Draft version July 15, 2022
Typeset using LaTeX/show your work! **modern** style in AASTeX631

Accumulation of Bayesian Evidence In Hierarchical Population Models

Will M. Farr[1,2] and and friends

[1] *Department of Physics and Astronomy, Stony Brook University, Stony Brook NY 11794, USA*
[2] *Center for Computational Astrophysics, Flatiron Institute, New York NY 10010, USA*

ABSTRACT

We consider how evidence accumulates from repeated observations in models that introduce a continuous deformation parameter to an underlying physical theory. We imagine that the true deformation of the theory is constant from observation to observation, and explore models where the deformation is assumed constant but unknown from observation to observation or where the deformation is drawn independently from a Gaussian with unknown mean and variance for each observation. Both these models can recover the true data generating process at specific parameter values, but also admit parameter values that do not match the true data generating process. In the limit of a large number of independent observations, these models have reduced Bayesian evidence relative to the true model for the data generating process. But we find that the log evidence for these models relative to the true model is reduced by a term that grows only *logarithmically* in the number of observations; the model with a constant deformation parameter for each observation suffers half the (logarithmic) reduction relative to the model where the deformation parameter is normally distributed across observations. But both models suffer much less reduction in log evidence compared to a model where the deformation is fixed to an incorrect value for each observation, for which the log evidence relative to the true model decreases *linearly* in the number of observations. Thus we conclude that both models are about equally "efficient" at detecting or ruling out deformations from the underlying physical theory; but we advocate for the model with normally distributed deformation parameters across observations because it is robust to true data generating processes where the deformation *is not* constant from observation to observation.

1. INTRODUCTION

Consider a setup designed to test a theory by introducing a continuous deformation parameter, x. When $x = 0$, the underlying theory is recovered, while for $x \neq 0$ the theory is modified in some way. Imagine that we make (noisy) observations of the parameter x in a number of different situations, and we wish to use these

observations to determine if $x \equiv 0$ is consistent with our observations or not. This is (approximately) the setup in many tests of GR using gravitational waves.

Suppose further that our observations of x are such that (a) the noise is independent in each observation and (b) in each observation the likelihood for the observed value of x, x_o, is normal about the latent value of x at the time of the observation with known uncertainty, σ_o:

$$x_o \sim N\left(x, \sigma_o\right). \tag{1}$$

We can build several different models for combining multiple observations to better constrain the value of x. If we are willing to make the restrictive assumption that there is one, single, true value of x across all our observations (perhaps our extension to the base theory has a single, common parameter that controls the deviation in all circumstances), then after $i = 1, \ldots, N$ observations, the combined likelihood of the N observations $x_{o,i}$ is

$$p\left(\{x_{o,i} \mid i = 1, \ldots, N\} \mid x\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_o^2}} \exp\left[-\frac{\left(x_{o,i} - x\right)^2}{2\sigma_o^2}\right]. \tag{2}$$

Alternately, if we are unwilling to commit to such a restrictive assumption, and want to allow for the possibility that there is a different (latent) value of x_i for each observation, Isi et al. (2019) recommend imposing a Gaussian population assumption for how the latent x_i parameters appear in our observations (Isi et al. 2019, 2022). Each observation has a different value of x, denoted x_i, which are drawn from a Gaussian distribution with mean μ and standard deviation σ:

$$x_i \sim N\left(\mu, \sigma\right). \tag{3}$$

In this model, the marginal likelihood for each observation $x_{o,i}$ is also Gaussian with

$$\begin{aligned} p\left(x_{o,i} \mid \mu, \sigma\right) &= \int \mathrm{d}x_i \, p\left(x_{o,i} \mid x_i\right) p\left(x_i \mid \mu, \sigma\right) \\ &= \frac{1}{\sqrt{2\pi\left(\sigma^2 + \sigma_o^2\right)}} \exp\left[-\frac{\left(x_{o,i} - \mu\right)^2}{2\left(\sigma^2 + \sigma_o^2\right)}\right]. \end{aligned} \tag{4}$$

Combining many observations to learn about μ and σ gives the joint likelihood

$$p\left(\{x_{o,i} \mid i = 1, \ldots, N\} \mid \mu, \sigma\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left(\sigma^2 + \sigma_o^2\right)}} \exp\left[-\frac{\left(x_{o,i} - \mu\right)^2}{2\left(\sigma^2 + \sigma_o^2\right)}\right] \tag{5}$$

The model with a single value of x common to all observations is recovered as a special case of this model when $\sigma \to 0$ and $\mu \to x$ (Isi et al. 2019, 2022).

Suppose that, in fact, there is a single true value of x common to all the observations, denoted x_t. In the limit of a large number of observations, the likelihoods in Eqs. (2) and (5) will asymptote to the exponential of N times the mean log-likelihood,

with additional terms at $o(N)$. Thus, to explore the asymptotic behavior of the two models it is sufficient to calculate the mean log likelihood. When there is a single, true value of $x \equiv x_t$, and the assumed Gaussian likelihood is the correct description of the data generating process, the distribution of x_o in each observation is

$$x_o \sim N(x_t, \sigma_o). \tag{6}$$

For the model with a single, shared value of x across all observations, the average log-likelihood in each observation is

$$\begin{aligned}
&\langle \log p(x_o \mid x) \rangle \\
&\quad = \int \mathrm{d}x_o \left(-\frac{1}{2}\log\left(2\pi\sigma_o^2\right) - \frac{(x_o - x)^2}{2\sigma_o^2} \right) \frac{1}{\sqrt{2\pi\sigma_o^2}} \exp\left[-\frac{(x_o - x_t)^2}{2\sigma_o^2} \right] \\
&\quad = -\frac{1}{2}\log\left(2\pi\sigma_o^2\right) - \frac{(x - x_t)^2}{2\sigma_o^2} - \frac{1}{2}.
\end{aligned} \tag{7}$$

So Eq. (2) becomes in the large N limit

$$p(\{x_{o,i} \mid i = 1, \ldots, N\} \mid x) \simeq \left(2\pi\sigma_o^2\right)^{-N/2} \exp\left[-\frac{N(x - x_t)^2}{2\sigma_o^2} - \frac{N}{2} \right]. \tag{8}$$

Similarly, for the model with Gaussian distributed x_i, the average log-likelihood in each observation is

$$\begin{aligned}
&\langle \log p(x_o \mid \mu, \sigma) \rangle \\
&\quad = \int \mathrm{d}x_o \left(-\frac{1}{2}\log\left(2\pi\left(\sigma^2 + \sigma_o^2\right)\right) - \frac{(x_o - \mu)^2}{2\left(\sigma^2 + \sigma_o^2\right)} \right) \frac{1}{\sqrt{2\pi\sigma_o^2}} \exp\left[\frac{(x_o - x_t)^2}{2\sigma_o^2} \right] \\
&\quad = -\frac{1}{2}\log\left(2\pi\left(\sigma^2 + \sigma_o^2\right)\right) - \frac{(\mu - x_t)^2 + \sigma_o^2}{2\left(\sigma^2 + \sigma_o^2\right)}.
\end{aligned} \tag{9}$$

So Eq. (5) becomes in the large N limit

$$p(\{x_{o,i} \mid i = 1, \ldots, N\} \mid \mu, \sigma) \simeq \left(2\pi\left(\sigma^2 + \sigma_o^2\right)\right)^{-N/2} \exp\left[-N\frac{(\mu - x_t)^2 + \sigma_o^2}{2\left(\sigma^2 + \sigma_o^2\right)} \right]. \tag{10}$$

Both models give asymptotically consistent maximum likelihood estimates for their parameters. Eq. (8) is maximized when $x = x_t$. Eq. (10) is maximized when $\mu = x_t$ and $\sigma^2 = \sigma_o^2/(N-1) \to 0$ as $N \to \infty$. (Note that we must maximize over the parameter σ^2; maximizing over σ instead introduces an essential singularity at $\sigma = \sqrt{\sigma^2} = 0$, precisely where the true model is recovered.)

The evidence in each model after N observations is an integral over the shared parameters (x or μ and σ) of the likelihoods derived above and a prior on x or μ and σ. If we assume that we have enough observations to constrain x or μ and σ to scales

much smaller than the prior (i.e. that we are likelihood dominated), then the prior is approximately constant, and the evidence becomes the integral of the likelihood times the prior evaluated at the peak likelihood parameters. For the model with a single, shared value of x, we have

$$\begin{aligned} &p\left(\{x_{o,i} \mid i = 1, \ldots, N\} \mid \text{shared}\right) \\ &\qquad \simeq p_{\text{shared}}\left(x_t\right) \int \mathrm{d}x \, \left(2\pi\sigma_o^2\right)^{-N/2} \exp\left[-\frac{N\left(x - x_t\right)^2}{2\sigma_o^2} - \frac{N}{2}\right] \\ &\qquad = \frac{p_{\text{shared}}\left(x_t\right)}{\sqrt{N}\left(2\pi\sigma_o^2\right)^{(N-1)/2}} \exp\left(-\frac{N}{2}\right), \end{aligned} \tag{11}$$

with $p_{\text{shared}}\left(x\right)$ the prior we impose on the shared parameter x.

For the model with Gaussian distributed values of x_i in each observation, we have[1]

$$\begin{aligned} &p\left(\{x_{o,i} \mid i = 1, \ldots, N\} \mid \text{Gaussian}\right) \\ &\qquad \simeq p_{\text{Gaussian}}\left(x_t, \frac{\sigma_o^2}{N-1}\right) \\ &\qquad \times \int \mathrm{d}\mu \, \mathrm{d}\sigma^2 \, \left(2\pi\left(\sigma^2 + \sigma_o^2\right)\right)^{-N/2} \exp\left[-N\frac{\left(\mu - x_t\right)^2 + \sigma_o^2}{2\left(\sigma^2 + \sigma_o^2\right)}\right] \\ &\qquad = \frac{1}{2\pi N} \frac{p_{\text{Gaussian}}\left(x_t, \frac{\sigma_o^2}{N-1}\right)}{\left(\pi N \sigma_o^2\right)^{(N-3)/2}} \left(\Gamma\left(\frac{N-3}{2}\right) - \Gamma\left(\frac{N-3}{2}, \frac{N}{2}\right)\right), \end{aligned} \tag{12}$$

where $p_{\text{Gaussian}}\left(\mu, \sigma^2\right)$ is the prior we impose on μ and σ^2, $\Gamma\left(\cdot\right)$ is the Gamma function, and $\Gamma\left(\cdot, \cdot\right)$ is the incomplete Gamma function. This expression is not terribly illuminating. If instead we use a Gaussian approximation to the marginal likelihood for σ^2 about $\sigma^2 = \sigma_o^2 / \left(N - 1\right)$ with width $2N\sigma_o^2 / \left(N - 1\right)^{3/2}$, and keep only the leading terms as $N \to \infty$, we find[2]

$$\begin{aligned} &p\left(\{x_{o,i} \mid i = 1, \ldots, N\} \mid \text{Gaussian}\right) \\ &\qquad \simeq p_{\text{Gaussian}}\left(x_t, \frac{\sigma_o^2}{N-1}\right) \frac{1}{N\sqrt{2\pi}\left(2\pi\sigma_o^2\right)^{(N-1)/2}} \exp\left(-\frac{N}{2}\right). \end{aligned} \tag{13}$$

Because the model with the single, shared value of x for each observation is actually the true model in the situation we are considering, it is not surprising that it achieves larger evidence (up to prior terms); the Gaussian population model only recovers the truth for a specific point in its two-dimensional parameter space. However, the loss of log-evidence from the Gaussian model is only logarithmic in the number of observations (the $1/N$ term instead of $1/\sqrt{N}$ pre-factor), while the overall log-evidence for each model accumulates linearly in N, as expected.

[1] Once again, we are treating σ^2 as the parameter of interest not σ.

[2] Note that the peak of the marginal likelihood for σ^2 is much closer to 0 than the width, so we have divided the usual Gaussian approximation by 2 to account for the (almost) half-normal integral.

We can compare both modes to the true model used to generate the observations, which is the shared *and known* value of $x \equiv x_t$. The evidence for this zero-parameter model is given by Eq. (2) with $x = x_t$, which is

$$p\left(\{x_{o,i} \mid i = 1, \ldots, N\} \mid x = x_t\right) \simeq \left(2\pi\sigma_o^2\right)^{-N/2} \exp\left[-\frac{N}{2}\right]. \tag{14}$$

Both the model with shared but unknown x and the model with Gaussian distributed x suffer logarithmic penalties in the log-evidence compared to this "perfect" model. In the large-N limit, the log-evidence penalty for the shared but unknown x is half the penalty suffered by the Gaussian model; but both models have much smaller penalty than, say, the base physical model ($x \equiv 0$) when $x_t \neq 0$, which has evidence

$$p\left(\{x_{o,i} \mid i = 1, \ldots, N\} \mid x = 0\right) \simeq \left(2\pi\sigma_o^2\right)^{-N/2} \exp\left[-\frac{N x_t^2}{2\sigma_o^2} - \frac{N}{2}\right] \tag{15}$$

and suffers a linear penalty in log-evidence relative to the true model at large N.

We thank Tom Callister for comments on the manuscript.

REFERENCES

Isi, M., Chatziioannou, K., & Farr, W. M. 2019, PhRvL, 123, 121101, doi: 10.1103/PhysRevLett.123.121101

Isi, M., Farr, W. M., & Chatziioannou, K. 2022, arXiv e-prints, arXiv:2204.10742. https://arxiv.org/abs/2204.10742