Draft version July 2, 2021
Typeset using LaTeX **modern** style in AASTeX63

Re-Weighting Existing Samples to a Population Analysis

Will M. Farr[1,2] and Thomas A. Callister[2]

[1] *Department of Physics and Astronomy, Stony Brook University, Stony Brook NY 11794, United States*

[2] *Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York NY 10010, United States*

ABSTRACT

We show how to re-weight pre-existing parameter samples to obtain samples distributed according to a hierarchical population analysis. We discuss the behavior of the marginalized distribution for single-event parameters in a hierarchical population analysis.

For an alternative presentation of essentially identical material, see Hogg et al. (2010).

It is a common problem in astronomy to have a collection of observations of some objects from a population in need of a simultaneous analysis of the population properties and object properties. Such analyses are called "hierarchical" because they naturally separate into several distinct "levels." To be a bit more mathematically precise: we are presented with a set of data,

$$D \equiv \left\{ d_i \mid i = 1, \ldots, N \right\}, \tag{1}$$

consisting of N distinct data sets, d_i, each representing some measurement of an object. Each object may have parameters, θ_i, that are of interest. We think that the set of parameters,

$$\Theta \equiv \left\{ \theta_i \mid i = 1, \ldots, N \right\}, \tag{2}$$

comes to us as fair draws[1] from a population distribution, which may in turn depend on some paramteters, λ:

$$\theta_i \sim p\left(\theta \mid \lambda\right). \tag{3}$$

will.farr@stonybrook.edu

wfarr@flatironinstitute.org

thomas.callister@flatironinstitute.org

[1] If the draws are not fair, then the sample is said to suffer from *selection effects*; in this case see Loredo (2004); Messenger & Veitch (2013); Mandel et al. (2019). Loredo (2004); Mandel et al. (2019) also describe how to fit a *rate* of objects from such observations.

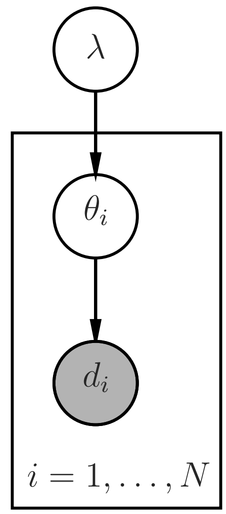

Figure 1. A graphical description of our hierarchical model. Each node in the graph is a variable. Shaded nodes are *observed* variables whose values are conditioned on in the analysis. An arrow connecting two nodes represents a distribution of the target node conditioned on the source node's value.

We assume that we know enough about the data generating, or measurement, process that we can write down a probabilistic description of the data conditioned on the parameters θ; this function is commonly called the "likelihood:"

$$d_i \sim p\left(d \mid \theta_i\right). \tag{4}$$

(Note that θ_i can contain both parameters that are *intrinsic* to the object, and also parameters that describe the measurement process for that object—detector noise levels, or calibration parameters, for example—and that the population distribution can be both an *intrinsic* population and also a model of the *distribution* of such measurement parameters.) A graphical description of our hierarchical model appears in Figure 1.

If all these distributions are available to us, then it is straightforward to sample over the joint distribution of Θ and λ given the data D. Impose a prior on λ, $p\left(\lambda\right)$; then[2]

$$p\left(\Theta, \lambda \mid D\right) \propto \left[\prod_{i=1}^{N} p\left(d_i \mid \theta_i\right) p\left(\theta_i \mid \lambda\right)\right] p\left(\lambda\right), \tag{5}$$

and your favorite stochastic sampling method™ can be used to draw samples in the (possibly high-dimensional) space of Θ and λ. However, the more common situation is that we are provided instead with a *catalog* of objects and parameters already

[2] We are here implicitly assuming that the data generating process is such that *once conditioned on parameters* θ successive observations are independent of each other. This is almost certainly false, but may be "true enough" for our purposes, particularly if there are parameters describing systematic or "calibration" effects in our instrument in each θ_i.

inferred from them according to some prior. Thus, we have a set of M samples,

$$\left\{\theta_i^{(j)} \mid j = 1, \ldots, M\right\}, \tag{6}$$

where each sample is drawn from a posterior with a prior p_0:

$$\theta_i^{(j)} \sim p\left(d_i \mid \theta\right) p_0\left(\theta\right). \tag{7}$$

A common trick in this situation (Hogg et al. 2010) is to give up on sampling in Θ, and integrate the θ_i out of Eq. (5):

$$p\left(\lambda \mid D\right) \propto \left[\prod_{i=1}^{N} \int \mathrm{d}\theta_i p\left(d_i \mid \theta_i\right) p\left(\theta_i \mid \lambda\right)\right] p\left(\lambda\right) \equiv p\left(D \mid \lambda\right) p\left(\lambda\right). \tag{8}$$

The integrals inside the product can be approximated (up to ignorable constants) using importance sampling with the samples from our catalog, $\theta_i^{(j)}$:

$$p\left(\lambda \mid D\right) \propto \left[\prod_{i=1}^{N} \left\langle \frac{p\left(\theta \mid \lambda\right)}{p_0\left(\theta\right)} \right\rangle_{\theta_i^{(j)}}\right] p\left(\lambda\right), \tag{9}$$

where the average is taken over the samples $\theta_i^{(j)}$. Here the ratio of the population distribution to the catalog prior is the *importance weight* for each of the $\theta_i^{(j)}$. Depending on the number of samples associated to each object in the catalog and the relative widths of the likelihood, catalog prior, and population, this procedure can go awry; but often it is good enough, and its simplicity combined with the (often dramatic) reduction in dimensionality (the components of Θ often dominate the number of degrees of freedom in λ) argues for using it when possible.

Sometimes we are only interested in the population parameters, λ; in this case the individual-object parameters Θ are a nuisance anyway, and we need not worry about integrating them out. However, we are often interested in *both* the population and individual-level parameters; and sometimes we are only using the population to improve estimates of individual-level parameters by (partially) pooling information across observations (Lieu et al. 2017). In this case we must recover samples of Θ after we have samples from the marginal distribution over population parameters.

Recall that it is a theorem of probability that

$$p\left(\Theta, \lambda \mid D\right) = p\left(\Theta \mid \lambda, D\right) p\left(\lambda \mid D\right). \tag{10}$$

If we are given samples, $\lambda^{(k)}$, drawn from $p\left(\lambda \mid D\right)$,

$$\lambda^{(k)} \sim p\left(\lambda \mid D\right), \tag{11}$$

then augmenting each $\lambda^{(k)}$ with a set $\Theta^{(k)}$, where each $\theta_i^{(k)}$ is drawn from

$$\theta_i^{(k)} \sim p\left(d_i \mid \theta\right) p\left(\theta \mid \lambda^{(k)}\right), \tag{12}$$

will produce a draw from the joint distribution over Θ and λ. This works because

$$p\left(d_i \mid \theta\right) p\left(\theta \mid \lambda^{(k)}\right) \propto p\left(\theta \mid \lambda, D\right), \tag{13}$$

with the constant of proportionality *independent* of Θ at fixed D and λ. A draw from Eq. (12) can be accomplished by choosing randomly one of the existing catalog samples $\theta_i^{(j)}$ with weight, $w_i^{(j)}$, proportional to (Hogg et al. 2010)

$$w_i^{(j)} \propto \frac{p\left(\theta_i^{(j)} \mid \lambda^{(k)}\right)}{p_0\left(\theta_i^{(j)}\right)}. \tag{14}$$

So, to summarize, here is the algorithm for sampling from the joint distribution of Θ and λ given D and a catalog of samples, $\theta_i^{(j)}$ drawn from a catalog posterior with prior $p_0\left(\theta\right)$.

1. Use a stochastic sampler to draw samples $\lambda^{(k)}$ from Eq. (9).
2. For each sample, λ_k, and each object i, draw a random catalog sample, $\theta_i^{(k)}$ from the $\theta_i^{(j)}$ with weights given by Eq. (14).

The pairs of $\lambda^{(k)}$, and the associated set of catalog draws, $\Theta^{(k)}$, constitute a sample from the joint distribution on Θ and λ defined in Eq. (5), and can be used to estimate population properties, individual-event properties informed by a population, correlations between population properties and individual-event properties, etc.

In the case where we are not interested in samples λ at all, we can also integrate out λ in Eq. (10). We will need the properly-normalized version of Eq. (12) since the normalization depends on λ:

$$p\left(\theta_i \mid \lambda, D\right) = \frac{p\left(d_i \mid \theta_i\right) p\left(\theta_i \mid \lambda\right)}{p\left(d_i \mid \lambda\right)} = \frac{p\left(d_i \mid \theta_i\right) p\left(\theta_i \mid \lambda\right)}{\int \mathrm{d}\theta\, p\left(d_i \mid \theta\right) p\left(\theta \mid \lambda\right)}. \tag{15}$$

Then the marginal distribution for θ_i is

$$p\left(\theta_i \mid D\right) \propto p\left(d_i \mid \theta_i\right) \int \mathrm{d}\lambda\, \frac{1}{p\left(d_i \mid \lambda\right)} p\left(\theta_i \mid \lambda\right) p\left(\lambda \mid D\right). \tag{16}$$

Note that the evidence for object i modifies what otherwise would be the posterior expectation over λ of the population distribution for θ_i. Assuming, again, that we have posterior samples $\theta_i^{(k)}$ drawn from a catalog with prior $p_0\left(\theta\right)$ and posterior samples for λ drawn from the marginal posterior, $\lambda^{(l)} \sim p\left(\lambda \mid D\right)$, this can be approximated as

$$p\left(\theta_i \mid D\right) \propto p\left(d_i \mid \theta_i\right) \left\langle p\left(\theta_i \mid \lambda^{(l)}\right) \left[\left\langle \frac{p\left(\theta_i^{(k)} \mid \lambda^{(l)}\right)}{p_0\left(\theta_i^{(k)}\right)} \right\rangle_{\theta_i^{(k)}} \right]^{-1} \right\rangle_{\lambda^{(l)}}, \tag{17}$$

or, expressed as importance weights for resampling the $\theta_i^{(k)}$

$$w_i^{(k)} \propto \frac{1}{p_0\left(\theta_i^{(k)}\right)} \left\langle p\left(\theta_i^{(k)} \mid \lambda^{(l)}\right) \left[\left\langle \frac{p\left(\theta_i^{(k')} \mid \lambda^{(l)}\right)}{p_0\left(\theta_i^{(k')}\right)} \right\rangle_{\theta_i^{(k')}} \right]^{-1} \right\rangle_{\lambda^{(l)}}, \quad (18)$$

where we have introduced the index k' to the inner expectation value to emphasize that it should be taken independently of the outer expectation over samples of θ_i. This expression is equivalent to Eq. (6) of Callister (2019).

Eq. (16) and its implementation in Eqs. (17) and (18) are equivalent to imposing a prior on θ_i that comes from the population distribution weighted by the "leave one out" posterior on λ:[3]

$$p\left(\theta_i \mid D\right) \propto p\left(d_i \mid \theta_i\right) \int \mathrm{d}\lambda \, p\left(\theta_i \mid \lambda\right) p\left(\lambda \mid D \backslash d_i\right). \quad (19)$$

[3] It seems like this must be a well-known fact—probably it appears in Gelman et al. (2013)—but I am not familiar with it.

REFERENCES

Callister, T. 2019, Reweighting Single Event Posteriors with Hyperparameter Marginalization, Technical Report T2100301, LIGO. https://dcc.ligo.org/T2100301

Gelman, A., Carlin, J. B., Stern, H. S., et al. 2013, Bayesian Data Analysis, Chapman & Hall

Hogg, D. W., Myers, A. D., & Bovy, J. 2010, ApJ, 725, 2166, doi: 10.1088/0004-637X/725/2/2166

Lieu, M., Farr, W. M., Betancourt, M., et al. 2017, MNRAS, 468, 4872, doi: 10.1093/mnras/stx686

Loredo, T. J. 2004, in American Institute of Physics Conference Series, Vol. 735, American Institute of Physics Conference Series, ed. R. Fischer, R. Preuss, & U. V. Toussaint, 195–206, doi: 10.1063/1.1835214

Mandel, I., Farr, W. M., & Gair, J. R. 2019, MNRAS, 486, 1086, doi: 10.1093/mnras/stz896

Messenger, C., & Veitch, J. 2013, New Journal of Physics, 15, 053027, doi: 10.1088/1367-2630/15/5/053027