Statistics for Astronomers: Lecture 23, 2019.05.22

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: Incorporating detection limits

see Appendix in Sawicki 2012 PASP 124, 1208

A faint source may be undetected during an observation. The detection limit is the upper limit to the true flux of a faint source.

Assuming that the background noise is normally distributed, the detection limit usually specified in terms of the background noise. *E.g.*, "the 3-σ upper limit is 0.5 μJy" means that the probability that the true flux of the faint source is below 0.5 μJy is about 98.75%.

Suppose our model predicts a flux of $F_{\rm mod}$ for this source. The probability that the source flux is drawn from this model is

$$P \propto \int_{-\infty}^{F_{\rm lim}} dF \; \exp\left[-\frac{1}{2}\left(\frac{F - F_{\rm mod}}{\sigma^2}\right)^2\right] = \sqrt{\frac{\pi}{2}}\sigma\left[1 + \mathrm{erf}\left(\frac{F_{\rm lim} - F_{\rm mod}}{\sigma}\right)\right].$$

For multi-band photometry combining detections and non-detections,

$$\chi^2 = \sum_i \left(\frac{F_i - F_{{\rm mod},i}}{\sigma_i}\right)^2 + \sum_j \ln\left\{\sqrt{\frac{\pi}{2}}\sigma_j\left[1 + \mathrm{erf}\left(\frac{F_{{\rm lim},j} - F_{{\rm mod},j}}{\sigma_j}\right)\right]\right\}.$$

χ^2 minimisation can now be performed numerically to obtain the best-fit model parameters.

Same technique can be applied to bright sources – the integration limits then go from $F_{\rm sat}$ to ∞.

Recall: Bayesian Outlier rejection

We discuss one example of robust regression, Bayesian Outlier Rejection (see Hogg et al. (2010) and Section 8.9 in the AstroML book). The method is similar to assuming a Gaussian mixture model for the data.

Assumptions:

– The uncertainties associated with “true data” are distributed according to $\mathcal{N}(0, \sigma^2)$.

– Outliers are generated from a Gaussian distribution with mean Y_b with variance V_b (substantially larger than σ).

– The probability that a given data point is an outlier in the resulting Gaussian mixture is P_b. Or, equivalently, we can flag each point according to whether or not we think it is an outlier. Each point then has an associated flag variable q_i ($q_i = 0$ if the point is “bad”, 1 if “good”).

We want to fit a straight line to the “good” points. In addition to m and b, we now have $N + 3$ extra parameters. The q_i are nuisance parameters which we can marginalise over. BUT for a given point j we could also marginalise over all other parameters except q_j to see if it was flagged as a true data point or an outlier! This is the strength of the Bayesian method.

Recall: marginalisation in Bayesian outlier rejection

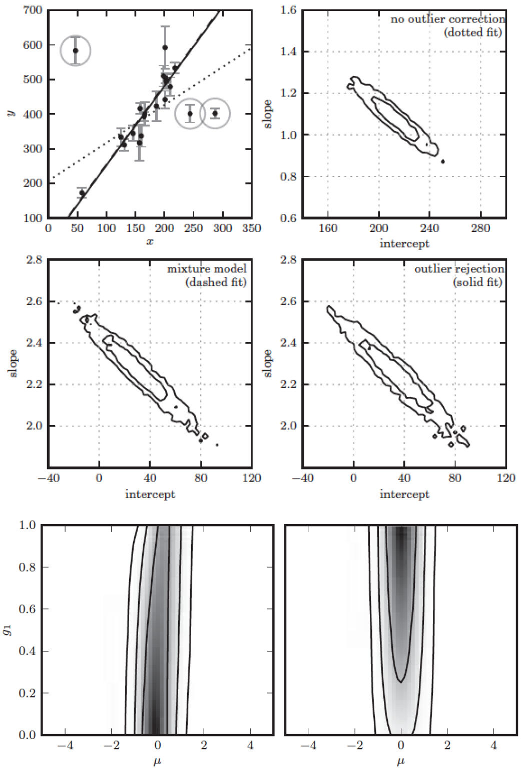
The posterior is $\propto$ likelihood $\times$ the priors.

We can marginalise this posterior over the nuisance parameters q_i to obtain the joint distribution of m and b.

Since the q_i are discrete (they take values 0 or 1), marginalising over them means summing over these possible values instead of integrations.

Once this is done, we also marginalise over P_b, V_b, and Y_b.

This is a multidimensional problem, perfect for MCMC. The implementation is part of the AstroML book (Section 8.9). Part of the homework this week.



source: AstroML book Sections 5.6.7 and 8.9

Goodness-of-fit and unknown uncertainties

Is my [linear] model a good fit to the data?

Common (frequentist) procedure: compute χ^2, and check if its value is outside $(N-2) \pm \sqrt{2(N-2)}$.

In the Bayesian context, we can't reject the model, we can only say if one model family is better than another.

For the assumptions: linear model + Gaussian uncertainties,

– Because of shape of χ^2 distribution, significantly higher values than $(N-2)$ are easier to obtain than significantly lower values. The latter is not impossible, though.

– Do not reject models based on the χ^2 value. Under/overestimated or correlated uncertainties can cause bad models to have reasonable χ^2 and vice versa. Underestimated uncertainties usually result in high χ^2 for reasonable models.

If you don't trust your estimated uncertainties, use the Bayesian method – infer and marginalise them away!

Bivariate measurement uncertainties + intrinsic scatter

What to do when both the independent and dependent variables have measurement error and intrinsic scatter?

These are called error-in-variables or measurement error models in that they account for uncertainties in the independent variable as well.

For example, Kelly 2007 lists a large number of available methods for this purpose.

Some computational resources

Read through the introduction here: https://github.com/rsnemmen/BCES. The package includes a BCES (bivariate correlated errors and intrinsic scatter) code for linear fitting, and also includes scripts to plot confidence bands for the fits.

The README file also provides some background on newer Bayesian methods for this purpose. Of course, the AstroML contribution is one of the most recent `Python` codes for this purpose: http://www.astroml.org/book_figures/chapter8/fig_total_least_squares.html.

Let's look at this method (same as that of Sections 7 and 8 in Hogg et al. 2010) in some detail...

Fitting bivariate uncertainties

"Total least squares" or "errors-in-variables" fitting (because it accounts for bivariate correlated uncertainties).

Each point now has a covariance tensor component $\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_{xi}^2 & \sigma_{xyi} \\ \sigma_{xyi} & \sigma_{yi}^2 \end{bmatrix}$, with $\sigma_{xyi} = \rho_{xyi}\sigma_{xi}\sigma_{yi}$.

Either the uncertainties are given to be Gaussian, or we can assume they are (given only the variance, the Gaussian is the maximum-entropy distribution).

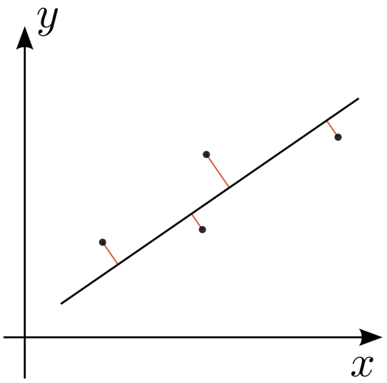
The probability of observing a pair (x_i, y_i) given the true data values (x, y) is a bivariate Gaussian

$$p(x_i, y_i|x, y, \mathbf{\Sigma}_i) = \frac{1}{2\pi\sqrt{\mathrm{Det}(\mathbf{\Sigma}_i)}} \exp\left[-\frac{1}{2}(\mathbf{Z}_i - \mathbf{Z})^T {\mathbf{\Sigma}_i}^{-1}(\mathbf{Z}_i - \mathbf{Z})\right], \text{ with } \mathbf{Z} = \begin{bmatrix} x \\ y \end{bmatrix}, \mathbf{Z}_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$$

In this sense, "error bars" can be shown as ellipses. The orientation of each uncertainty ellipse is decided by the off-diagonal elements of $\mathbf{\Sigma}_i$.

Fitting bivariate uncertainties (contd.)

One way to describe this problem is to minimise the perpendicular distance of each point from the best-fit line. The model assumes that the true (x, y) pairs lie exactly on the line, and any deviation is due to the measurement uncertainties.



(credit: User:Netheril96/CC BY-SA 3.0)

A unit vector along the line is

$$\hat{\mathbf{v}} = \frac{1}{\sqrt{1+m^2}}\begin{bmatrix} -m \\ 1 \end{bmatrix} = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}$$

where $\theta = \tan^{-1} m$.

Orthogonal displacement of a point $\mathbf{Z}_i$ from the line:
$\mathbf{\Delta}_i = \hat{\mathbf{v}}^T\mathbf{Z}_i - b\cos\theta \equiv \hat{\mathbf{v}}^T\mathbf{Z}_i - b_\perp$
($b_\perp$ is the perpendicular distance of the line from the origin).

"Orthogonal variance": $\sigma_{\mathrm{orth},i}^2 = \hat{\mathbf{v}}^T\mathbf{\Sigma}_i\hat{\mathbf{v}}$.

$$\Longrightarrow \text{Log-likelihood: } \ln\mathscr{L} = \text{const.} - \frac{1}{2}\sum_{i=1}^{N}\frac{\Delta_i^2}{\sigma_{\mathrm{orth},i}^2}.$$

Perform MLE (can solve numerically for the parameters).
Advice: maximise w.r.t. $(\theta, b_\perp)$, not (m, b).

For example, $\theta \sim U(0, \pi/2)$ samples the first quadrant uniformly in angle but $m \sim U(0, \infty)$ produces more samples at higher slopes (because of the higher dynamic range).

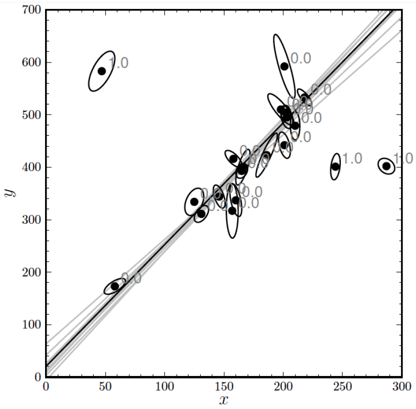
Fitting bivariate uncertainties and intrinsic scatter

Consider the special case where we quantify the intrinsic scatter orthogonal to the line.

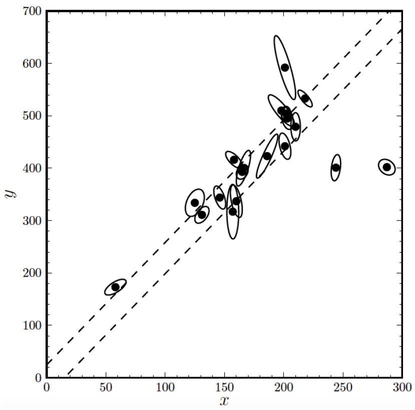
Recall that σ_i^2 was the orthogonal variance due to measurement uncertainties in the i^{th} observation. Let V be the orthogonal variance due to intrinsic scatter. We now have three parameters to estimate: $\theta, b_\perp$, and V.

The extension of the previous analysis is simple: add the orthogonal variances together.

Log-likelihood: $\ln \mathscr{L} = \mathrm{const.} - \frac{1}{2}\sum_{i=1}^{N} \ln\left(\sigma^2_{orth,i} + V\right) - \frac{1}{2}\sum_{i=1}^{N} \frac{\Delta_i^2}{\sigma^2_{orth,i} + V}.$



Hogg et al. (2010) fits incorporating bivariate measurement uncertainties (left) and intrinsic scatter (right).



Notes

The method presented here does not account for variation along the line, whether due to measurement uncertainties or intrinsic variances. There are many other methods available. Read through Kelly (2007) and the notes in Hogg et al. (2010) for references.

If you want to learn how to fit your data, you must at the very least run through the exercises in Hogg et al. (2010). There are many other ways of treating this problem.

Look through the AstroML book and/or Modern Statistical Methods for Astronomy by Feigelsen & Babu.

Tomorrow: review

Go through the course notes and homework questions. Is there anything still bugging you?

Are there any topics you had questions about, especially concerning your research, that are still unanswered?