Bayesian luminosity calibration
Exercise instructions

Anthony Brown

Sterrewacht Leiden

brown@strw.leidenuniv.nl

11.06.2013

Abstract. In this exercise you will reproduce the 'Bayesian luminosity calibration' example from chapter 16 in van Altena (2013).

Revision History

Issue	Rev. No.	Date	Author	Comments
0	2	11.06.2013	AB	More explanation on the goals of the exercise.
0	1	07.06.2013	AB	Comments by BH included.
0	0	22.05.2013	AB	Creation of notes.

Contents

References

van Altena, W.F., 2013, *Astrometry for Astrophysics*, Cambridge University Press

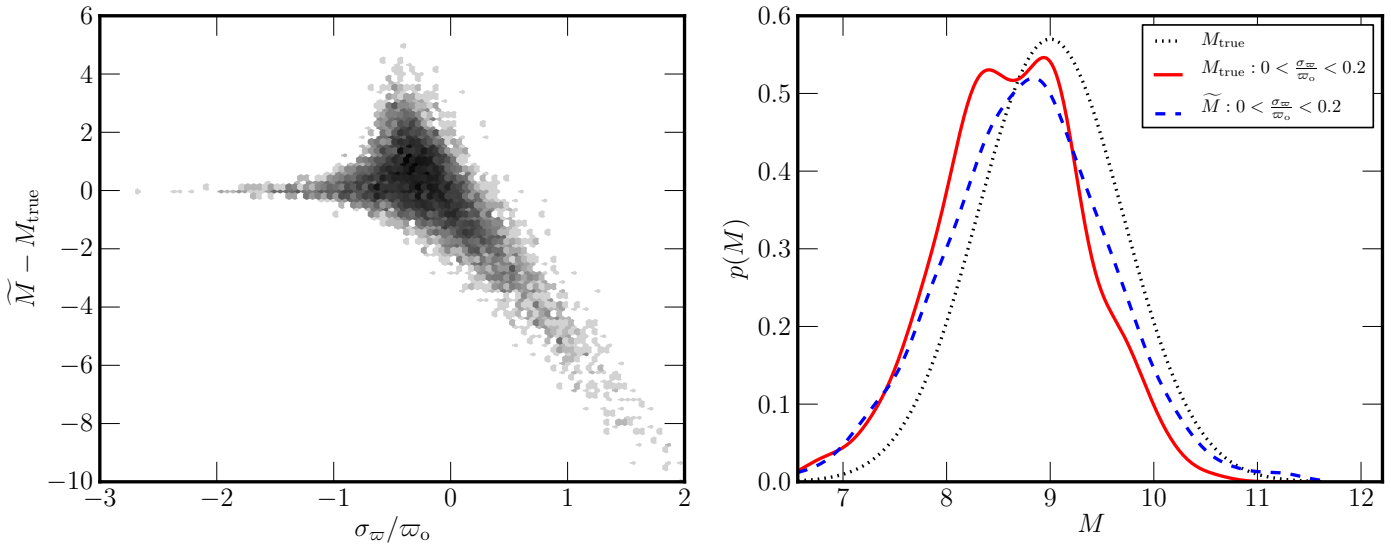

Figure 1: Results of the naive estimation of the absolute magnitudes of stars in our parallax survey. For clarity 10 000 stars were simulated which are uniformly distributed between 1 and 100 pc from the Sun. The true values of μ_M and σ^2_M are 9 and 0.49. The left panel shows how the simple estimates $\widetilde{M}_i$ for the individual absolute magnitudes can be severely wrong and that there is also a systematic bias as one goes to large relative parallax errors. The right panel shows that when selecting only the best parallaxes (relative errors less than 20%) a bias is introduced in the distribution of absolute magnitudes. The brightest stars are preferentially selected and this will skew the estimates of μ_M as well as σ^2_M.

1 Introduction

For this exercise we will interpret the data from a fictitious parallax survey of all stars of a particular luminosity class, which are distributed uniformly throughout a certain volume around the Sun. In addition we know that the luminosities of the stars have a normal distribution. The question we ask is: what is the mean absolute magnitude μ_M of this class of stars and what is the variance σ^2_M around the mean? The survey data provided are the measured parallaxes $\varpi_{\rm o}$, apparent magnitudes m and the corresponding observational errors σ_ϖ and σ_m. The latter vary with apparent magnitude, the brighter stars having smaller measurement errors. The straightforward approach to this problem would be to estimate for each star i its absolute magnitude as $\widetilde{M}_i = m_i + 5\log\varpi_{{\rm o},i} + 5$ and then determine the values of μ_M and σ^2_M from the resulting distribution of $\widetilde{M}_i$.

To illustrate that this naive approach is not a good idea we simulate a survey of 10 000 stars distributed uniformly between 1 and 100 pc from the Sun (i.e. 10 mas $< \varpi <$ 1000 mas). The values of values of μ_M and σ^2_M are 9 and 0.49, respectively. The errors on the observed parallaxes and apparent magnitudes are given by:

$$\sigma_{\varpi,i} = \begin{cases} a_\varpi \times 10^{0.2(m_i-m_0)} & m_i \geq m_0 \\ b_\varpi & m_i < m_0 \end{cases}, \qquad \sigma_{m,i} = \begin{cases} a_m \times 10^{0.2(m_i-m_0)} & m_i \geq m_0 \\ b_m & m_i < m_0 \end{cases}, \tag{1}$$

with σ_ϖ in mas and $a_\varpi = 0.2$, $b_\varpi = 0.2$ mas, $a_m = 0.006$, and $b_m = 0.001$. These parameters represent the slope of the increase of the errors with increasing m (photon noise driven) and a 'calibration floor' at the bright end. The calibration floor starts at $m_0 = 5$.

The results of using the naive estimates of M_i are shown in figure 1. The left panel shows the error in the magnitude estimate $\Delta M = \widetilde{M}_i - M_i$ as a function of the relative parallax error $\sigma_\varpi/\varpi_{\rm o}$. Depending on the value of the relative error the estimated magnitudes are on average over- or severely underestimated. The distribution of the points in ΔM vs. $\sigma_\varpi/\varpi_{\rm o}$ is caused by the combination non-linear relation between $\widetilde{M}_i$ and $\varpi_{\rm o}$ and the steep increase of the number of stars with decreasing parallax. This means that any average taken from these estimates may also be strongly biased. Now, one could try to remedy this problem by only considering stars with good quality parallaxes (here with $\sigma_\varpi/\varpi_{\rm o} < 0.2$). However this leads to preferentially selecting the intrinsically

brighter stars and thus to an unrepresentative distribution of absolute magnitudes as illustrated in the right panel of figure 1.

The problems associated with simply inverting parallaxes to estimate distances or luminosities have been discussed at length in the astronomical literature. More discussion can be found in chapter 16 of van Altena (2013) and references therein. The only real solution is to take a forward modelling approach and this is where the Bayesian methodology comes in.

2 Bayesian approach to luminosity calibration

We again use the simple parallax survey described above and now ask the slightly different question: what are the most likely values of μ_M and σ_M^2 given the observations $\mathbf{o} = \{\varpi_{o,i}, m_i\}$ $(i = 0, \ldots, N-1)$? In order to answer this question we will formulate it probabilistically using Bayes' theorem:

$$P(\mu_M, \sigma_M^2, \mathbf{t}|\mathbf{o}) = \frac{P(\mathbf{o}|\mu_M, \sigma_M^2, \mathbf{t})P(\mu_M, \sigma_M^2, \mathbf{t})}{P(\mathbf{o})}\,. \tag{2}$$

The left hand side of the equation is the joint probability of μ_M, σ_M^2, and the true values of the parallax and absolute magnitude of each star ($\mathbf{t} = \{\varpi_i, M_i\}$), given the observations $\mathbf{o}$. $P(\mathbf{o}|\mu_M, \sigma_M^2, \mathbf{t})$ is the probability of the data given the true values of the observables, i.e. the likelihood. $P(\mu_M, \sigma_M^2, \mathbf{t})$ is the joint probability distribution of the luminosity distribution model parameters and the true values of the observables. This term represents our prior information on the distribution of stars in space and luminosity. The term $P(\mathbf{o})$ is the probability of obtaining the data (evidence) and can be considered a normalizing constant which can be ignored for the optimization problem considered here.

In order to obtain estimates for μ_M and σ_M^2 we now seek to maximize the *posterior probability* $P(\mu_M, \sigma_M^2, \mathbf{t}|\mathbf{o})$, which we can write as:

$$\begin{aligned} &P(\mu_M, \sigma_M^2, \mathbf{t}|\mathbf{o}) \propto \\ &\prod_i \exp\left[-\tfrac{1}{2}\left(\tfrac{\varpi_{o,i}-\varpi_i}{\sigma_{\varpi,i}}\right)^2\right] \times \exp\left[-\tfrac{1}{2}\left(\tfrac{m_i - M_i + 5\log\varpi_i + 5}{\sigma_{m,i}}\right)^2\right] \times \\ &\varpi_i^{-4} \times \tfrac{1}{\sigma_M\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}\left(\tfrac{M_i-\mu_M}{\sigma_M}\right)^2\right] P(\mu_M)P(\sigma_M^2)\,. \end{aligned} \tag{3}$$

This represents the joint probability of $(\mu_M, \sigma_M^2, \mathbf{t})$ given $\mathbf{o}$. The first two terms express the assumption that the errors in the measured parallaxes and apparent magnitudes are Gaussian. The ϖ_i^{-4} term expresses our assumption that the stars are distributed uniformly in space around the Sun and the third Gaussian is the assumption on the luminosity distribution of the stars in our sample. The latter two terms constitute our model for the luminosity calibration problem. Finally $P(\mu_M)$ and $P(\sigma_M^2)$ represent our prior information on plausible values of μ_M and σ_M^2 (i.e. our model parameters).

Our job now is to work out the posterior probability from the above equation. We can then determine the probability distribution of (μ_M, σ_M^2) by marginalizing over the 'nuisance parameters' $\mathbf{t}$ (we are not interested in the true values of the parallaxes and absolute magnitudes of the individual stars). From this distribution we can then make estimates $\widetilde{\mu_M}$ and $\widetilde{\sigma_M^2}$, for example by taking the mean or the point with the maximum a-posteriori probability (MAP). The problem is of course that $P(\mu_M, \sigma_M^2, \mathbf{t}|\mathbf{o})$ is a function over a very high dimensional space and it will not be possible to calculate it analytically.

There are, however, numerical methods that allow one to construct a sampling of the posterior distribution. A popular example is the so-called Markov Chain Monte Carlo (MCMC) method which will be used in the example discussed here. Basically, the MCMC method produces an intelligent sampling of $P(\mu_M, \sigma_M^2, \mathbf{t}|\mathbf{o})$ by making a controlled random walk through the parameter space $(\mu_M, \sigma_M^2, \mathbf{t})$. The sampling will result in a distribution of points in the parameter space for which the density is proportional to the posterior distribution. From this distribution of points one can construct histograms of μ_M and σ_M^2 values. These histograms then represent a sampling of the posterior probability distributions for μ_M and σ_M^2 *marginalized over all other model parameters*. This means that these distributions represent our posterior knowledge of the luminosity function parameters taking into account *all* the uncertainties due to observational errors and the vagueness of our prior information.

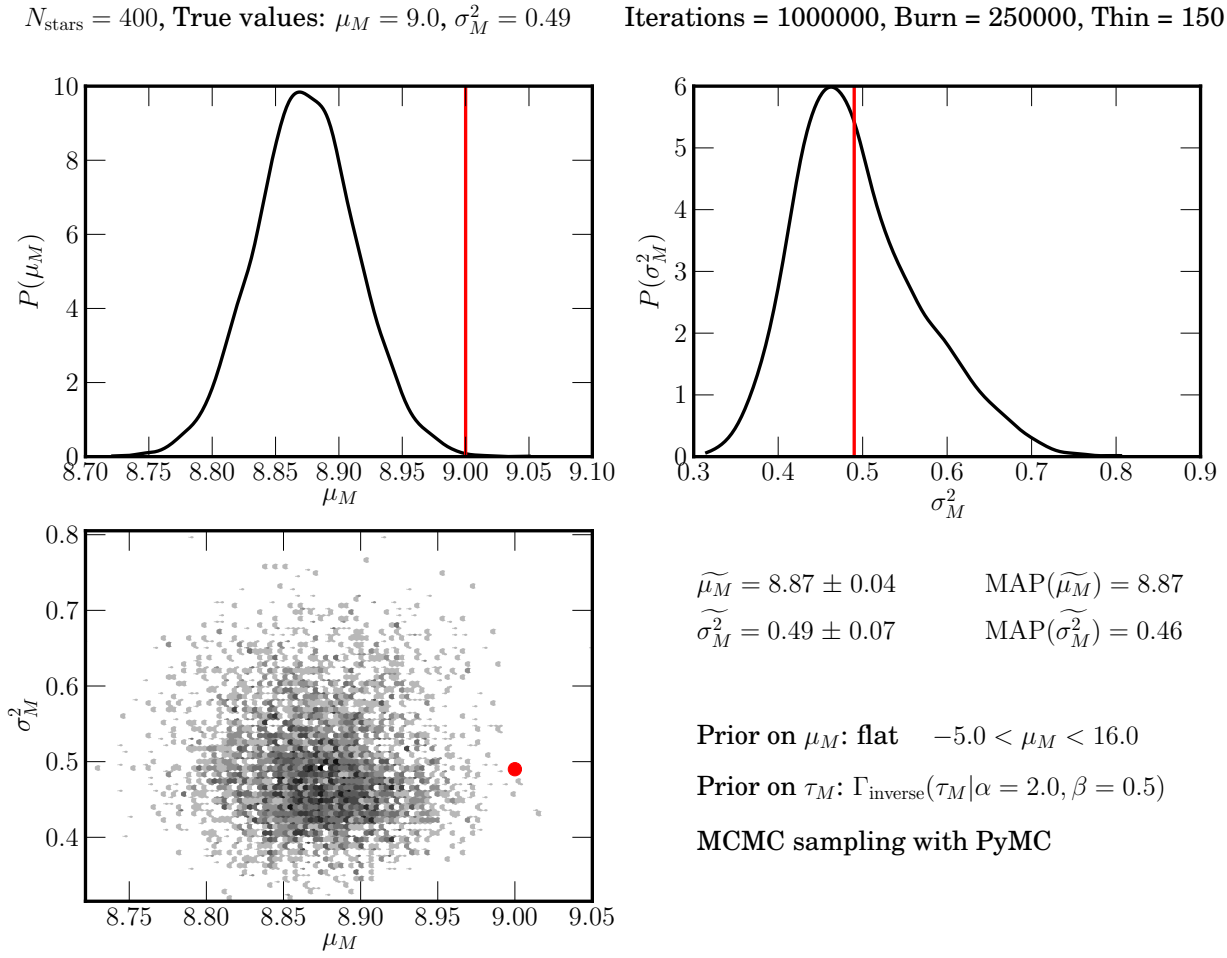

Figure 2: Example result of the MCMC sampling of the posterior distribution for μ_M and σ_M^2. In this case the `PyMC` package was used and the Γ-prior for σ_M^2. Note that in `PyMC` one has to sample the *precision* $\tau = 1/\sigma_M^2$ rather than the variance. The prior on τ is given by the inverse-Γ distribution with parameters $\alpha = k$ and $\beta = 1/\theta$. The panels show the marginal distributions of μ_M and σ_M^2 as well as the joint distribution. The points estimates obtained from the marginal distribution are also shown. The vertical red lines and the red dot indicate the true values.

We will apply this now to our simulated survey containing stars of a single type. The survey contains $N = 400$ stars for which the absolute magnitudes are drawn from a Gaussian luminosity function with $\mu_M = 9$ and $\sigma_M^2 = 0.49$. The stars are uniformly distributed between distances of 1 and 100 pc (i.e., 10 mas $\leq \varpi_i \leq$ 1000 mas) and the survey is assumed to be volume complete. The known observational errors $\sigma_{\varpi,i}$ and $\sigma_{m,i}$ vary as a function of the apparent magnitude as described above. We furthermore assume that we know the distribution of stars to be uniform in space and that we know the limits on the true parallax distribution. Hence, the only unknowns to estimate from the data are μ_M and σ_M^2.

Before proceeding we still need to specify the priors on μ_M and σ_M^2. The prior on μ_M is assumed to be flat between $\mu_M = -5$ and $\mu_M = +16$, while for σ_M^2 the prior distribution is assumed to be proportional to $1/\sigma_M^2$ or to be given by the $\Gamma(\sigma_M^2|k,\theta)$ distribution. The first choice of prior probability distribution for σ_M^2 represents a so-called non-informative prior for a scale parameter (σ_M^2 setting the width of the luminosity distribution of the stars). However, this prior strongly favours small values of σ_M^2, while in contrast the Γ function leads to low probabilities for very small values of σ_M^2 and the shape can be tuned to make it fairly flat at larger values of σ_M^2. In addition the Γ function is properly normalized.

Figure 2 shows an example of the MCMC sampling of the posterior probability distribution of μ_M and σ_M^2. The prior on σ_M^2 was the Γ function with $k = 2$ and $\theta = 1$. The MCMC method was run on the expression for

the posterior probability given in equation (3), using the `PyMC` package and 10^6 iterations with 250 000 'burn in' steps, storing every 150th sample.

3 Python technicalities

I have prepared python code that uses either the `PyMC` (`https://github.com/pymc-devs/pymc`) or the `emcee` (`http://dan.iel.fm/emcee/`) package for the MCMC sampling. You should install these packages and in addition you will need to install the `acor` (`https://pypi.python.org/pypi/acor`) and `PyTables` (`http://pytables.github.io/`) packages. `PyTables` may already be part of your python installation. Note that in all cases you can use the following command to install the package without needing root rights:

```
python setup.py install --user
```

The installation will then end up in the `~/.local` folder. If you do have root rights (on your notebook for example) you do not need the `--user` option.

NOTE that `numpy` **version 1.6.2 or higher** is required.

The code for this exercise can be downloaded from `https://github.com/agabrown/AstroStats-II`. Just unpack the code into your folder of choice and start using it. No need to install it. The following python scripts can be used to run the MCMC sampling of the luminosity calibration model.

`runLCM.py`	Uses `PyMC` and inverse-Γ prior on $\tau = 1/\sigma_M^2$
`runLCMBook.py`	Uses `PyMC` and $1/\tau$ prior on $\tau = 1/\sigma_M^2$
`runLCMemcee.py`	Uses `emcee` and Γ prior on σ_M^2
`runLCMemceeBook.py`	Uses `emcee` and $1/\sigma_M^2$ prior on σ_M^2

In all cases you can invoke the script with `python script.py --help` to get documentation on the necessary command line options. The `emcee` based scripts do not store the results and produce a plot straight away. The `PyMC` based scripts store the results in HDF5 files (for which `PyTables` is needed) called:

```
LumCalResults-40-10.0-1000.0-9.0-0.49.h5
LumCalSimSurvey-40-10.0-1000.0-9.0-0.49.h5
```

for example. The first file contains the MCMC results and the second the simulated parallax survey data.

Example runs for `PyMC`:

```
python runLCM.py --mcmc 100000 25000 15 --survey 40 10 1000 9.0 0.49 Inf --priors -5 16 2 0.5
```

or

```
python runLCMBook.py --mcmc 100000 25000 15 --survey 40 10 1000 9.0 0.49 Inf --priors -5 16 1 100
```

Plot results with:

```
python plotLCM-MCMC-results.py LumCalSimSurvey-40-10.0-1000.0-9.0-0.49.h5 \
                                    LumCalResults-40-10.0-1000.0-9.0-0.49.h5
```

Example runs for `emcee`

```
python runLCMemcee.py --mcmc 100 100 1 10 --survey 40 10 1000 9.0 0.49 Inf --priors -5 16 2 0.5
```

or

```
python runLCMemceeBook.py --mcmc 100 100 1 10 --survey 40 10 1000 9.0 0.49 Inf --priors -5 16 1 100
```

You will note that with 40 stars it is not easy to estimate the parameters of the luminosity distribution, especially the variance. The run-time of `emcee` is much shorter but it seems to have more difficulty to converge on the correct answer.

This is a difficult MCMC problem!

4 Purpose of this exercise

What to do with the code provided? The idea of this exercise is to illustrate a more complex application of Bayesian inference in the context of a problem one will surely encounter with Gaia. Simply try out different samples sizes, distance limits, values of μ_M (faint, bright), values of σ_M^2 (small, large). Try to also get a feeling for the effect of the priors on μ_M and σ_M^2; you can even code your own different priors.

Note how this method is capable of dealing with negative parallaxes and errors varying as a function of apparent magnitude. All information is used!

What I have not done for this exercise is to tailor the MCMC sampling algorithms to the problem at hand. This goes somewhat beyond the scope of the lectures but is probably needed in order to deal with the effect of the steep probability density function for the parallaxes.

Some questions to consider:

- How would you deal with unknown distance limits?
- How could a different distribution of the stars (i.e., non-uniform) be accomodated?
- What to do in the releastic case of a magnitude limit imposed on the survey?