MNRAS **000**, 1–4 (2017) Preprint 20 May 2017 Compiled using MNRAS LaTeX style file v3.0

Correcting for data truncation in Bayesian analyses

Samuel R. Hinton,[1,2]* Alex Kim,[3] Tamara M. Davis,[1,2]

[1] *School of Mathematics and Physics, The University of Queensland, Brisbane, QLD 4072, Australia*

[2] *ARC Centre of Excellence for All-sky Astrophysics (CAASTRO)*

[3] *Lawrence Berkeley National Labs*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Lifetimes are strictly positive. Fainter objects are harder to see with a telescope than brighter ones. We cut our data to remove background noise. Truncated or biased datasets are inescapable in modern physics. In this paper, we present a simple overview of a Bayesian consideration of truncation, giving a solution to both analytically tractable and intractable models. This can be accomplished via a combination of analytic approximations and Monte Carlo integration, in which dataset simulation is efficiently used to correct for issues in the observed dataset. Toy models are included, along with numerical considerations and optimisations for implementation.

1 INTRODUCTION

Truncated data is a problem in many areas of scientific inquiry. It is one of the primary difficulties when performing supernovae cosmology analysis, as our telescopes have visual limits that truncate our observed supernovae distribution from the actual underlying distribution. This bias, termed Malmquist bias, is source of much investigation (Butkevich et al. 2005). It is considered during analysis by either modifying the observed data to remove the expected bias (Betoule et al. 2014; Conley et al. 2011), or by incorporating the expected bias into the underlying model (Rubin et al. 2015). Truncated data is also commonly encountered in biological fields, where data such as mortality rates are left-truncated (Colchero & Clark 2012). Simplified and generalised examples have been investigated in numerous fashions (Woodroofe 1985; Gull 1989; Grogger & Carson 1991; O'Neill & Barry 1995) and with different fitting algorithms (Gelfand et al. 1992). Whilst generalised resources exist that provide a comprehensive overview of truncated data and analysis techniques (Klein & Moeschberger 2005), these sources are often opaque due to volume and mathematical complexity.

This work provides a simple treatment of truncated data in a common Bayesian technique. Section 2 discusses the ever-elusive case of perfect data without truncation to provide a common basis for Sections 3 and 4, which respectively cover analytically correctable data truncation and analytically intractable models. Section 5 details numeric concerns and tricks to be aware of for effective implementation of Monte Carlo corrections applied to analytic approximations.

2 THE PERFECT WORLD: NO TRUNCATION

In a perfect world, data is neither biased nor truncated. The data is perfect. Uncertainties are well quantified and normally distributed around true values. Presumably everything is also spherical and in a vacuum. Let us create a mock model in this perfect world. Let us observe a series of independent and identically distributed events $\vec{x}$, which are drawn from a normal distribution such that

$$\vec{x} \sim \mathcal{N}(\mu, \sigma). \tag{1}$$

If, having collected our observations $\vec{x}$, we wanted to constrain μ and σ, this would be a simple task of modelling the posterior surface. Taking uniform priors on both parameters we simply wish to map the surface

$$P(\theta|\text{data}) \propto P(\text{data}|\theta)P(\theta), \tag{2}$$

where our model parameters $\theta = \{\mu, \sigma\}$ and our data is given by $\vec{x}$.

$$P(\mu, \sigma|\vec{x}) \propto P(\vec{x}|\mu, \sigma)P(\mu, \sigma) \tag{3}$$

With uniform priors, $P(\mu, \sigma) =$ constant, and can be absorbed into the constant of proportionality. Expanding our observation vector, the posterior surface is given by

$$P(\mu, \sigma|\vec{x}) \propto \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \sigma). \tag{4}$$

Generating a hundred data points with $\mu = 100$, $\sigma = 10$, we can recover our input parameters easily, as shown in Figure 1.

3 THE IMPERFECT WORLD: ANALYTIC TRUNCATION

In a slightly imperfect world we may have to deal with something like truncated data. For an example, consider the previous model, but with an instrumentation deficiency such that we can only observe events above a certain threshold, such that $x > \alpha$. We assign a value $\alpha = 85$ for convenience. If we do not take this truncation into account, we will recover biased parameter estimates, as shown in Figure 2. However,

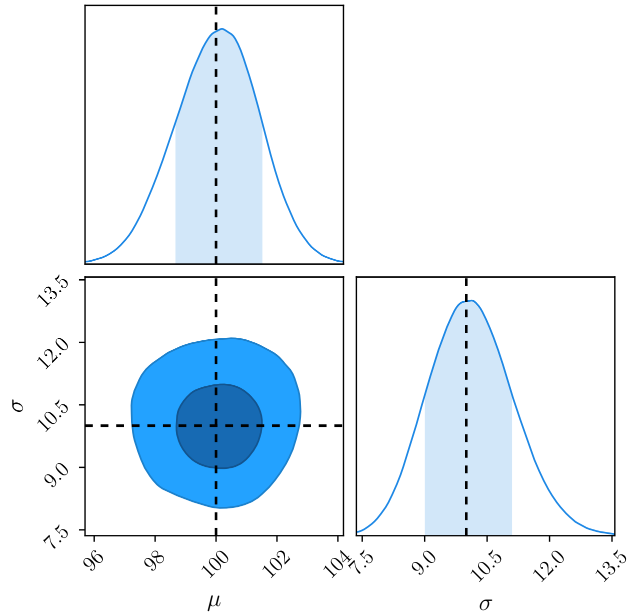

Figure 1. A systematic test of our perfect model, done by stacking the output chains from fitting 100 independent realisations of our 100 data points. Any systematic offset in our model would be revealed by a shift in the stacked results away from the true parameter values.

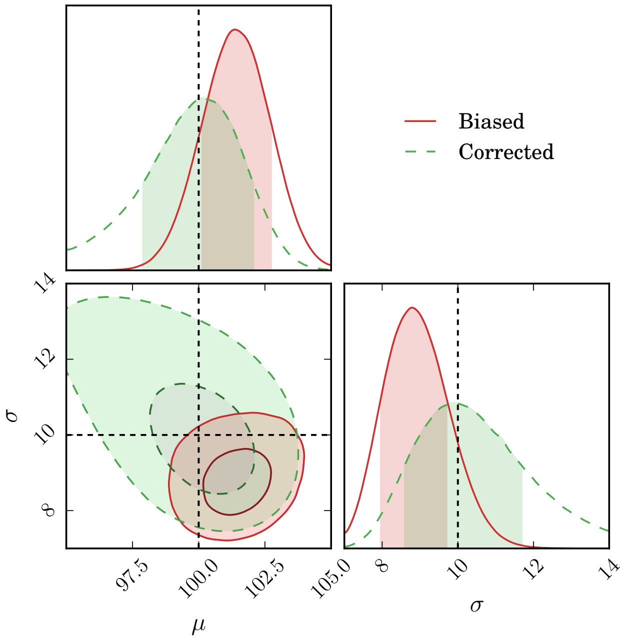

Figure 2. A systematic test of our imperfect model, done by stacking the output chains from fitting 100 independent realisations of our 100 data points, subject to our thresholding. The bias shown in the red 'Biased' contour can be corrected to via the techniques shown in Section 3 to recover unbiased surfaces.

we can correct for this truncation. If we restate our likelihood as the probability of observation given our model parameters *and* our selection effects S, we have

$$\mathcal{L} = P(\text{data}|\theta, S) \tag{5}$$

$$= \frac{P(S|\text{data}, \theta) P(\text{data}|\theta)}{P(S|\theta)}. \tag{6}$$

Introducing an integral over all possible data to make $P(S|\theta)$ physical,

$$= \frac{P(S|\text{data}, \theta) P(\text{data}|\theta)}{\int P(S, D|\theta)\, dD} \tag{7}$$

$$= \frac{P(S|\text{data}, \theta) P(\text{data}|\theta)}{\int P(S|D, \theta) P(D|\theta)\, dD}, \tag{8}$$

where the integral in the denominator has the same dimensionality is the experimental data. **Equation 8 is the generalised likelihood of experiments with sample selection.** In our example, our data is one dimensional (as we only observe one variable, x) and the selection efficiency is the step function $P(S|x, \theta) = \mathcal{H}(x - \alpha)$. Having successfully observed x, it follows that $x > \alpha$ and so $P(S|x, \theta) = 1$. To substitute in our normal model,

$$\mathcal{L}_i = \frac{\mathcal{N}(x_i|\mu, \sigma)}{\int_{-\infty}^{\infty} \mathcal{H}(D - \alpha) \mathcal{N}(D|\mu, \sigma)\, dD} \tag{9}$$

$$= \frac{\mathcal{N}(x_i|\mu, \sigma)}{\int_{\alpha}^{\infty} \mathcal{N}(D|\mu, \sigma)\, dD} \tag{10}$$

$$= \frac{\mathcal{N}(x_i|\mu, \sigma)}{\frac{1}{2}\text{erfc}\left[\frac{\alpha-\mu}{\sqrt{2}\sigma}\right]}, \tag{11}$$

where in the last line we have evaluated the integral in the case $\mu > \alpha$. Note that this is for a single observation of x, and so for a set of independent observations we need to introduce the product found in equation (4).

$$\mathcal{L} = \prod_{i=1}^{N} \frac{\mathcal{N}(x|\mu, \sigma)}{\frac{1}{2}\text{erfc}\left[\frac{\alpha-\mu}{\sqrt{2}\sigma}\right]}, \tag{12}$$

However, as our selection efficiency correction is observationally independent it is identical for all observations, allowing us to take it outside the product.

$$\mathcal{L} = 2\left(\text{erfc}\left[\frac{\alpha - \mu}{\sqrt{2}\sigma}\right]\right)^{-N} \prod_{i=1}^{N} \mathcal{N}(x|\mu, \sigma), \tag{13}$$

We can add this correction to our model, and note that we now recover unbiased parameter estimates, also shown in Figure 2.

4 THE REAL WORLD: ANALYTICALLY INTRACTABLE TRUNCATION

Unfortunately it is a rare scenario when dealing with nature and all her faults for us to have an analytic selection function, let alone a function encapsulated by a single parameter. A more realistic scenario involves a selection efficiency instead would take the form of non-analytic function of many model parameters. And the function would probably be stochastic too, just to throw another wrench in the works. Provided a method of forward modelling or simulating observations, the solution is to combine an analytic approximate correction with Monte Carlo integration.

So let us modify our imperfect toy model. Instead of observing just one variable, x, we also observe a new independent variable, y, which is drawn from its own distribution

$y \sim \mathcal{N}(\mu_y, \sigma_y)$. Our selection efficiency can now become a combination of x and y, such that we only observe events that satisfy $x + \beta y > \alpha$, giving $P(S|x, y, \theta) = \mathcal{H}(x + \beta y - \alpha)$. Our likelihood for such a toy model becomes now the combination of probabilities for observing both x and y, with the denominator becoming an integral over all possible X and Y observations subject to our selection effects.

$$\mathcal{L}_i = \frac{\mathcal{N}(x_i|\mu,\sigma)\mathcal{N}(y_i|\mu_y,\sigma_y)}{\iint_{-\infty}^{\infty} \mathcal{H}(X + \beta Y - \alpha)\mathcal{N}(X|\mu,\sigma)\mathcal{N}(Y|\mu_y,\sigma_y)\,dX\,dY} \tag{14}$$

Assume that we cannot solve this integral analytically, and must resort to numeric solutions. These often clash with sampling methods, especially for high dimensional integrals. Inserting Monte Carlo integration into fitting algorithms can drastically slow them down, and algorithms such as Hamiltonian MCMC that require continuous surfaces can easily fail on surfaces that fluctuate from Monte Carlo integration. Even by fixing the samples used in MC integration (thereby giving a continuous surface), the complexity of the surface derivatives will pose almost insurmountable problems for any algorithms that utilise surface gradients. One solution is to find an approximate, analytic correction we can utilise in our fitting algorithm which seeks to shift the region of parameter space sampled by the sampler closer to the correct area.

In our example, if $\beta \ll 1$, such that the majority of selection effect is encapsulated by x and not y, our approximate correction can take the form found in the previous correction from Section 3. Having true values of $\mu = 100$, $\sigma = 10$, $\mu_y = 30$, $\sigma_y = 5$, and a known $\beta = 0.2$, we can give a concrete example. Assuming some prior, imperfect knowledge of μ_y (perhaps we believe it is approximately 20) we estimate that the average contribution from βy is around $20\beta = 4$ (which is close to the correct value of 6), and from this our analytic correction to our likelihood is

$$w_{\rm approx} = \frac{1}{2}{\rm erfc}\left[\frac{\alpha - \mu - 4}{\sqrt{2}\sigma}\right]. \tag{15}$$

Further, let us explicitly break our likelihood into two parts, $\mathcal{L}_i = \mathcal{L}_{i1}\mathcal{L}_{i2}$, with the parts given by

$$\mathcal{L}_{i1} = \frac{\mathcal{N}(x_i|\mu,\sigma)\mathcal{N}(y_i|\mu_y,\sigma_y)}{w_{\rm approx}} \tag{16}$$

$$\mathcal{L}_{i2} = \frac{w_{\rm approx}}{\iint_{-\infty}^{\infty} \mathcal{H}(x + \beta y - \alpha)\mathcal{N}(X|\mu,\sigma)\mathcal{N}(Y|\mu_y,\sigma_y)\,dX\,dY}. \tag{17}$$

$\mathcal{L}_1$ can thus be fitted with a traditional sampler without numeric difficulty or slowdown, and $\mathcal{L}_2$ allows us to calculate the weight of each sample. We are effectively importance sampling our likelihood evaluations. The computational benefits of this should not be understated either - each sample in our chains can be reweighted independently, providing a task that is trivially parallelisable. Evaluating $\mathcal{L}_2$ using Monte Carlo integration of n samples, we have

$$\mathcal{L}_{i2} = \frac{w_{\rm approx} n}{\sum_{j=1}^{n} \mathcal{H}(X_j + \beta Y_j - \alpha)\mathcal{N}(X_j|\mu,\sigma)\mathcal{N}(Y_j|\mu_y,\sigma_y)}. \tag{18}$$

We can now easily move from a single observation to a set of N observations.

$$\mathcal{L}_1 = w_{\rm approx}^{-N} \prod_{i=1}^{N} \mathcal{N}(x_i|\mu,\sigma)\mathcal{N}(y_i|\mu_y,\sigma_y) \tag{19}$$

$$\mathcal{L}_2 = \left(\frac{w_{\rm approx} n}{\sum_{j=1}^{n} \mathcal{H}(X_j + \beta Y_j - \alpha)\mathcal{N}(X_j|\mu,\sigma)\mathcal{N}(Y_j|\mu_y,\sigma_y)}\right)^N \tag{20}$$

Thus we end up with a corrected posterior surface as shown in Figure 3.

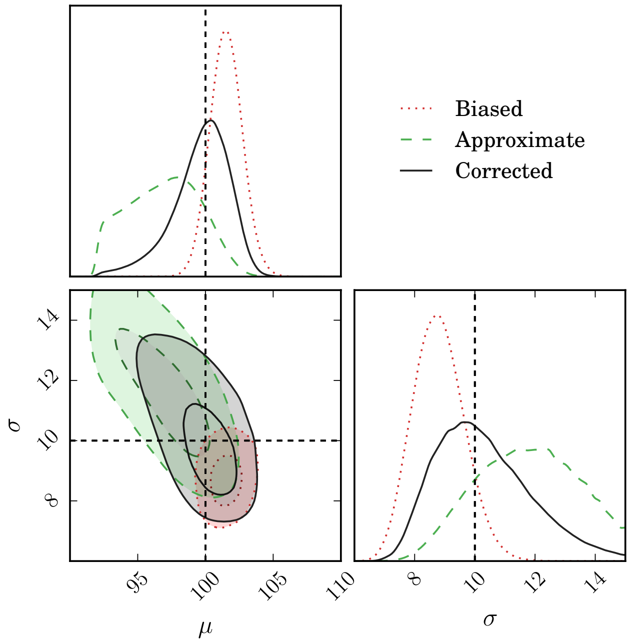

Figure 3. A systematic test of our more complicated model, done by stacking the output chains from fitting 100 independent realisations of our 100 data points, subject to our thresholding. The likelihood $\mathcal{L}_1$ was evaluated with the fitting algorithm emcee, and reweighted using Monte Carlo integration of a hundred thousand possible events as per $\mathcal{L}_2$. The truncated data with no correction is shown as 'Biased' in dotted red, the 'Approximate' only correction ($\mathcal{L}_1$) shown in dashed green, and the final reweighted chain shown in solid black as 'Corrected'.

5 NUMERICAL TRICKS

5.1 Importance Sampling

Further tricks can be used to increase the efficiency with which the samples are reweighted. Firstly, the overarching analytic model often provides functions which can be drawn from efficiently. In the case of our example, by drawing the random numbers X and Y respectively from the normal distributions $\mathcal{N}(\mu, \sigma)$ and $\mathcal{N}(\mu_y, \sigma_y)$ (ie traditional importance sampling) we need only evaluate the step function for our data points.

5.2 Precomputing selection

If evaluating the probability that an event is observed is numerically expensive (i.e. not a step function), it is easy to pregenerate a set of events and reuse them for all weights - provided that the number of events used when calculating the weights is sufficient to make the statistical error of Monte Carlo integration insignificant when compared to the constraining power of your dataset. This method is however only efficient when prior knowledge of parameter values is known to allow a reasonable initial draw of events. Without this prior information, samples need to span the entire posterior volume, which is numerically intractable even for low dimensional models.

Consider the imperfect example - where we observe x drawn from an underlying normal distribution, but utilise the Monte Carlo integration technique from Section 4 and do not have an analytic approximation (i.e. we set $w_{\rm approx} = 1$). We could estimate, given some prior knowledge, that variable $x \approx \mathcal{N}(\mu_{\rm guess}, \sigma_{\rm guess})$. We then drawn samples of x from this distribution, recording the probability of each draw and then calculating whether our potential observation of x would be observed in the experiment or not. That is, we assign $P(S|x,\theta) = P(S|x) = 1$ or 0 given it passed cuts or not. We discard all events with 0 weight (as they have 0 weight), and only track those events which pass. Then, when calculating the sample reweighting after running chains, $\mathcal{L}_2$ becomes

$$\mathcal{L}_2 \propto w_{\rm approx} \left[\sum_{i=1}^{n} \frac{\mathcal{N}(X_i|\mu,\sigma)}{\mathcal{N}(X_i|\mu_{\rm guess},\sigma_{\rm guess})} \right]^{-1}, \tag{21}$$

where you can see that we discard the constant n from equation (18) as we only care about likelihood proportionality. Provided our parameter estimate is reasonably well informed, the computation benefit this precomputation provides is enormous for any nontrivial selection function. Not only do we now waste no time when calculating $\mathcal{L}_2$ determining $P(S|{\rm data},\theta)$, and because we only save results that pass the cuts, we have no wasted evaluations of $\mathcal{N}(X|\mu,\sigma)$.

The astute reader may have picked up on one assumption - that selection efficiency of an observation is independent of model parameters θ. For most experimental cases this should hold, however if it does not this method cannot be used to increase efficiency. Gridding or interpolating the parameter space is strongly not recommended due the required accuracy of $\mathcal{L}_2$. Even a small error when raised to the power of N can spiral out of control.

5.3 Log-space

Following from the previous section, as our reweighting $\mathcal{L}_2$ is raised to the power of the number of our observations, they should definitely be computed in log-space, which turns the power into a linear factor. As most probabilistic work is already computed in log-space, this subsection barely needs to be stated. However, whilst working in log-space an efficient way of increasing the accuracy of the approximate analytic correction is to fit the correction such that the spread of the distribution $\log \mathcal{L}_2$ is minimised.

6 CONCLUSION

I keep trying to write this but just end up restarting the abstract almost verbatim. Do you have some good tips (for this write up and for the future) on proper ways to write the conclusion? How it should differ from the abstract and how to avoid repeating yourself too much?

ACKNOWLEDGMENTS

We gratefully acknowledge the input of the many researchers that were consulted during the creation of this paper.

REFERENCES

Betoule M., et al., 2014, A&A, 568, A22
Butkevich A. G., Berdyugin A. V., Teerikorpi P., 2005, Monthly Notices of the Royal Astronomical Society, 362, 321
Colchero F., Clark J. S., 2012, Journal of Animal Ecology, 81, 139
Conley A., et al., 2011, ApJS, 192, 1
Gelfand A. E., Smith A. F. M., Lee T.-M., 1992, Journal of the American Statistical Association, 87, 523
Grogger J. T., Carson R. T., 1991, Journal of applied econometrics, 6, 225
Gull S. F., 1989, in , Maximum Entropy and Bayesian Methods. Springer, pp 511–518
Klein J. P., Moeschberger M. L., 2005, Survival analysis: techniques for censored and truncated data. Springer Science & Business Media
O'Neill T. J., Barry S. C., 1995, Biometrics, pp 533–541
Rubin D., et al., 2015, ApJ, 813, 137
Woodroofe M., 1985, The Annals of Statistics, pp 163–177

This paper has been typeset from a TEX/LATEX file prepared by the author.