Statistical Tools and Techniques for Solar Astronomers

Alexander W Blocker Nathan Stein

SolarStat 2012

Outline

1. Introduction & Objectives
2. Statistical issues with astronomical data
3. Example: filter / hardness ratios
4. Standard approach
5. Building a statistical model
6. Using the model

Outline

1. Introduction & Objectives
2. Statistical issues with astronomical data
3. Example: filter / hardness ratios
4. Standard approach
5. Building a statistical model
6. Using the model

Introductions

Alex Blocker, Harvard Statistics

- Model-based stacking for low-count observations
- Event detection for massive light curve databases

Nathan Stein, Harvard Statistics

- Analysis of CMDs in stellar clusters
- Robust clustering methods for astronomical datas

Objectives for the day

Disclaimer: Will not make statisticians in two hours

Goals:

- Awareness of statistical issues and concepts
- Understanding of probability modeling approach
- Familiarity with computational tools

Basically, want informed statistical consumers

Background

- Assuming little statistical background
- However, should have basic understanding of probability
- *Not* assuming knowledge of Bayesian modeling, MCMC, etc.

Outline

1. Introduction & Objectives
2. Statistical issues with astronomical data
3. Example: filter / hardness ratios
4. Standard approach
5. Building a statistical model
6. Using the model

Sources of error

- Raw data typically consist of photon counts
- Measurement noise and intrinsic variation
- Contamination from background
- Inhomogeneous instrumental sensitivities

Forms of data

Starts with images, but many ways to extract numerical data (increasing order of structure and complexity):

- Point and area measurements (predefined)
- Light curves
- Spatiotemporal on predefined regions (local)
- Global spatiotemporal patterns

Focus

Focusing today on simplest structure — measurements on predefined regions

- Core modeling is shared across settings
- Computational strategies are similar, but greater sophistication is needed for more complex settings
- Analyses of light curves, whole images, etc. add layers of structure

Outline

1. Introduction & Objectives
2. Statistical issues with astronomical data
3. Example: filter / hardness ratios
4. Standard approach
5. Building a statistical model
6. Using the model

Problem definition

Interested in relative flux of source or region between two energies

- Filter ratios in solar
- Hardness ratios in high-energy

DEMs preferred to filter ratios for solar (e.g. Weber et al. 2005)

However, ratios provide a straightforward setting to work with; DEM analysis is an extension

Definitions

Denoting fluxes in hard and soft passbands as λ_H and λ_S

- Simple ratio

$$\mathcal{R} = \frac{\lambda_S}{\lambda_H}$$

- Color

$$\mathrm{C} = \log_{10}\left(\frac{\lambda_S}{\lambda_H}\right)$$

Data

Observations

- Photon counts from region of interest in hard (H) and soft (S) passbands — extracted from images
- Similar counts from area with background only in each passband (B_H and B_S)

Calibration

- Sensitivity of instrument to each band e_H and e_S (effective area)
- Relative effective area for background region r

Outline

1 Introduction & Objectives

2 Statistical issues with astronomical data

3 Example: filter / hardness ratios

4 Standard approach

5 Building a statistical model

6 Using the model

Simple case

Without background or other corrections, standard approach just substitutes counts for fluxes:

$$\mathcal{R} = \frac{S}{H}$$

$$\mathrm{C} = \log_{10}\left(\frac{S}{H}\right)$$

Corrections

Adjusting for background, standard approach would use:

$$\mathcal{R} = \frac{S - B_S/r}{H - B_H/r}$$

$$\mathrm{C} = \log_{10}\left(\frac{S - B_S/r}{H - B_H/r}\right)$$

Error estimates

Standard errors of these are usually propagated Gaussian approximation (linear approximation):

$$\sigma_{\mathcal{R}} = \frac{S - B_S/r}{H - B_H/r}\sqrt{\frac{\sigma_S^2 + \sigma_{B_S}^2/r^2}{(S - B_S/r)^2} + \frac{\sigma_H^2 + \sigma_{B_H}^2/r^2}{(H - B_H/r)^2}}$$

$$\sigma_{\mathrm{C}} = \frac{1}{\ln(10)}\sqrt{\frac{\sigma_S^2 + \sigma_{B_S}^2/r^2}{(S - B_S/r)^2} + \frac{\sigma_H^2 + \sigma_{B_H}^2/r^2}{(H - B_H/r)^2}}$$

where σ_S, σ_H, σ_{B_S}, and σ_{B_H} are typically approximated with the Gehrels prescription (Gehrels 1986)

$$\sigma_X \approx \sqrt{X + 0.75} + 1$$

From sigma to intervals

Typically not interested in σ for its own sake

- Want to summarize uncertainty about $\mathcal{R}$ or C
- Standard statistical approach is to use intervals
- Often constructed as $\hat{\theta} \pm k \cdot \sigma$
- Confidence interpretation: want interval to include true value at least as often as stated
- e.g., for Gaussian data, $\bar{X} \pm \sigma$ is a 68% interval; $\bar{X} \pm 1.96\sigma$ is 95%

Flaws

- Sigma does not summarize errors on $\mathcal{R}$ or C; actual uncertainty can be highly asymmetric
- Gaussian assumption is flawed for low-count observations; intervals are not valid
- Background subtraction leads to bias and inefficiency (van Dyk et al. 2001)
- Not accounting for differences in detector sensitivity effectively

Outline

1 Introduction & Objectives

2 Statistical issues with astronomical data

3 Example: filter / hardness ratios

4 Standard approach

5 Building a statistical model

6 Using the model

Models vs. procedures

- Classical approach is set of procedures; not derived from deeper framework
- Model-based approach starts from description of data-generating process
- Model is realistic (though not necessarily physical) description of underlying mechanisms
- Why model? Efficient use of information, consistency, and incorporation of complex error structure

Parameters vs. observations

Parameters regulate underlying processes (source and observation)

- Ideally invariant to detail of observation structure (e.g. flux, not expected counts)
- Target of inference
- For hardness ratios, parameters are source fluxes (λ_H and λ_S) and background fluxes (ξ_H and ξ_S)

Observations are noisy outputs of parameters

- S, H, B_S, and B_H in ratio problem
- Input for, not target of, inference

Distributions as connections

Connect parameters to observations through distributions

- Background counts depend only on background flux and exposure

$$\begin{aligned} B_S &\sim \text{Poisson}(r \cdot e_S \cdot \xi_S) \\ B_H &\sim \text{Poisson}(r \cdot e_H \cdot \xi_H) \end{aligned}$$

  where e is the effective area for the source region

- Source counts depend on both source and background fluxes

$$\begin{aligned} S &\sim \text{Poisson}(e_S \cdot (\lambda_S + \xi_S)) \\ H &\sim \text{Poisson}(e_H \cdot (\lambda_H + \xi_H)) \end{aligned}$$

Augmentation

Sometimes useful to expand model by expanding observations

- Looks like adding complication, but can simplify computation and help with interpretation
- Usually ask "what observations would make this problem easy?"
- For ratio case, would be easy if we knew which parts for S and H came from source vs. background
- So, augment with background counts

$$\eta_S \sim \text{Poisson}(e_S \cdot \lambda_S) \quad \text{and} \quad \eta_H \sim \text{Poisson}(e_H \cdot \lambda_H),$$
$$\beta_S \sim \text{Poisson}(e_S \cdot \xi_S) \quad \text{and} \quad \beta_H \sim \text{Poisson}(e_H \cdot \xi_H),$$
$$S = \eta_S + \beta_S \quad \text{and} \quad H = \eta_H + \beta_H$$

Outline

1 Introduction & Objectives

2 Statistical issues with astronomical data

3 Example: filter / hardness ratios

4 Standard approach

5 Building a statistical model

6 Using the model

Likelihood

Likelihood is at the core of model-based inference

Definition

Likelihood is the probability of observing your (fixed) sample, as a function of the parameters. If your sample is Y and parameters are θ, likelihood is $L(\theta) \propto Pr(Y|\theta)$.

- Likelihood is *not* the probability of your parameters taking on a particular value
- Higher values of likelihood indicate more support from data for given parameter value
- Likelihood function contains all information for inference with given model

Likelihood, in particular

For independent observations, likelihood is just the product of their probabilities.

So, for the ratio problem, the likelihood is:

$$\begin{aligned} L(\lambda_S, \lambda_H, \xi_S, \xi_H) \quad \propto \quad & P(B_S|\xi_S) \cdot P(S|\lambda_S, \xi_S) \cdot \\ & P(B_H|\xi_H) \cdot P(H|\lambda_H, \xi_H) \end{aligned}$$

Here, all of these probabilities take the form of the Poisson PMF

MLE

One way to use the likelihood is to find the parameter values that maximize it; known as *maximum likelihood estimation*

- Resembles χ^2 fitting, but error measures need not be squared
- Has some desirable properties in large samples (efficiency, known approximate errors, etc.)
- Requires numerical maximization for most realistic models
- Can be badly misleading for small samples and settings with highly asymmetric uncertainty

Bayesian inference

Uses Bayes Theorem to quantify uncertainty and perform estimation

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

- $P(\theta|Y)$ is posterior distribution of parameters
  - Estimates from mean, median, etc. of this distribution
  - Intervals, standard errors, etc. from its quantiles and spread
- Can derive or simulate posterior of any function of θ using this posterior
- Drawback: need prior $P(\theta)$
- Typically aim to choose prior that has little effect on results; check through sensitivity analysis

Implementation

On to Nathan for computation!