Introduction to Bayesian Methods and Uncertainty Quantification

Sarah Wesolowski
Salisbury University

APS Division of Nuclear Physics meeting, 2019

What is "Bayesian" inference?

Table 1.1 *Frequentist and Bayesian approaches to probability*.

Approach	Probability definition
FREQUENTIST STATISTICAL INFERENCE	$p(A)$ = long-run relative frequency with which A occurs in identical repeats of an experiment. "A" restricted to propositions about random variables.
BAYESIAN INFERENCE	$p(A\|B)$ = a real number measure of the plausibility of a proposition/hypothesis A, given (conditional on) the truth of the information represented by proposition B. "A" can be any logical proposition, *not* restricted to propositions about random variables.

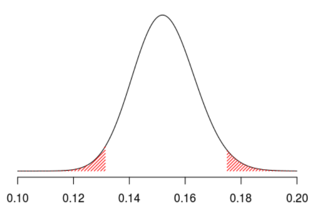
P. Gregory, "Bayesian Logical Data Analysis for the Physical Sciences"

Comparison of credible and confidence intervals

Bayesian probability:

- probabilities treated as degree of plausibility
- more natural interpretation for quantities like model parameters

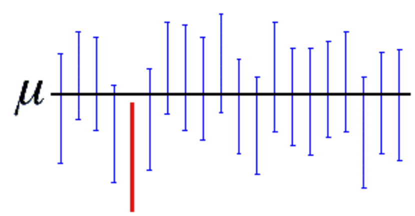
p% credible interval: there is p% probability that the true, unknown value lies in the interval



Frequentist probability:

- probability is long-run frequency
- relies on the idea of identical repeats

p% confidence interval: will cover the true value of the quantity over p% of experiments



See references for more nuance, philosophy, and debates.

Uncertainty Quantification in Nuclear Physics

To produce meaningful experimental measurements and theoretical predictions, it is essential to quantify uncertainties!

$$y_{\text{th}} + \delta y_{\text{th}} = y_{\text{exp}} + \delta y_{\text{exp}}$$

Theory discrepancy:

$$\delta y_{\text{th}}$$

Made up of the following:

- missing physics
- numerical/ method errors
- fitting to uncertain data

Notes:

- likely to be "systematic"
- not usually fully quantified
- often assumed to be normal

Experimental discrepancy:

$$\delta y_{\text{exp}}$$

Made up of the following:

- counting statistics
- background and selection effects
- systematic uncertainties

Notes

- systematic errors may not be well understood or inflated
- often assumed to be normal

Practical details

Probability of A being true given that B is true:

$$p(A|B)$$

- "Given information": inclusion of prior information (physics!)
- Bayesian pdfs follow all the same rules of probability:

$$p(A|B) + p(\bar{A}|B) = 1$$

$$\begin{aligned} p(A, B|C) &= p(A|C)p(B|A, C) \\ &= p(B|C)p(A|B, C) \end{aligned}$$

- Bayes theorem is a simple rearrangement of the product rule:

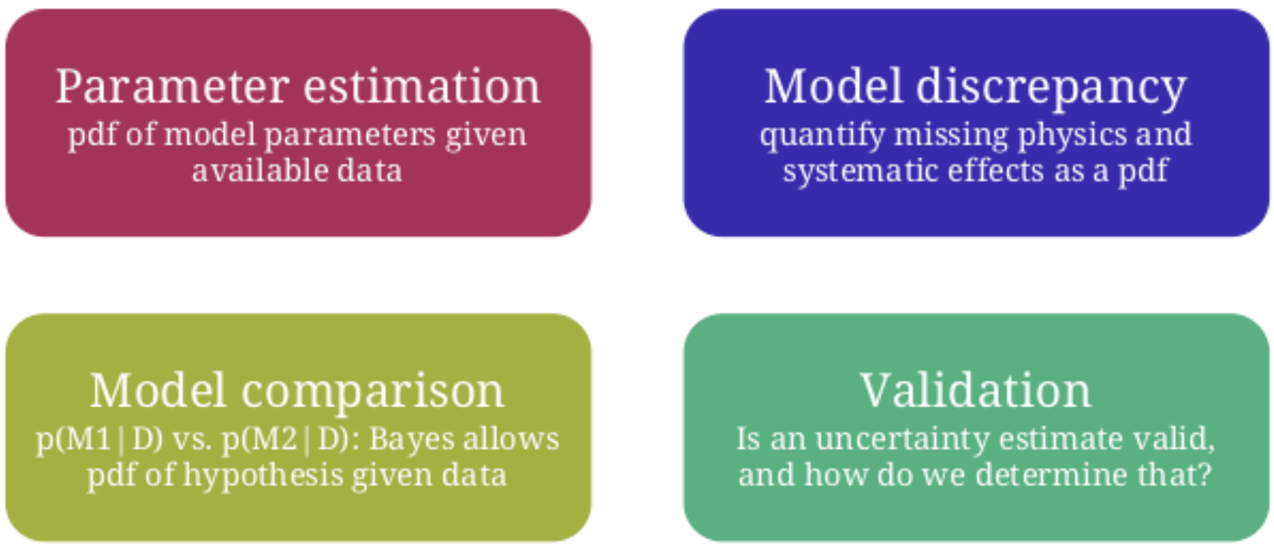
$$p(A|B, C) = \frac{p(B|A, C)p(A|C)}{p(B|C)}$$

- Can develop prescriptions for combining sources of uncertainty
- Many frequentist procedures have a clear Bayesian interpretation.

Uncertainty quantification issues

Main problem: given the available information, what is the probability distribution (pdf) of uncertainties?

Entangled problems of UQ



And more… Design of experiments, sensitivity analysis, etc.

Bayesian parameter estimation

Consider a model or theory with k parameters $\vec{a} = \{a_0, a_1, \dots, a_{k-1}\}$ which we wish to constrain with N measured data $D = \{d_1, d_2, \dots, d_N\}$, given background information I

Goal: estimate the pdf $\mathrm{pr}(\vec{a}|D, I)$

Bayes theorem allows us to actually compute this pdf:

$$\mathrm{pr}(\vec{a}|D, I) = \frac{\mathrm{pr}(D|\vec{a}, I)\ \mathrm{pr}(\vec{a}|I)}{\mathrm{pr}(D|I)}$$

Names for each of these terms:

- Posterior: $\mathrm{pr}(\vec{a}|D, I)$
- Likelihood: $\mathrm{pr}(D|\vec{a}, I)$
- Prior: $\mathrm{pr}(\vec{a}|I)$
- Evidence/ marginal likelihood (normalization factor):

$$\mathrm{pr}(D|I) = \int d\vec{a}\ \mathrm{pr}(D|\vec{a}, I)\ \mathrm{pr}(\vec{a}|I)$$

Example parameter estimation problem

The problem (developed to mock up an effective field theory expansion):

- Generate synthetic data with indep. Gaussian noise from "real-world" $g(x)$

$$g(x) = \left(\frac{1}{2} + \tan\left(\frac{\pi}{2}x\right)\right)^2 = 0.25 + 1.57x + 2.47x^2 + 1.29x^3 + \ldots$$

- Given data and series expansion, estimate coefficients of Taylor series (about $x = 0$) up to some order. Truncated polynomial is the theory $g_{\rm th}(x)$ with $k+1$ parameters $\vec{a}$

$$g_{\rm th}(x) = \sum_{n=0}^{k} a_n x^n$$

- This is linear optimization, unlike most problems in nuclear physics.

- Problem is simple but helpful for intuition about many statistical issues.

Schindler and Phillips, Annals Phys. 324, 3 (2009)
SW et al., JPG 43, 074001 (2016)

Example parameter estimation problem

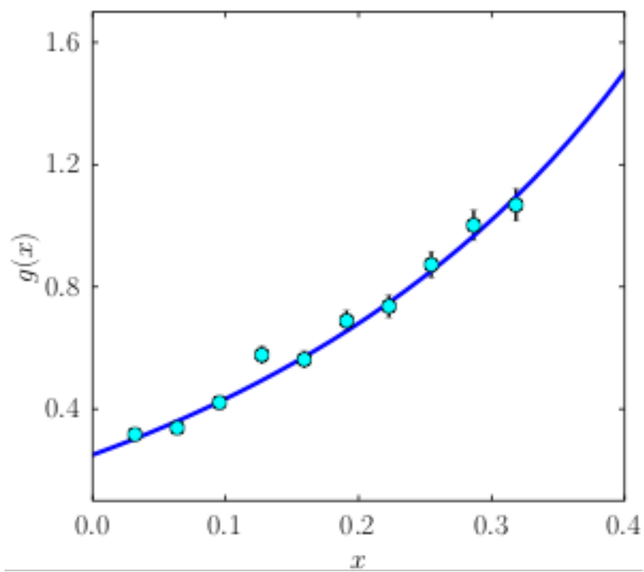
Given data D, estimate and plot the posterior pdf of the parameters $\vec{a}$

The resulting pdf $\mathrm{pr}(\vec{a}|D, I)$ can be used to propagate one piece of uncertainty to the final prediction $g = g_{\mathrm{th}} + \delta g_{\mathrm{th}}$. Here

$$\delta g_{\mathrm{th}} = (\delta g_{\mathrm{th}})_{\mathrm{params}} + (\delta g_{\mathrm{th}})_{\mathrm{trunc}}$$

Also have model discrepancy due to truncation of the Taylor polynomial!

The data: D and "real world" function:



Example parameter estimation problem

$$\text{pr}(\vec{a}|D, I) = \frac{\text{pr}(D|\vec{a}, I)\ \text{pr}(\vec{a}|I)}{\text{pr}(D|I)}$$

becomes, with normal, independent data $D = \{d_i\}_{i=1}^N$ with standard deviations $\{\sigma_i\}_{i=1}^N$ and a uniform (bounded) prior $\text{pr}(\vec{a}|I) \propto 1$

$$\text{pr}(\vec{a}|D, I) \propto e^{-\chi^2/2}$$

where

$$\chi^2(\vec{a}) = \sum_{i=1}^{N} \left(\frac{d_i - g_{\text{th}}(x_i; \vec{a})}{\sigma_i} \right)^2$$

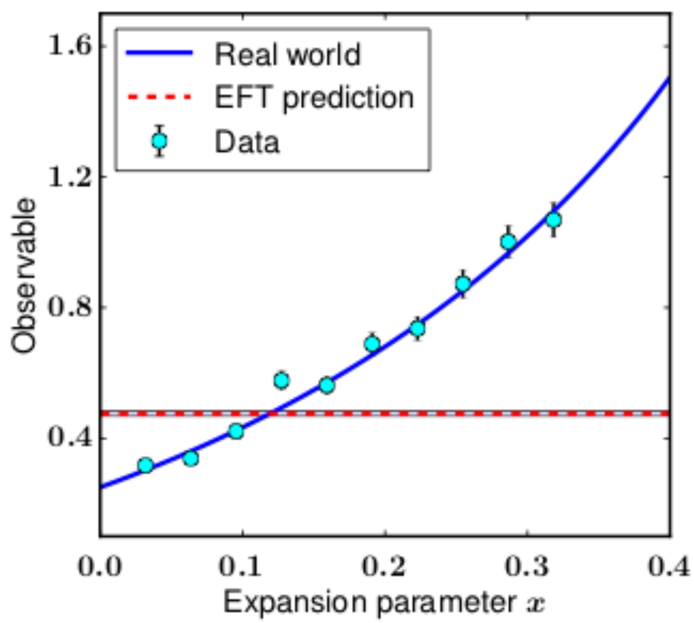
This is the standard least-squares optimization result. For this example, it can be solved analytically using standard results.

Simple parameter estimation problem

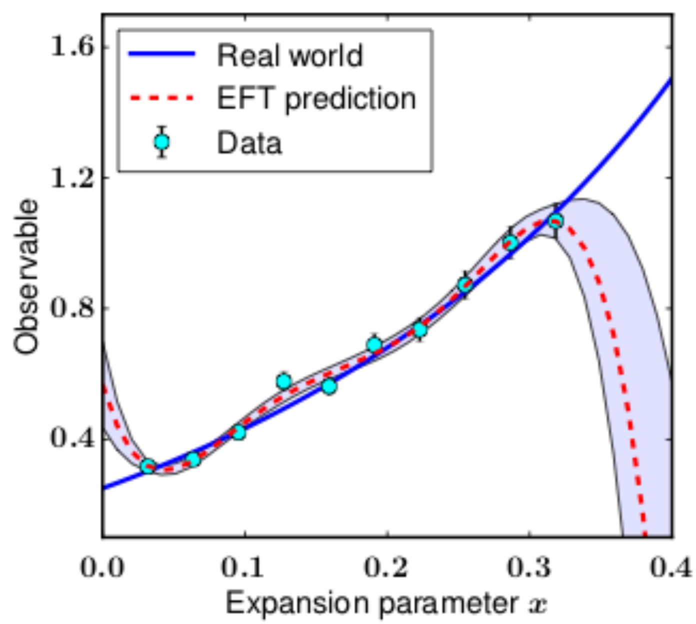
What happens to this problem when you use least-squares?

$$\mathrm{pr}(\vec{a}\,|\,D, I) \propto e^{-\chi^2/2}$$

Underfitting: $g_{\rm th} = a_0$



Overfitting: $g_{\rm th} = \sum_{n=0}^{5} a_n x^n$



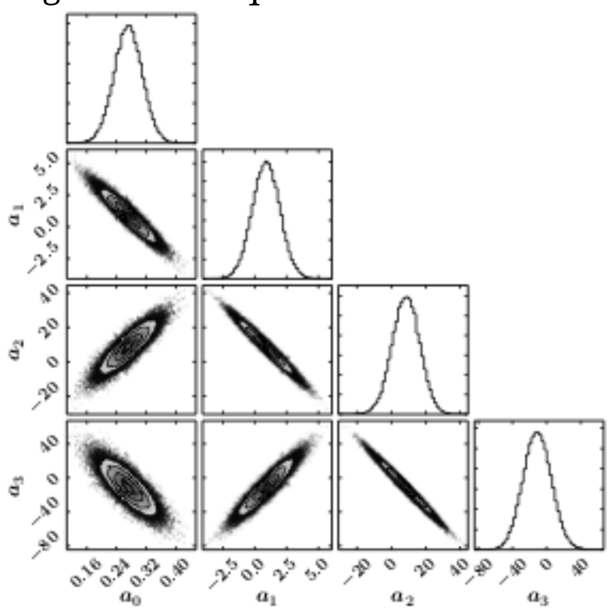
Simple parameter estimation problem

Use information that Taylor series coefficients are "natural" (EFT principle).

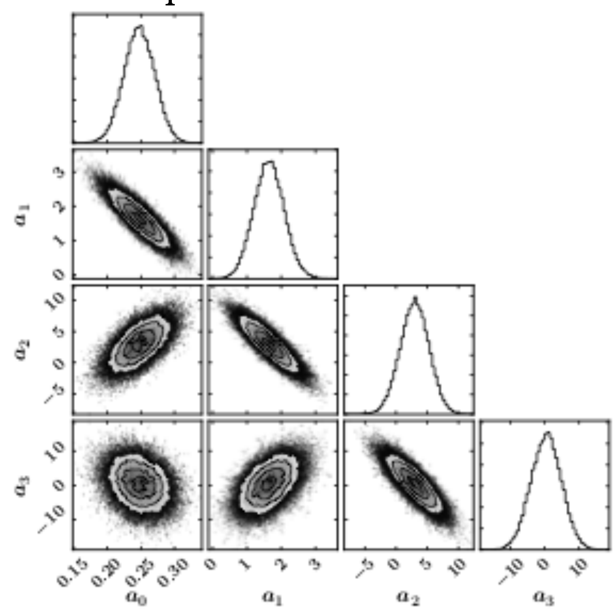
$$\mathrm{pr}(\vec{a}|I) \propto e^{-\vec{a}^2/(2\,\bar{a}^2))}$$

$$\mathrm{pr}(\vec{a}|D, I) \propto e^{-\chi^2/2-\vec{a}^2/(2\,\bar{a}^2)}$$

Regular least-squares



Gaussian prior with width 5



see also: regularized least-squares

More issues in this problem

- Check for robustness to the prior pdf $\mathrm{pr}(\vec{a}|I)$
- Include impact of higher-order terms (theory discrepancy)

$$g(x) = \sum_{n=0}^{k} a_n x^n + \sum_{n=k+1}^{\infty} a_n x^n$$

- Simple to do in this linear problem, but nonlinear problems:
    - not analytic
    - sampling objective function can be costly

$$\chi^2(\vec{a}) = \sum_{i=1}^{N} \left(\frac{d_i - g_{\mathrm{th}}(x_i; \vec{a})}{\sigma_i} \right)^2$$

    - have to result to sampling: Markov Chain Monte Carlo (MCMC)
    - non-normality, multimodality
- Validation
- UQ: theory discrepancy and parameter uncertainty (and everything else)

Marginalization

Marginalization "integrates out" nuisance parameters by summing over a complete set of possibilities:

$$\mathrm{pr}(A) = \int dB p(A, B) = \int dB p(A|B)p(B)$$

Previous example: include effects of unconstrained higher-order terms!

Useful for introducing auxiliary parameters, e.g.:

$$\mathrm{pr}(\vec{a}|I) = \int d\bar{a}\ \mathrm{pr}(\vec{a}|\bar{a}, I)\,\mathrm{pr}(\bar{a}|I)$$

where $\bar{a}$ is the natural width of the prior. This can be used to avoid too tightly specifying a single value for such a parameter.

But we must now specify a prior on the hyperparameter $\bar{a}$.

Previous plots $\mathrm{pr}(\bar{a}|I) = \delta(\bar{a} - 5)$

Marginalization is used to plot and study pdfs as well, for example:

$$\mathrm{pr}(a_0|D, I) = \int da_1 \cdots da_k\, \mathrm{pr}(\vec{a}|D, I)$$

Approximating a pdf as a histogram

Sampling

$$\mathrm{pr}(\vec{a}|D, I)$$

may be expensive for nonlinear problems and intensive observables.

We use MCMC sampling to evaluate the pdf at many values of $\vec{a}$, and histogram the samples obtained

Many flavors of MCMC on the market in a variety of languages.
Some python packages: emcee [used here], pymc3, pySTAN.

Also nested sampling: pyMultiNest

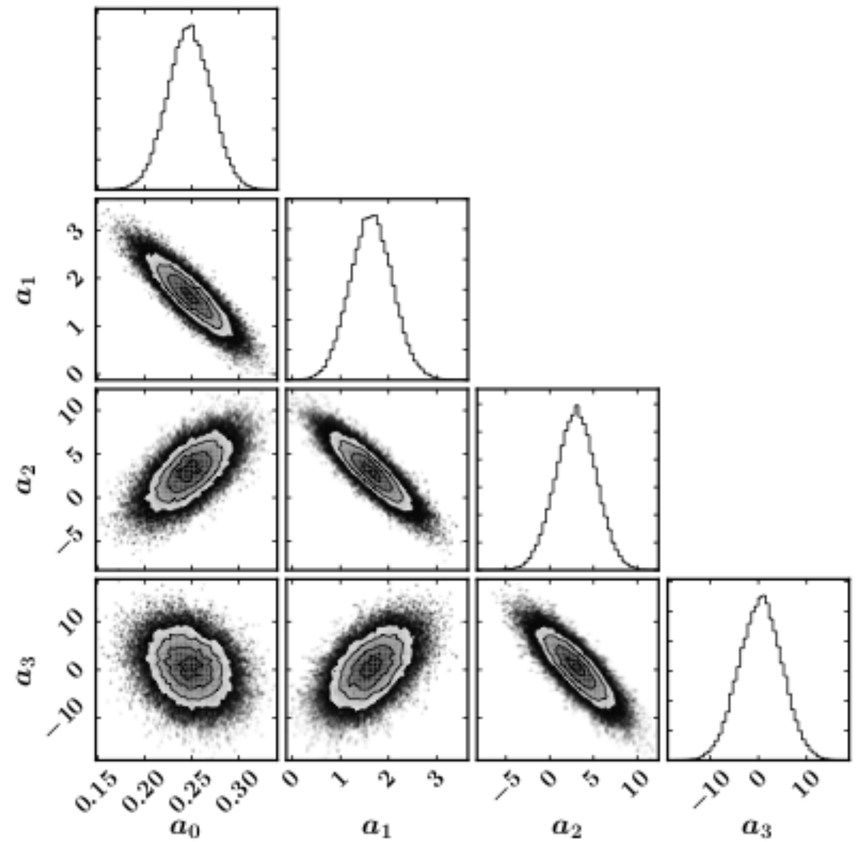
Sampling can still be infeasible. Get around the problem with emulation:

- Bayesian optimization Ekström et al. JPhysG 46, 9 (2019)
- Eigenvector continuation Frame et al., PRL 121 032501 (2018)

Conjugate priors can simplify things because posterior is analytic
see Melendez, SW, et al. PRC 100 (2019)

Approximating a pdf as a histogram

Marginalization is trivial over various parameters, just use samples in the parameters you want:



Quantitative model comparison with Bayes

Very generic formulation: model 1 (M_1) and model 2 (M_2). Examples:

- totally different theories for a phenomenon
- hierarchical models (one order vs. next)
- anything else that can be used to compute data

Using Bayes theorem and assuming models are *a priori* equally likely

$$\frac{\mathrm{pr}(M_1|D, I)}{\mathrm{pr}(M_2|D, I)} = \frac{\int d\vec{a}_1\, \mathrm{pr}(\vec{a}_1|D, I)}{\int d\vec{a}_2\, \mathrm{pr}(\vec{a}_2|D, I)}$$

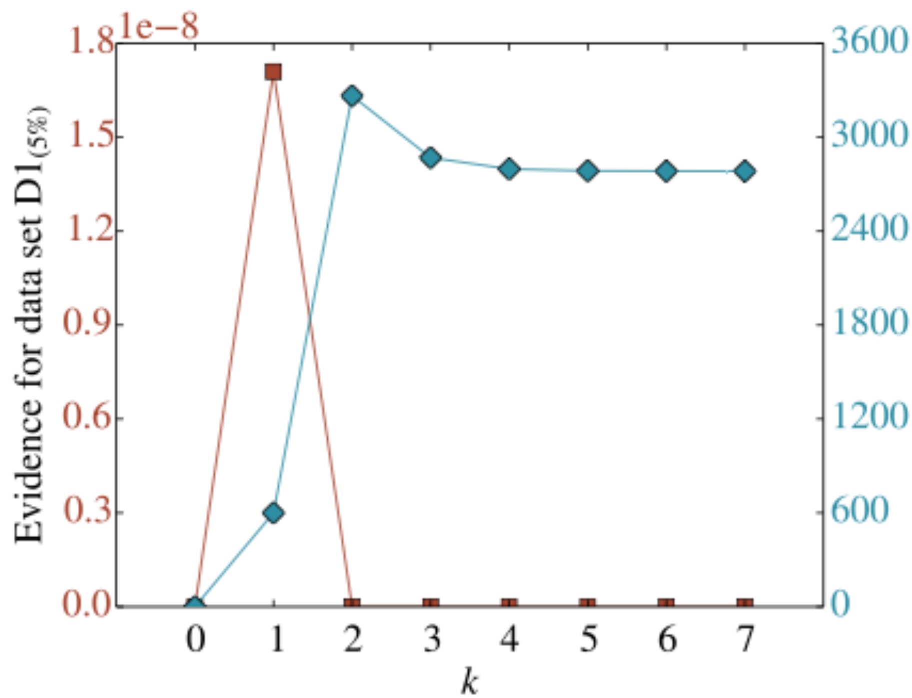
"Evidence ratio" or "Odds factor"

- Expensive to compute if integrals are not analytic
- MCMC samples of the posterior don't help (need normalization!)
- Tricky to interpret

In our simple example, though, it is computable and has a clear interpretation

Evidence calculation for Taylor series model

Compute evidence at ascending orders in theory $k = 0, 1, 2, \dots ,$



Compare least-squares (brown squares) with Gaussian prior (blue diamonds)

Natural result of Occam's razor for LS result. Saturates for Gaussian prior.

Our work using Bayes for χEFT

- The BUQEYE (Bayesian Uncertainty Quantification: Errors in Your EFT) collaboration work with low-energy nuclear EFTs:
  - parameter estimation of EFT low-energy constants
  - using Gaussian processes to model EFT truncation error
  - diagnostics to validate uncertainty estimates

Summary

- Bayesian statistics is ideal for UQ problems in physics
- Explicit, quantitative incorporation of prior information
- Like traditional methods, it can be expensive to fully implement
- But taking the time to understand and sample pdfs can yield dividends
    - proper inclusion of theory errors
    - are pdfs Gaussian, and are covariance approximations justified?
- Many approximation methods (like MCMC sampling) and emulation schemes are possible
- Bayesian methods and statistical analysis give a new window into nuclear theory, allowing diagnostics and validation of theory expectations from a data-driven perspective!
- Visit the BUQEYE collaboration online at buqeye.github.io
- A very nice place to learn more about Bayes for nuclear physicists: nucleartalent.github.io/Bayes2019/

References

- P. Gregory, "Bayesian Logical Data Analysis for the Physical Sciences"
- Schindler and Phillips Annals Phys. 324, 3 (2009)
- Wesolowski et al., JPG 43, 074001 (2016): "Bayesian parameter estimation for effective field theories"
- Ekström et al. JPhysG 46, 9 (2019): "Bayesian optimization in ab initio nuclear physics"
- Frame et al., PRL 121 032501 (2018): "Eigenvector Continuation with Subspace Learning"
- Melendez et al. PRC 100 (2019): "Quantifying Correlated Truncation Errors in Effective Field Theory"

General Bayes/Machine Learning references

Bayesian statistics

- D.S. Sivia and J. Skilling, "Data Analysis: A Bayesian Tutorial"
- P. Gregory, "Bayesian Logical Data Analysis for the Physical Sciences"
- Gelman et al., "Bayesian Data Analysis" (3rd edition)
- R. Trotta, "Bayes in the sky: Bayesian inference and model selection in cosmology"

Other stuff (like Gaussian processes)

- Rasmussen and Williams, "Gaussian processes for machine learning"
- D. J.C. MacKay, "Introduction to Gaussian processes"
- D. J.C. MacKay, "Information Theory, Inference, and Learning Algorithms"