MNRAS **000**, 1–4 (2017) Preprint 11 May 2017 Compiled using MNRAS LATEX style file v3.0

Bayesian pipeline to address the problem of faint flux measurement

Krzysztof Suberlak,[1]⋆ Željko Ivezić [1]
[1]*Department of Astronomy, University of Washington, Seattle, WA, United States*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
A technical report to outline the methods used to reprocess forced photometry measurement below a certain small signal-to-noise ratio. By imposing a non-negative prior on faint and even negative fluxes we are able to recover upper measurement limits and thus provide a more physical interpretation of the underlying brightness of the measured source at faint epochs.

1 INTRODUCTION

Many objects in the universe, from stellar to extragalactic scales, vary on timescales less than a few hundred years. Light curves carry a wealth of information allowing one to infer various physical properties of a planet or a galaxy. If a light curve is poorly sampled, the inferred characteristics are less certain (see **?** for review). Yet, since all astronomical observations suffer from a detection threshold, a very faint variable object in some epochs may be undetectable. Forced photometry rescues the information from very faint epochs by performing a measurement in all epochs in a location from the co-added images. An inherent challenge to such set of measurements is an interpretation of noise-dominated flux. To circumvent this problem many studies apply a magnitude cutoff few magnitudes above the detection limit, which reduces the amount of available data. Indeed, in order to fully utilize information present in time-domain surveys, such as Large Scale Synoptic Telescope, Palomar Transient Factory, or Sloan Digital Sky Survey, and properly characterize faint variable objects, we need to properly handle the faint flux measurements. A new methodology would allow an unbiased study of such faint variable objects, including quasars, RR Lyrae, Cepheids, and a wealth of other variable sources. With the advent of precision time-domain astronomy surveys it is crucial to apply the best possible faint forced photometry algorithms and thus make full use of the data.

In this report we describe how faint measurements may arise, specifically as a result of performing forced photometry on background-subtracted data. We suggest a fully Bayesian approach to recalculating flux at each faint epoch, and provide example usage.

2 FAINT SOURCES

Forced photometry in a background-subtracted epoch may yield unphysical, negative values. Such negative pixel value may originate from the variation of background across the image. If we model the background counts as a Gaussian centered around the mean $B_0 : B - B_0 \sim \mathcal{N}(0, \sigma_B)$, then background in some parts of the image will be above, and in other parts below the mean value. After subtraction of the mean background value, regions with previously lower than average background will have negative pixel value. This means that a forced photometry on a location where an object is undetected in single epoch may be measuring the background noise. Apart from creating low pixel values, background oscillation may also cause spurious detections where it is above the mean. For a large number of photons hitting the detector the width of the background noise distribution is proportional to the square-root of counts: $\sigma_B \propto \sqrt{B}$. This yields a 1 in a million chance that a pixel has a background value larger than $5\sigma_B$. Thus on a 16 megapixel CCD, like those used in SDSS, we anticipate about 16 spuriously bright pixels per CCD. Therefore, since forced photometry is affected by the background noise, a special care must be taken of faint, noise-dominated measurements.

We remedy the unphysically low, even negative measurements for all 'faint' ($< 2\sigma$) sources and recalculate their fluxes. Each flux measurement can be thought of as a mean of the 'intrinsic' flux likelihood function $L(F)$. L determines the probability that the flux has a value F. In our treatment we assume that L is a Gaussian, and therefore its width corresponds to the measurement error : $L(F) \sim \mathcal{N}(F, \sigma_F)$. This means that bright sources with high signal-to-noise ratio have very narrow L, and faint sources, dominated by noise, have L with very wide tails. Only for faint sources a significant portion of L may be negative, corresponding to non-zero likelihood of flux being negative. This stands in conflict with our prior knowledge that no physical flux can be negative. We resolve this problem by recalculating single-epoch flux for all sources where $F < 2\sigma_F$. We calculate for each epoch the mean of the truncated L, such that $L(F) = 0$ for $F < 0$. This shifts upward all measurements for faint epochs, and remedies the unphysicality of faint forced photometry fluxes.

In our treatment we are explicitly using a prior understanding of the flux behavior of any astrophysical object.

© 2017 The Authors

Without any further knowledge about the nature of the source, flat prior is the least informative Bayesian prior. Any additional information about the nature of object, and thus expected variability pattern, could affect the choice of prior to be more specific. For instance, consider a sinusoidal flux variability. If the flux of an object over many epochs is expected to vary in a sinusoidal fashion, i.e. $F(t) = F_{min} + sin(t)$, the probability of a given flux measurement is a cosine, ranging from F_{min} to F_{max}. With that prior, without any measurement taken, the flux of the object is most likely $(F_{min} + F_{max})/2$, i.e. at the peak of the cosine likelihood function. However, as soon as one measures (F_i, e_i) from that source, the probability distribution of a flux at that epoch becomes a convolution of cosine prior information with the Gaussian curve of width e_i, centered on F_i (assuming Gaussian errors). However, without any a-priori information about the variability pattern of the considered object, the least informative Bayesian prior we can impose is a flat one : $p(F) = 0$ for $F < 0$, and 1 elsewhere. Thus the posterior probability is

$$p(F|data) \propto L(F|data)p(F) \tag{1}$$

To test the method we generate fiducial light curves (DRW / sinusoidal), with a uniform sampling ($N = 100 \div 1000$). Based on the generated flux (F_{true}) we define the 5σ level as the robust 25-th percentile (or median) of the ensemble F_{true} distribution : $\sigma_F = (1/5)F_{25\%}$ (in reality, σ increases for fainter observations, but this is a good approximation). We define $F_{obs} = F_{true} + F_{noise}$, where the Gaussian noise $F_{noise} = \sigma_F \mathcal{N}(0,1)$ was added to each point. For a weak signal, defined as $F^i_{obs} < 2\sigma F$, we consider $p(F)$ - a Gaussian likelihood associated with i-th measurement: $p_i(F) = \mathcal{N}(\mu = F^i_{obs}, \sigma = \sigma_F)$. Each measurement F_{obs} is a mean of this likelihood: $F_{obs} = \langle p_i(F) \rangle$. We call it $p(F)$ for short :

For each epoch, based on the raw forced photometry measurement, we calculate new descriptors of faint fluxes. We define a faint measurement by $F_i < 2\sigma F_i$, i.e. where the flux is less than twice the flux error. We assume that the flux is the mean of the Gaussian likelihood $p_i(F) = \mathcal{N}(\mu = F_i, \sigma = \sigma_{F_i})$:

$$p(F) = \frac{1}{\sqrt{2\pi\sigma_F^2}} \exp\left(-\frac{(F-\mu)^2}{2\sigma_F^2}\right) \tag{2}$$

so that $F_i = \langle p(F) \rangle$. For faint measurements we truncate the negative part of $p(F)$, and recalculate the mean, median, rms, and 2σ level. Thus the mean is

$$F_{mean} = \frac{\int_0^\infty F p(F) dF}{\int_0^\infty p(F) dF} \tag{3}$$

where we normalized the truncated Gaussian likelihood.

We define the median as

$$\int_0^{F_{median}} p(F) dF = \int_{F_{median}}^\infty p(F) dF \tag{4}$$

The rms level is

$$F^2_{rms} = \frac{\int_0^\infty (F - F_{mean})^2 p(F) dF}{\int_0^\infty p(F) dF} \tag{5}$$

Finally, since for a Gaussian distribution the area contained between $\mu \pm \sigma$ is 95.5% of the total area under the curve, for the truncated Gaussian we define the 2σ level as:

$$\int_{F_{2\sigma}}^\infty p(F) dF = 0.05 * \int_0^\infty p(F) dF \tag{6}$$

Both for median and for the 2σ level the normalization cancels out (see Sec. 3)

3 TREATMENT OF FAINT SOURCES

In our calculations we used the `scipy` implementation of the following often used integrals of Gaussian distributions :

- cumulative density function, that is an area under the Gaussian distribution from $-\infty$ to x_0 :

$$\mathrm{cdf}(x_0, \mu, \sigma) = \int_{-\infty}^{x_0} \mathcal{N}(\mu, \sigma) dx = \int_{-\infty}^{x_0} \frac{\exp(x-\mu)^2/2\sigma^2}{\sqrt{2\pi\sigma^2}} dx \tag{7}$$

- point percent function , that is an inverse of the cumulative density function of the unit Gaussian: if $A = \mathrm{cdf}(x_0, 1, 0)$, then $x_0 = \mathrm{ppf}(A)$

- survival function (also known as the complementary cumulative distribution function), that is an area under a Gaussian distribution from x_0 to ∞

$$\mathrm{sf}(x_0, \mu, \sigma) = \int_{x_0}^\infty \mathcal{N}(\mu, \sigma) dx = \int_{-\infty}^\infty \mathcal{N}(\mu, \sigma) dx - \int_{-\infty}^{x_0} \mathcal{N}(\mu, \sigma) dx = 1 - \mathrm{cdf}(x_0) \tag{8}$$

Note that we use these functions here to employ the `scipy.stats.norm` methods of `cdf`, `sf`, and other functions that are related to integrating a Normal distribution. Therefore we reduce all expressions by appropriate substitutions and translation, to arrive at the integration of a unit Gaussian :

$$\mathcal{N}(0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \tag{9}$$

Therefore we use a shorthand notation :

$$\mathrm{sf}(x_0) = \mathrm{sf}(x_0, 1, 0) = \int_{x_0}^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \tag{10}$$

and likewise for

$$\mathrm{cdf}(x_0) = \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \tag{11}$$

In our faint flux treatment we assume that each flux measurement has an associated Gaussian likelihood, and that the width and mean of the likelihood correspond to the measured flux and the measurement error respectively.

For a source where signal-to-noise < 2 (in our case, ratio of flux to error), we remove the negative portion of the likelihood, since there is no physical likelihood that a flux would be negative. Thus for mean, we integrate from 0 instead of $-\infty$:

$$F_{mean} = \frac{\int_0^\infty F p(F) dF}{\int_0^\infty p(F) dF} = I_M / I_N \tag{12}$$

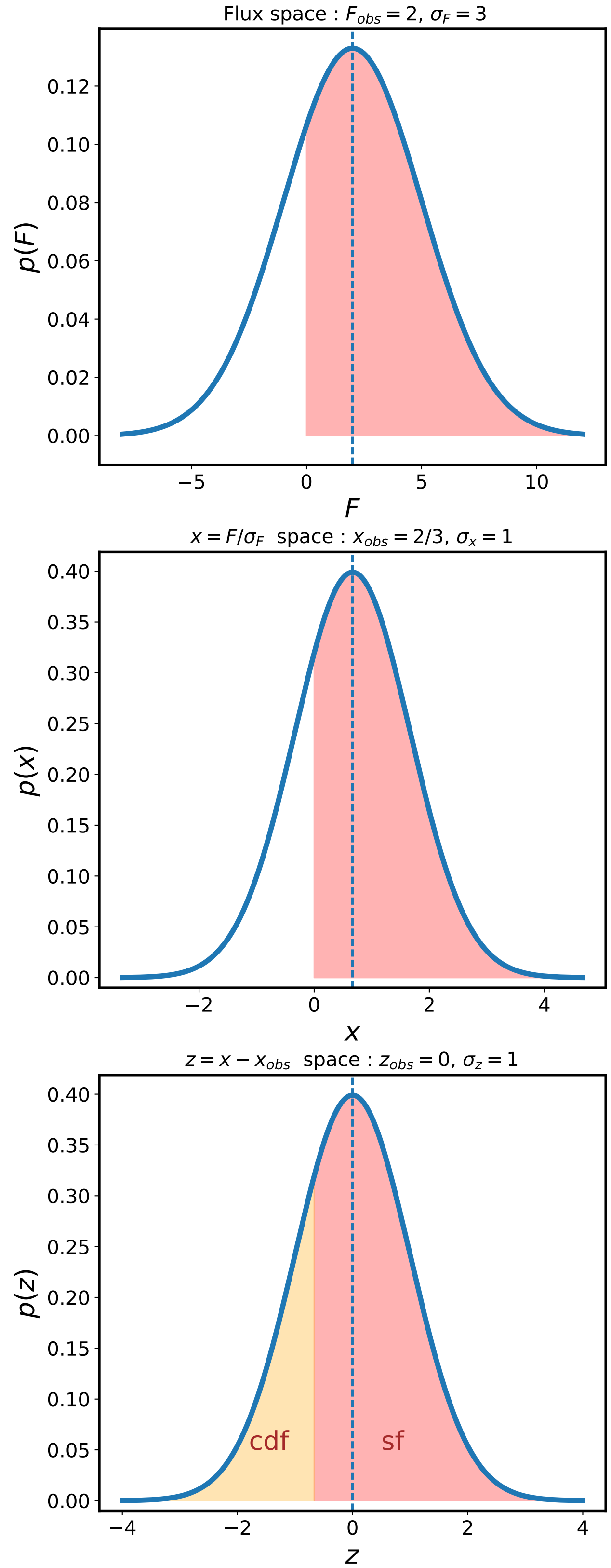

Figure 1. The three panels illustrate three main steps taken in expressing the analytic expressions for mean, median, RMS of the flux likelihood distribution as functions of a unit Gaussian. The top panels shows the flux space, where the likelihood for flux measurement is centered on the reported value of flux F_{obs}, with a width of the reported error σ_F. The middle panel shows the same expressed in the S/N space, called x-space here, since $x = S/N = F/\sigma_F$. Here the likelihood is a unit Gaussian centered on x_{obs}. The bottom panel depicts the z-space, which is the translation of x-space by x_{obs}, to allow calculations using unit Gaussian functions. In this space the likelihood is a unit Gaussian, centered on 0.

where we need to normalize by the integral over the positive part of the Gaussian likelihood.

We evaluate

$$I_M = \int_0^\infty \frac{F}{\sqrt{2\pi\sigma_F^2}} \exp\left(-\frac{(F-F_{obs})^2}{2\sigma_F^2}\right) dF = \frac{\sigma_F}{\sqrt{2\pi}} \exp\left(-\frac{F_{obs}^2}{2\sigma_F^2}\right) + F_{obs}\,\mathrm{sf}\left(\frac{-F_{obs}}{\sigma_F}\right) \quad (13)$$

and

$$I_N = \int_0^\infty p(F)dF = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_F^2}} \exp\left(-\frac{(F-F_{obs})^2}{2\sigma_F^2}\right) dF = \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-x_{obs})^2}{2}\right) dx \quad (14)$$

where we subsituted $F/\sigma_F = x$, and $F_{obs}/\sigma_F = x_{obs}$. Now, apply translation $z = x - x_{obs}$:

$$I_N = \int_{-x_{obs}}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = \mathrm{sf}(-x_{obs}) \quad (15)$$

so that

$$x_{mean} = \frac{\exp(-x_{obs}^2/2)}{\mathrm{sf}(-x_{obs})\sqrt{2\pi}} + x_{obs} \quad (16)$$

where we scaled F_{obs} by σ_F (i.e. $F_{mean} = x_{mean} \cdot \sigma_F$).

To find the median and the 2σ level we transform from F space to x space , scaling by σ_F, so that $x = F/\sigma_F$, and thus the likelihood $p(x) \sim \mathcal{N}(x_{obs}, 1)$ is :

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-x_{obs})^2}{2}\right) \quad (17)$$

We then transform from x to z space, with a translation by $x_{obs} = F_{obs}/\sigma_F$: $z = x - x_{obs}$, so that now $p(z) \sim \mathcal{N}(0,1)$:

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad (18)$$

In z-space, the median from

$$\int_0^{x_{med}} p(x)dx = \int_{x_{med}}^\infty p(x)dx \quad (19)$$

becomes

$$\int_{x_0}^{z_{med}} p(z)dz = \int_{z_{med}}^\infty p(z)dz \quad (20)$$

with $x_0 = -x_{obs}$

We find z_{med} analytically - the right hand side of Eq. 20 is the survival function :

$$RHS = \int_{z_{med}}^\infty p(z)dz = \mathrm{sf}(z_{med}) \quad (21)$$

and the left hand side, assuming that the median $z_{med} > x_0$, is :

$$LHS = \int_{x_0}^{z_{med}} p(z)dz = \int_{-\infty}^{z_{med}} p(z)dz - \int_{-\infty}^{x_0} p(z)dz = \mathrm{cdf}(z_{med}) - \mathrm{cdf}(x_0) \quad (22)$$

Comparing left and right hand sides :

$$\mathrm{cdf}(z_{med}) - \mathrm{cdf}(x_0) = \mathrm{sf}(z_{med})$$
$$\mathrm{cdf}(z_{med}) - \mathrm{cdf}(x_0) = 1 - \mathrm{cdf}(z_{med})$$
$$2\,\mathrm{cdf}(z_{med}) = 1 + \mathrm{cdf}(x_0) \quad (23)$$

Rearranging, and using the percent point function (ppf) we find:

$$\therefore z_{med} = \mathrm{cdf}^{-1}\left(\frac{1+\mathrm{cdf}(x_0)}{2}\right) = \mathrm{ppf}\left(\frac{1+\mathrm{cdf}(x_0)}{2}\right) \quad (24)$$

and transforming back to F space:

$$F_{med} = F_{obs} + \sigma_F\ \mathrm{ppf}\left(\frac{1+\mathrm{cdf}(-F_{obs}/\sigma_F)}{2}\right) \quad (25)$$

For the calculation of 2σ level , as in Eq. 6, we look for a point such that the area contained under the likelihood between that point and infinity (B) is 0.05 of the area as calculated under the curve from 0 to infinity (A) . In other words, B = 0.05A . Since we use unit Gaussian likelihood, both areas can be expressed as the survival functions: A = $\mathrm{sf}(x_0)$, and B = $\mathrm{sf}(z_B)$, i.e.

$$\mathrm{sf}(z_B) = 0.05\,\mathrm{sf}(x_0) \quad (26)$$

so to find z_B we use the inverse survival function isf : $z_B = \mathrm{isf}(0.05\mathrm{A})$. Transforming back to F-space we have:

$$F_{2\sigma} = F_{obs} + \sigma_F\,(\,\mathrm{isf}(\,0.05\,\mathrm{sf}(x_0)) \quad (27)$$

Finally, we find the root-mean-square:

$$F_{rms}^2 = \frac{\int_0^\infty (F - F_{mean})^2 p(F)dF}{\int_0^\infty p(F)dF} = I_R/I_N \quad (28)$$

this can be evaluated by numerical integration, scaling by σ_F, so that $x_{mean} = F_{mean}/\sigma_F$, $x_{obs} = F_{obs}/\sigma_F$:

$$F_{rms}^2 = \frac{\sigma_F^2 \int_0^\infty (x - x_{mean})^2 exp(-(x-x_{obs})^2/2)dx}{\int_0^\infty exp(-(x-x_{obs})^2/2)dx} \quad (29)$$

based on Eq. 28 we derive analytical expression for rms :

$$x_{rms} = \frac{F_{rms}}{\sigma_F} = \left(\frac{I_R}{I_N \sigma_F^2}\right)^{1/2} \quad (30)$$

where I_N is our normalization , as in calculation of F_{mean}, and as

$$\frac{I_R}{\sigma_F^2} = \frac{1}{2}\,\mathrm{erf}\left(\frac{x_{obs}}{\sqrt{2}}\right) + \frac{1}{\sqrt{2\pi}} e^{(-x_{obs}^2/2)}(2\Delta x - x_{obs}) + (\Delta x)^2\,\mathrm{sf}(-x_{obs}) \quad (31)$$

with $\Delta x = x_{obs} - x_{mean}$

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