DRAFT VERSION APRIL 24, 2019
Typeset using LATEX **modern** style in AASTeX62

Accuracy Requirements for Empirically-Measured Selection Functions

WILL M. FARR[1,2]

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

[2]*Center for Computational Astronomy, Flatiron Institute, New York NY 10010, United States*

When conducting a population analysis on a catalog of objects the effect of the selection function must be incorporated to avoid so-called "Malmquist bias" (Malmquist 1922; Loredo 2004; Mandel et al. 2018). Suppose we have a catalog consisting of data d_i, $i = 1, \ldots, N_{\mathrm{obs}}$, that constrain the parameters θ_i of a set of N_{obs} objects. We wish infer the population distribution function

$$\frac{\mathrm{d}N}{\mathrm{d}\theta}\left(\lambda\right), \tag{1}$$

which can depend on some population-level parameters λ. The joint posterior for the object-level parameters θ_i and population-level parameters is (Loredo 2004; Mandel et al. 2018)

$$\pi \propto \prod_{i=1}^{N_{\mathrm{obs}}} \left[p\left(d_i \mid \theta_i\right) \frac{\mathrm{d}N}{\mathrm{d}\theta_i}\left(\lambda\right) \right] \exp\left[-\Lambda\left(\lambda\right)\right] p\left(\lambda\right). \tag{2}$$

$p\left(d \mid \theta\right)$ is the likelihood function that describes the measurement process for the catalog, $p\left(\lambda\right)$ is a prior, and Λ is the expected number of detections:

$$\Lambda\left(\lambda\right) \equiv \int_{\{d \mid f(d) > 0\}} \mathrm{d}d\, \mathrm{d}\theta\, \frac{\mathrm{d}N}{\mathrm{d}\theta}\left(\lambda\right) p\left(d \mid \theta\right). \tag{3}$$

f represents the selection function; an observation will be included in the catalog if and only if it generates data such that $f(d) > 0$. We factor an overall normalization out of the population distribution so that

$$\frac{\mathrm{d}N}{\mathrm{d}\theta}\left(\lambda\right) = R\xi\left(\theta \mid \tilde{\lambda}\right), \tag{4}$$

with the amplitude of ξ fixed in some way; $\tilde{\lambda}$ is the set of parameters that remain once the amplitude of the population distribution is fixed. In this re-parameterization, $\Lambda = Rx$, where x is given by

$$x\left(\tilde{\lambda}\right) \equiv \int_{\{d \mid f(d) > 0\}} \mathrm{d}d\, \mathrm{d}\theta\, \xi\left(\theta \mid \tilde{\lambda}\right) p\left(d \mid \theta\right). \tag{5}$$

will.farr@stonybrook.edu
wfarr-vscholar@flatironinstitute.org

If ξ integrates to one over all θ, then x is the *fraction* of sources from a population described by $\tilde{\lambda}$ that are detectable.

In simple cases the integral in Eq. (5) can be evaluated analytically. But for most realistic applications it is not possible to analytically evaluate f (see e.g. Burke et al. 2015; Christiansen et al. 2015; Abbott et al. 2016b,a; Burke & Catanzarite 2017). Instead, the detection efficiency must be estimated by drawing synthetic objects from a fiducial distribution, $p_{\mathrm{draw}}(\theta)$, drawing corresponding data from the likelihood function $p(d \mid \theta)$, and "injecting" these data into the pipeline used to produce the catalog, recording which observations are detected (Tiwari 2018). This procedure introduces uncertainty in the estimation of the selection integral; we must have enough draws that this uncertainty does not alter the shape of the posterior π very much.

Given a set of detected objects with parameters θ_j, $j = 1, \ldots, N_{\mathrm{det}}$ generated from a total number of draws N_{draw} the integral in Eq. (5) can be estimated via

$$x \simeq \frac{1}{N_{\mathrm{draw}}} \sum_{j=1}^{N_{\mathrm{det}}} \frac{\xi\left(\theta_j \mid \tilde{\lambda}\right)}{p_{\mathrm{draw}}(\theta_j)}. \tag{6}$$

Under repeated samplings x will follow an approximately normal distribution

$$x \sim N(\mu, \sigma), \tag{7}$$

with

$$\mu \simeq \frac{1}{N_{\mathrm{draw}}} \sum_{j=1}^{N_{\mathrm{det}}} \frac{\xi\left(\theta_j \mid \tilde{\lambda}\right)}{p_{\mathrm{draw}}(\theta_j)}, \tag{8}$$

and

$$\sigma^2 \equiv \frac{\mu^2}{N_{\mathrm{eff}}} \simeq \frac{1}{N_{\mathrm{draw}}^2} \sum_{i=1}^{N_{\mathrm{det}}} \left[\frac{\xi\left(\theta_j \mid \tilde{\lambda}\right)}{p_{\mathrm{draw}}(\theta_j)}\right]^2 - \frac{\mu^2}{N_{\mathrm{draw}}}. \tag{9}$$

We have introduced the parameter N_{eff} that gives the *effective* number of independent draws that contribute to the estimate of x.

Given a particular sampling of the selection function, we should marginalize over the uncertainty in x. Eq. (2) becomes

$$\pi \propto \prod_{i=1}^{N_{\mathrm{obs}}} \left[p(d_i \mid \theta_i)\, \xi\left(\theta_i \mid \tilde{\lambda}\right)\right] \int \mathrm{d}x\, R^{N_{\mathrm{obs}}} \exp\left[-Rx\right] N(x \mid \mu, \sigma). \tag{10}$$

Integrating over $-\infty < x < \infty$, we obtain

$$\pi \propto \prod_{i=1}^{N_{\mathrm{obs}}} \left[p(d_i \mid \theta_i)\, \xi\left(\theta_i \mid \tilde{\lambda}\right)\right] R^{N_{\mathrm{obs}}} \exp\left[\frac{R\mu\left(R\mu - 2N_{\mathrm{eff}}\right)}{2N_{\mathrm{eff}}}\right]. \tag{11}$$

The divergence of this expression as $R \to \infty$ reflects that the normal approximation permits non-zero probability of $x < 0$. Eq. (11) has stationary points in R at

$$R = R_{\pm} = \frac{N_{\mathrm{eff}} \pm \sqrt{N_{\mathrm{eff}}\left(N_{\mathrm{eff}} - 4N_{\mathrm{obs}}\right)}}{2\mu}. \tag{12}$$

Provided $N_{\rm eff} > 4N_{\rm obs}$ these stationary points will occur for real, positive R. In this case, the stationary point at R_- is a local maximum; at R_+ we have a minimum associated with the "unphysical" transition to the divergent behavior as $R \to \infty$. We have

$$R_- = \frac{N_{\rm obs}}{\mu}\left(1 + \frac{N_{\rm obs}}{N_{\rm eff}} + 2\left(\frac{N_{\rm obs}}{N_{\rm eff}}\right)^2 + \mathcal{O}\left(\frac{N_{\rm obs}}{N_{\rm eff}}\right)^3\right). \tag{13}$$

$R = N_{\rm obs}/\mu$ is the point estimate for the detection efficiency in Eq. (6). Near $R = R_-$ a normal approximation holds for the posterior as a function of R with $\mu_R = R_-$ and

$$\sigma_R = \frac{\sqrt{N_{\rm obs}}}{\mu}\left(1 + \frac{3}{2}\frac{N_{\rm obs}}{N_{\rm eff}} + \frac{31}{8}\left(\frac{N_{\rm obs}}{N_{\rm eff}}\right)^2 + \mathcal{O}\left(\frac{N_{\rm obs}}{N_{\rm eff}}\right)^3\right). \tag{14}$$

Marginalizing the normal approximation over R imposing a flat-in-log R prior gives

$$\log \pi \propto \sum_{i=1}^{N_{\rm obs}} \log p\left(d_i \mid \theta_i\right) \xi\left(\theta_i \mid \tilde{\lambda}\right) - N_{\rm obs} \log \mu + \frac{3N_{\rm obs} + N_{\rm obs}^2}{2N_{\rm eff}} + \mathcal{O}\left(N_{\rm eff}\right)^{-2}. \tag{15}$$

The term involving μ would appear in an analysis that ignores the rate R and works entirely with population distributions (Mandel et al. 2018; Fishbach et al. 2018); the term involving $N_{\rm eff}$ is a correction to account for the uncertainty in our estimate of the selection integral.

The uncertainty in parameters is driven by the *differences* in the log-posterior. The R-dependent terms contribute to such differences through

$$\Delta \log \pi = \ldots - N_{\rm obs}\left(\frac{\partial \log \mu}{\partial \tilde{\lambda}} - \frac{N_{\rm obs}}{2N_{\rm eff}}\frac{\partial \log N_{\rm eff}}{\partial \tilde{\lambda}}\right)\Delta\tilde{\lambda}. \tag{16}$$

Both derivatives are independent of $N_{\rm eff}$, so the relative contribution of the second term to the parameter estimates is $\mathcal{O}\left(N_{\rm obs}/N_{\rm eff}\right)$.

If $N_{\rm eff}$ becomes close to $4N_{\rm obs}$ for any relevant set of population parameters then the posterior no longer peaks in R and more injections must be obtained for an accurate analysis.

A worked example, along with the LaTeX source for this document, can be found at https://github.com/farr/SelectionAccuracy.

REFERENCES

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2016a, The Astrophysical Journal Supplement Series, 227, 14

—. 2016b, ApJ, 833, L1

Burke, C. J., & Catanzarite, J. 2017, Planet Detection Metrics: Per-Target Detection Contours for Data Release 25, Technical Report KSCI-19111-002, NASA Ames Research Center, https://exoplanetarchive.ipac.caltech.edu/docs/KSCI-19111-002.pdf

Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, ApJ, 809, 8

Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2015, ApJ, 810, 95

Fishbach, M., Holz, D. E., & Farr, W. M. 2018, ApJ, 863, L41

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

Malmquist, K. G. 1922, Meddelanden fran Lunds Astronomiska Observatorium Serie I, 100, 1

Mandel, I., Farr, W. M., & Gair, J. R. 2018, ArXiv e-prints, arXiv:1809.02063

Tiwari, V. 2018, Classical and Quantum Gravity, 35, 145009