log \frac{\int d\mathbf{a}_C \exp[-\frac{1}{2}(\chi_{C,\text{aug}}^2 + \chi_{C,i}^2)]}{\int d\mathbf{a}_C \exp[-\frac{1}{2}\chi_{C,\text{aug}}^2]} + 2d_C. \quad (\text{F18})$$

The ‘‘K’’ integrals can be approximated using the NLO Laplace approximation as in the previous sections, leading to the same IC formulas that we found previously over the kept data, but with a subtle difference: the inverse covariance matrix appearing in $\hat{\chi}_K^2$ is the sub-block of the full

inverse matrix $(\hat{\Sigma}^{-1})_K$, as opposed to the inverse of the sub-block covariance matrix $(\hat{\Sigma}_K)^{-1}$. Under our block-diagonal assumption $\Sigma^{-1} \approx \text{diag}(\Sigma_K^{-1}, \Sigma_C^{-1})$, these matrices are identical; even when off-diagonal correlations are present, the inverse of the sub-block is often used in practice to define $\hat{\chi}_K^2$. For further discussion of this point, see Sec. V.

We will also compute the ‘‘C’’ integrals using our Laplace approximation formulas, but since we can take $M_{\text{perf},\mu}$ to be linear (e.g., a polynomial of degree $d_C - 1$), the BPIC formula will be exact; the PPIC is not exact here, but the result in Sec. V is exact. Beginning with the BPIC, we have the result

$$\frac{\int d\mathbf{a}_C \exp[-\frac{1}{2}\chi_{\text{aug},C}^2] \tilde{\chi}_C^2}{\int d\mathbf{a}_C \exp[-\frac{1}{2}\chi_{\text{aug},C}^2]} = \text{tr}[\tilde{\Sigma}_C^{-1} \Sigma_C^*], \quad (\text{F19})$$

taking $\Sigma^* \approx \text{diag}(\Sigma_K^*, \Sigma_C^*)$ which follows from our assumption that $\Sigma^{-1} \approx \text{diag}(\Sigma_K^{-1}, \Sigma_C^{-1})$. Since we are working in the limit of infinitely diffuse priors over the cut data, we have $(\Sigma_C^*)^{-1} = \hat{\Sigma}_C^{-1} + \tilde{\Sigma}_C^{-1} \rightarrow \hat{\Sigma}_C^{-1}$, and thus

$$\text{tr}[\tilde{\Sigma}_C^{-1} \Sigma_C^*] \rightarrow \text{tr}[\hat{\Sigma}_C^{-1} \hat{\Sigma}_C] \rightarrow 0. \quad (\text{F20})$$

So the only additional contribution to the BPIC for data subset selection is a penalty term of $+3d_C$. We turn next to the PPIC, where the Laplace approximation formula gives us the result:

$$\sum_{i=1}^N \log \frac{\int d\mathbf{a}_C \exp[-\frac{1}{2}(\chi_{\text{aug},C}^2 + \chi_{C,i}^2)]}{\int d\mathbf{a}_C \exp[-\frac{1}{2}\chi_{\text{aug},C}^2]} = \sum_{i=1}^N \log \left[1 + \frac{1}{2} \text{tr} \left[\left(\frac{1}{4} (g_{Ci})(g_{C,i})^T - \frac{1}{2} H_{C,i} \right) \Sigma_C^* \right] \right]. \quad (\text{F21})$$

In summary, the ICs (up to constant terms) for data subset selection are

$$\text{BAIC}_{\mu,P} = \hat{\chi}^2(\mathbf{a}^*) + 2k + 2d_C, \quad (\text{F22})$$

$$\text{BPIC}_{\mu,P} \approx \hat{\chi}^2(\mathbf{a}^*) - \frac{1}{2} \tilde{H}_{ba}(\Sigma^*)_{ab} + \frac{1}{2} \tilde{g}_d T_{cba}(\Sigma_2^*)_{abcd} + 3k + 3d_C, \quad (\text{F23})$$

$$\begin{aligned} \text{PPIC}_{\mu,P} \approx & \hat{\chi}^2(\mathbf{a}^*) + 2k + 2 \left(1 + \frac{1}{2N} \right) d_C \\ & - 2 \sum_{i=1}^N \log \left[1 + \frac{1}{2} \left(\frac{1}{4} (g_i)_b (g_i)_a - \frac{1}{2} (H_i)_{ba} \right) (\Sigma^*)_{ab} + \frac{1}{4} (g_i)_d T_{cba}(\Sigma_2^*)_{abcd} \right], \end{aligned} \quad (\text{F24})$$

where we have dropped all ‘‘K’’ subscripts.

It is worth noting that the PPIC in Eq. (F24) disagrees with Eq. (127) even to $O(N^{-1})$. This disagreement comes from higher-order terms in the bias correction of the PPIC on the perfect model. By computing the KL_C exactly, the result in Eq. (127) accounts for these corrections where as Eq. (F24) does not. Studying higher-order corrections to IC bias could be an interesting direction for future work.

APPENDIX G: RELEVANT DERIVATIVES

In this appendix, we give expressions for the relevant derivative tensors used in Sec. VI. We write these in terms of derivatives of the model function $f(\mathbf{a})$ (with the model index μ suppressed), which we calculated using auto-differentiation

(specifically using the Python package JAX [98]) to obtain the results in Sec. VI. For brevity, we only give expressions for derivatives for χ_{aug}^2 , as the other relevant derivatives (i.e., those of $\tilde{\chi}^2$, $\hat{\chi}^2$, and χ_i^2) can be deduced from

$$\chi_{\text{aug}}^2(\mathbf{a}) = \tilde{\chi}^2(\mathbf{a}) + \hat{\chi}^2(\mathbf{a}) = \tilde{\chi}^2(\mathbf{a}) + \sum_{i=1}^N \chi_i^2(\mathbf{a}) - (N-1)d. \quad (\text{G1})$$

The derivative tensors are written in index summation notation where indices at the beginning of the (roman) alphabet (i.e., a, b, c, d) denote parameter dimensions and at the end of the alphabet (i.e., x, y) data dimensions. The expressions are as follows:

$$\chi_{\text{aug}}^2 = (\mathbf{a} - \tilde{\mathbf{a}})_a (\tilde{\Sigma}^{-1})_{ab} (\mathbf{a} - \tilde{\mathbf{a}})_b + (\bar{y} - f)_x (\hat{\Sigma}^{-1})_{xy} (\bar{y} - f)_y, \quad (\text{G2})$$

$$\frac{\partial \chi_{\text{aug}}^2}{\partial a_a} = 2 \left[(\tilde{\Sigma}^{-1})_{ab} (\mathbf{a} - \tilde{\mathbf{a}})_b - \left(\frac{\partial f}{\partial a_a} \right)_x (\hat{\Sigma}^{-1})_{xy} (\bar{y} - f)_y \right], \quad (\text{G3})$$

$$\frac{\partial^2 \chi_{\text{aug}}^2}{\partial a_a \partial a_b} = 2 \left[(\tilde{\Sigma}^{-1})_{ab} - \left(\frac{\partial^2 f}{\partial a_a \partial a_b} \right)_x (\hat{\Sigma}^{-1})_{xy} (\bar{y} - f)_y + \left(\frac{\partial f}{\partial a_a} \right)_x (\hat{\Sigma}^{-1})_{xy} \left(\frac{\partial f}{\partial a_b} \right)_y \right], \quad (\text{G4})$$

$$\begin{aligned} \frac{\partial^3 \chi_{\text{aug}}^2}{\partial a_a \partial a_b \partial a_c} = & 2 \left[- \left(\frac{\partial^3 f}{\partial a_a \partial a_b \partial a_c} \right)_x (\hat{\Sigma}^{-1})_{xy} (\bar{y} - f)_y + \left(\frac{\partial^2 f}{\partial a_a \partial a_b} \right)_x (\hat{\Sigma}^{-1})_{xy} \left(\frac{\partial f}{\partial a_c} \right)_y \right. \\ & \left. + \left(\frac{\partial^2 f}{\partial a_a \partial a_c} \right)_x (\hat{\Sigma}^{-1})_{xy} \left(\frac{\partial f}{\partial a_b} \right)_y + \left(\frac{\partial^2 f}{\partial a_b \partial a_c} \right)_x (\hat{\Sigma}^{-1})_{xy} \left(\frac{\partial f}{\partial a_a} \right)_y \right]. \end{aligned} \quad (\text{G5})$$

Note that we omit $(\partial^4 \chi_{\text{aug}}^2)/(\partial a_a \partial a_b \partial a_c \partial a_d)$ as it cancels to NLO in any case (see Sec. IV D for details).

- [1] L. Wasserman, Bayesian model selection and model averaging, *J. Math. Psychol.* **44**, 92 (2000).
- [2] D. Parkinson and A. R. Liddle, Bayesian model averaging in astrophysics: A review, *Stat. Anal. Data Min.* **6**, 3 (2013).
- [3] T. M. Fragoso, W. Bertoli, and F. Louzada, Bayesian model averaging: A systematic review and conceptual classification, *Int. Stat. Rev.* **86**, 1 (2018).
- [4] E. E. Leamer, *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (Wiley, New York, 1978).
- [5] A. Racine, A. Grieve, H. Fluhler, and A. Smith, Bayesian methods in practice: Experiences in the pharmaceutical industry, *Appl. Stat.* **35**, 93 (1986).
- [6] D. Madigan and A. E. Raftery, Model selection and accounting for model uncertainty in graphical models using occam's window, *J. Am. Stat. Assoc.* **89**, 1535 (1994).
- [7] R. E. Kass and A. E. Raftery, Bayes factors, *J. Am. Stat. Assoc.* **90**, 773 (1995).
- [8] A. E. Raftery and Y. Zheng, Discussion: Performance of Bayesian model averaging, *J. Am. Stat. Assoc.* **98**, 931 (2003).
- [9] H. Akaike, Information theory as an extension of the maximum likelihood principle, in *Proceedings of the 2nd International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki (Academiai Kiado, Budapest, 1973).
- [10] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Autom. Control* **19**, 716 (1974).
- [11] H. Akaike, Information theory and an extension of the maximum likelihood principle, in *Selected Papers of Hirotugu Akaike* (Springer, New York, 1998), pp. 199–213.
- [12] S. Zhou, Bayesian model selection in terms of Kullback-Leibler discrepancy, Ph.D. thesis, Columbia University, 2011.
- [13] S. Zhou, Posterior averaging information criterion, [arXiv:2009.09248](https://arxiv.org/abs/2009.09248).
- [14] Y. Chen, S.-J. Dong, T. Draper, I. Horvath, K.-F. Liu, N. Mathur, S. Tamhankar, C. Srinivasan, F. X. Lee, and J.-b. Zhang, The Sequential empirical Bayes method: An adaptive constrained-curve fitting algorithm for lattice QCD, [arXiv:hep-lat/0405001](https://arxiv.org/abs/hep-lat/0405001).
- [15] C. Davies, K. Hornbostel, I. Kendall, G. Lepage, C. McNeile, J. Shigemitsu, and H. Trotter (HPQCD Collaboration), Update: Accurate determinations of $\alpha(s)$ from realistic lattice QCD, *Phys. Rev. D* **78**, 114507 (2008).
- [16] M. R. Schindler and D. R. Phillips, Bayesian methods for parameter estimation in effective field theories, *Ann. Phys. (Amsterdam)* **324**, 682 (2009); **324**, 2051(E) (2009).
- [17] S. Dürr *et al.* (Budapest-Marseille-Wuppertal Collaboration), Lattice QCD at the physical point meets SU(2) chiral perturbation theory, *Phys. Rev. D* **90**, 114504 (2014).
- [18] E. Berkowitz *et al.*, An accurate calculation of the nucleon axial charge with lattice QCD, [arXiv:1704.01114](https://arxiv.org/abs/1704.01114).
- [19] C. Chang *et al.*, A per-cent-level determination of the nucleon axial coupling from quantum chromodynamics, *Nature (London)* **558**, 91 (2018).
- [20] E. Rinaldi, S. Syritsyn, M. L. Wagman, M. I. Buchoff, C. Schroeder, and J. Wasem, Lattice QCD determination of neutron-antineutron matrix elements with physical quark masses, *Phys. Rev. D* **99**, 074510 (2019).
- [21] N. Miller *et al.*, F_K/F_π from Möbius domain-wall fermions solved on gradient-flowed HISQ ensembles, *Phys. Rev. D* **102**, 034507 (2020).
- [22] S. Borsanyi *et al.*, Leading hadronic contribution to the muon 2 magnetic moment from lattice QCD, *Nature (London)* **593**, 51 (2021).
- [23] D. R. Phillips *et al.*, Get on the BAND Wagon: A Bayesian framework for quantifying model uncertainties in nuclear dynamics, *J. Phys. G* **48**, 072001 (2021).
- [24] M. A. Connell, I. Billig, and D. R. Phillips, Does Bayesian model averaging improve polynomial extrapolations? Two toy problems as tests, *J. Phys. G* **48**, 104001 (2021).
- [25] S. Borsanyi *et al.* (BMW Collaboration), *Ab initio* calculation of the neutron-proton mass difference, *Science* **347**, 1452 (2015).
- [26] W. I. Jay and E. T. Neil, Bayesian model averaging for analysis of lattice field theory results, *Phys. Rev. D* **103**, 114502 (2021).
- [27] R. J. Larsen and M. L. Marx, *An Introduction to Mathematical Statistics* (Prentice Hall, New York, 2005).
- [28] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. R. C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, *Phys. Rep.* **810**, 1 (2019).
- [29] S. Kullback and R. A. Leibler, On information and sufficiency, *Ann. Math. Stat.* **22**, 79 (1951).
- [30] S. Konishi and G. Kitagawa, Generalised information criteria in model selection, *Biometrika* **83**, 875 (1996).
- [31] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, Bayesian measures of model complexity and fit, *J. R. Stat. Soc. Ser. B* **64**, 583 (2002).
- [32] T. Ando, Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical bayes models, *Biometrika* **94**, 443 (2007).
- [33] K. Takeuchi, Distribution of information statistics and criteria for adequacy of models, *Suri-Kagaku (Math. Sci.)* **153**, 12 (1976).
- [34] M. Stone, An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *J. R. Stat. Soc. Ser. B* **39**, 44 (1977).
- [35] R. Shibata, Statistical aspects of model selection, in *From Data to Model*, edited by J. Willems (Springer-Verlag, Berlin, Heidelberg, 1989), pp. 215–240.
- [36] M. Dixon and T. Ward, Takeuchi's information criteria as a form of regularization, *Entropy* **23**, 1419 (2021).
- [37] H. White, Maximum likelihood estimation of misspecified models, *Econometrica* **50**, 1 (1982).
- [38] S. Konishi and G. Kitagawa, Bayesian information criteria, in *Information Criteria and Statistical Modeling* (Springer, New York, 2008), Chap. 9, pp. 211–237.
- [39] H. Akaike, Likelihood and the Baye procedure, in *Bayesian Statistics*, edited by N. J. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (University Press, Valencia, Spain, 1980), pp. 141–166.
- [40] G. Kitagawa, Information criteria for the predictive evaluation of Bayesian models, *Commun. Stat., Theory Methods* **26**, 2223 (1997).
- [41] G. Schwarz, Estimating the dimesnion of a model, *Ann. Stat.* **6**, 461 (1978).

- [42] C. P. Robert and D. M. Titterton, Discussion of a paper by D. J. Spiegelhalter et al., *J. R. Stat. Soc. Ser. B* **64**, 621 (2002).
- [43] T. Ando, Bayesian model averaging and Bayesian predictive information criterion for model selection, *J. Japan Stat. Soc.* **38**, 243 (2008).
- [44] J. L. W. V. Jensen, Sur les fonctions convexes et les inégalités entre les valeurs moyennes, *Acta Math.* **30**, 175 (1906).
- [45] J. K. Hunter and B. Nachtergaele, *Applied Analysis* (World Scientific Publishing Company, Singapore, 2001).
- [46] A. Gelman, J. Hwang, and A. Vehtari, Understanding predictive information criteria for bayesian models, *Stat. Comput.* **24**, 997 (2014).
- [47] G. Lepage, B. Clark, C. Davies, H. K., P. Mackenzie, C. Morningstar, and H. Trotter, Constrained curve fitting, *Nucl. Phys. B, Proc. Suppl.* **106**, 12 (2002).
- [48] M. R. Schindler and D. R. Phillips, Bayesian methods for parameter estimation in effective field theories, *Ann. Phys. (Amsterdam)* **324**, 682 (2009).
- [49] E. T. Neil and J. W. Sitison, Model averaging approaches to data subset selection, *Phys. Rev. E* **108**, 045308 (2023).
- [50] W. D. Kirwin, Higher asymptotics of Laplace's approximation, *Asymptotic Analysis* **70**, 231 (2010).
- [51] L. Tierney and J. B. Kadane, Accurate approximations for posterior moments and marginal densities, *J. Am. Stat. Assoc.* **81**, 82 (1986).
- [52] H. Poincaré, Sur les intégrales irrégulières des équations linéaires, *Acta Math.* **8**, 295 (1886).
- [53] M. V. Berry and C. J. Howls, Hyperasymptotics for integrals with saddles, *Proc. R. Soc. A* **434**, 657 (1991).
- [54] M. Berry, *Asymptotics, Superasymptotics, Hyperasymptotics...* (Springer, Boston, MA, 1991), pp. 1–14.
- [55] J. P. Boyd, The devil's invention: Asymptotic superasymptotic and hyperasymptotic series, *Acta Appl. Math.* **56**, 1 (1999).
- [56] R. B. Paris, *Hadamard Expansions and Hyperasymptotic Evaluation: An Extension of the Method of Steepest Descents* (Cambridge University Press, Cambridge, England, 2011), Vol. 141.
- [57] G. G. Stokes, On the numerical calculation of a class of definite integrals and infinite series, *Trans. Cambridge Philos. Soc.* **9**, 379 (1847).
- [58] O. Costin and M. D. Kruskal, On optimal truncation of divergent series solutions of nonlinear differential systems; berry smoothing, *Proc. R. Soc. A* **455**, 1931 (1999).
- [59] M. D. Kruskal and H. Segur, Asymptotics beyond all orders in a model of crystal growth, *Stud. Appl. Math.* **85**, 129 (1991).
- [60] R. B. Dingle, *Asymptotic Expansion: Their Derivation and Interpretation* (Academic Press, New York and London, 1973).
- [61] C. J. Howls, Hyperasymptotics for multidimensional integrals, exact remainder terms and the global connection problem, *Proc. R. Soc. A* **453**, 2271 (1997).
- [62] T. Appelquist et al. (Lattice Strong Dynamics Collaboration), Nonperturbative investigations of SU(3) gauge theory with eight dynamical flavors, *Phys. Rev. D* **99**, 014509 (2019).
- [63] S. R. Beane et al. (NPLQCD, QCDSF Collaborations), Charged multihadron systems in lattice QCD + QED, *Phys. Rev. D* **103**, 054504 (2021).
- [64] J. Hartlap, P. Simon, and P. Schneider, Why your model parameter confidences might be too optimistic. unbiased estimation of the inverse covariance matrix, *Astron. Astrophys.* **464**, 399 (2007).
- [65] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. (John Wiley & sons, Inc., Hoboken, New Jersey, 2003).
- [66] S. I. Vrieze, Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), *Psychol. Methods* **17**, 228 (2012).
- [67] R. Shibata, Asymptotic mean efficiency of a selection of regression variables, *Ann. Inst. Stat. Math.* **35**, 415 (1983).
- [68] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, *J. Multivariate Anal.* **88**, 365 (2004).
- [69] O. Ledoit and M. Wolf, Nonlinear shrinkage estimation of large-dimensional covariance matrices, *Ann. Stat.* **40**, 1024 (2012).
- [70] O. Ledoit and M. Wolf, Direct nonlinear shrinkage estimation of large-dimensional covariance matrices, Technical Report, Working Paper, 2017.
- [71] A. Bazavov et al. (Fermilab Lattice, MILC Collaborations), D-meson semileptonic decays to pseudoscalars from four-flavor lattice QCD, *Phys. Rev. D* **107**, 094516 (2023).
- [72] G. Lepage, lsqfit, <https://github.com/gplepage/lsqfit> (2021).
- [73] G. Lepage, gvar, <https://github.com/gplepage/gvar> (2022).
- [74] A. Bazavov et al. (Fermilab Lattice, MILC Collaboration), $B_{(s)}^0$ -mixing matrix elements from lattice QCD for the standard model and beyond, *Phys. Rev. D* **93**, 113016 (2016).
- [75] G. P. Lepage, The analysis of algorithms for lattice field theory, in *Theoretical Advanced Study Institute in Elementary Particle Physics* (World Scientific Publishing Company, Singapore, 1989).
- [76] D. Grabowska, D. B. Kaplan, and A. N. Nicholson, Sign problems, noise, and chiral symmetry breaking in a QCD-like theory, *Phys. Rev. D* **87**, 014504 (2013).
- [77] M. L. Wagman and M. J. Savage, Statistics of baryon correlation functions in lattice QCD, *Phys. Rev. D* **96**, 114508 (2017).
- [78] G. P. Lepage, A new algorithm for adaptive multidimensional integration, *J. Comput. Phys.* **27**, 192 (1978).
- [79] G. P. Lepage, Adaptive multidimensional integration: VEGAS enhanced, *J. Comput. Phys.* **439**, 110386 (2021).
- [80] G. Lepage, vegas, <https://github.com/gplepage/vegas> (2022).
- [81] S. Mondal, R. Gupta, S. Park, B. Yoon, T. Bhattacharya, B. Joó, and F. Winter (Nucleon Matrix Elements (NME) Collaboration), Nucleon momentum fraction, helicity and transversity from 2 + 1-flavor lattice QCD, *J. High Energy Phys.* **04** (2020) 004.
- [82] C. M. Hurvich and C.-L. Tsai, Regression and time series model selection in small samples, *Biometrika* **76**, 297 (1989).
- [83] C. M. Hurvich and C.-L. Tsai, Model selection for extended quasi-likelihood models in small samples, *Biometrics* **51**, 1077 (1995).

- [84] N. Sugiura, Further analysis of the data by Akaike's information criterion and the finite corrections, *Commun. Stat.-Theor. Methods* **7**, 13 (1978).
- [85] D. Anderson and K. Burnham, *Model Selection and Multi-Model Inference*, 2nd ed. (Springer-Verlag, New York, 2004), 63, 10.
- [86] R. Shibata, Bootstrap estimate of Kullback-Leibler information for model selection, *Statistica Sinica* **7**, 375 (1997).
- [87] A. Gelman and C. R. Shalizi, Philosophy and the practice of Bayesian statistics, *Br. J. Math. Stat. Psychol.* **66**, 8 (2013).
- [88] V. Dose, Bayesian inference in physics: Case studies, *Rep. Prog. Phys.* **66**, 1421 (2003).
- [89] D. V. Lindley, A statistical paradox, *Biometrika* **44**, 187 (1957).
- [90] H. Jeffreys, *The Theory of Probability* (OUP, Oxford, 1998).
- [91] R. D. Cousins, The Jefferys-Lindley paradox and discovery criteria in high energy physics, *Synthese* **194**, 395 (2017).
- [92] O. E. Barndorff-Nielsen and D. R. Cox, *Asymptotic Techniques for Use in Statistics* (Chapman and Hall, London, 1989).
- [93] A. Takeshi, *Advanced Econometrics* (Harvard University Press, Cambridge, MA, 1985).
- [94] R. I. Jennrich, Asymptotic properties of non-linear least squares estimators, *Ann. Math. Stat.* **40**, 644 (1969).
- [95] C.-F. Wu, Asymptotic theory of nonlinear least squares estimation, *Ann. Stat.* **9**, 501 (1981).
- [96] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory* (Springer Science & Business Media, New York, 2013).
- [97] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals* (Courier Corporation, New York, 1974).
- [98] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX: Composable transformations of Python + NumPy programs (2018).