Data Analysis Methods

Successes, Opportunities, and Challenges

Chad M. Schafer
Department of Statistics & Data Science
Carnegie Mellon University
cschafer@cmu.edu

The Astro/Data Science Community

LSST Informatics and Statistics Science Collaboration
Over 75 Members

International Astrostatistics Association
Over 400 Members

American Stat. Assoc. Astrostatistics Interest Group
Over 50 Members

Individuals motivated by **advancing science** with **novel data analysis tools**

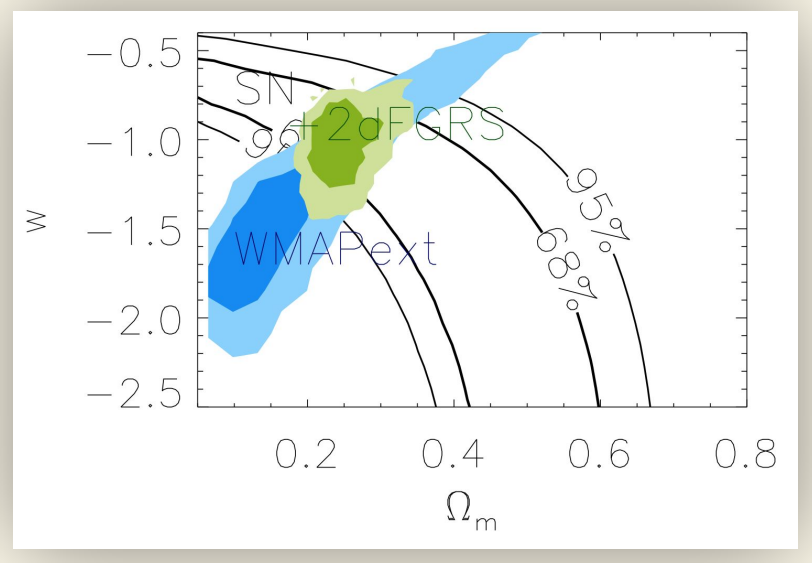
Success Story: MCMC in Cosmology

Markov Chain Monte Carlo (MCMC) - Class of methods for parameter estimation, approximation of posterior

Algorithms date to mid-20th Century, but development took off in Statistics during 1990s

By the 2000s, MCMC playing key role in cosmology, including WMAP analysis. (Spergel 2013)

Focus on development of resources for MCMC in cosmology (e.g., CosmoMC)

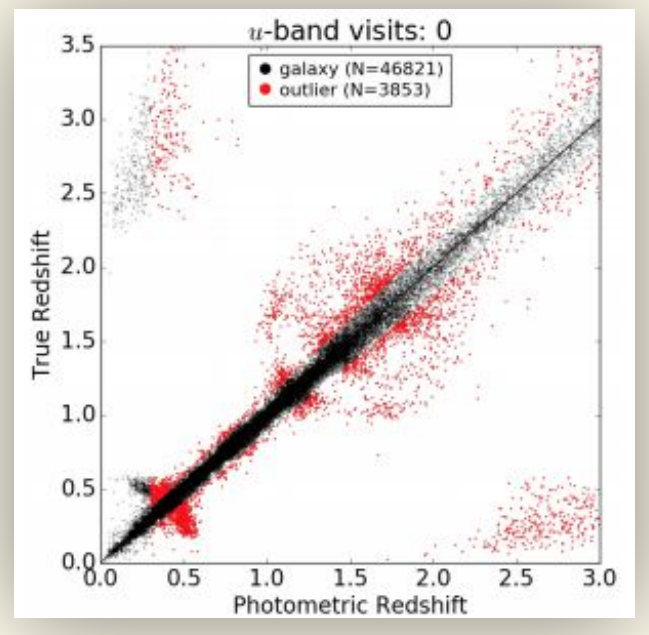


Success Story: Redshift Estimation

Estimation of Redshifts using Photometry -
Crucial challenge faced by photometric surveys

Range of methods proposed and attempted, using **training sets** built from spectroscopic samples.

Graham, et al. (2017): “Photometric redshifts are an essential part of every cosmological science goal of the LSST.”



Success Stories

What do these successes have in common?

Smaller gap between existing, advanced methods and the application

- **MCMC** naturally addressed challenges of estimation of cosmological parameters with CMB
- **Photometric redshift estimation** can be viewed as a standard supervised learning problem

In both cases the methodology is crucial to the science

The Potential

Theme: Making full use of data from large surveys

Sample size alone does not quantify information content

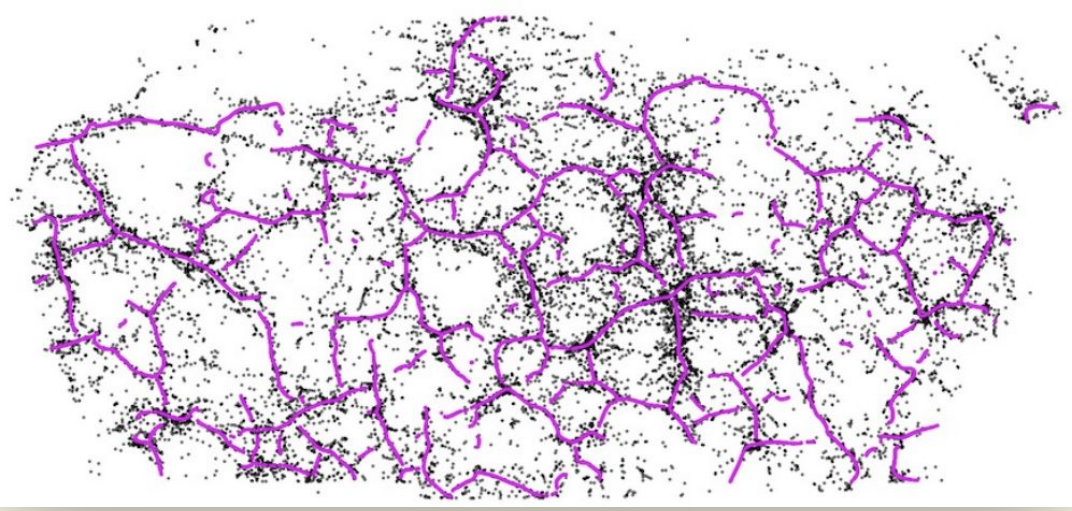
More data reduces variance, **better modelling** reduces bias

Specific areas for potential gains:

- Information preservation when “representing” raw data
- Improved propagation of errors
- Lessening bias from incorrect distributional assumptions
- Joint analyses of data sets

Example: Filament Catalog

Chen, et al. (2015): Density ridges in galaxy map lead to finding **filamentary structure**



SDSS filament catalog: https://sites.google.com/site/yenchicr/catalogue

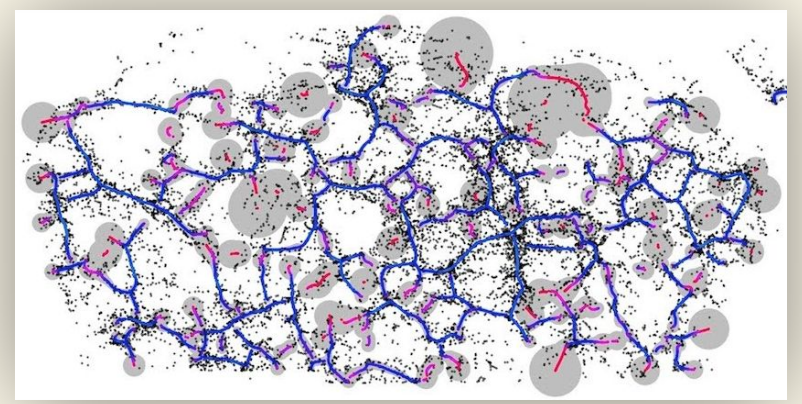
Plans to create for LSST

Example: Filament Catalog

Example of **complex data representation**

Nonparametric - Does not make restrictive assumptions

Emphasis on **error quantification**



Broad Potential: Create catalogs of rich scientific information, avoiding overcompression

Example: Model Misspecification Bias

Long (2017): Improving period estimation from light curves through better adjustment for measurement errors

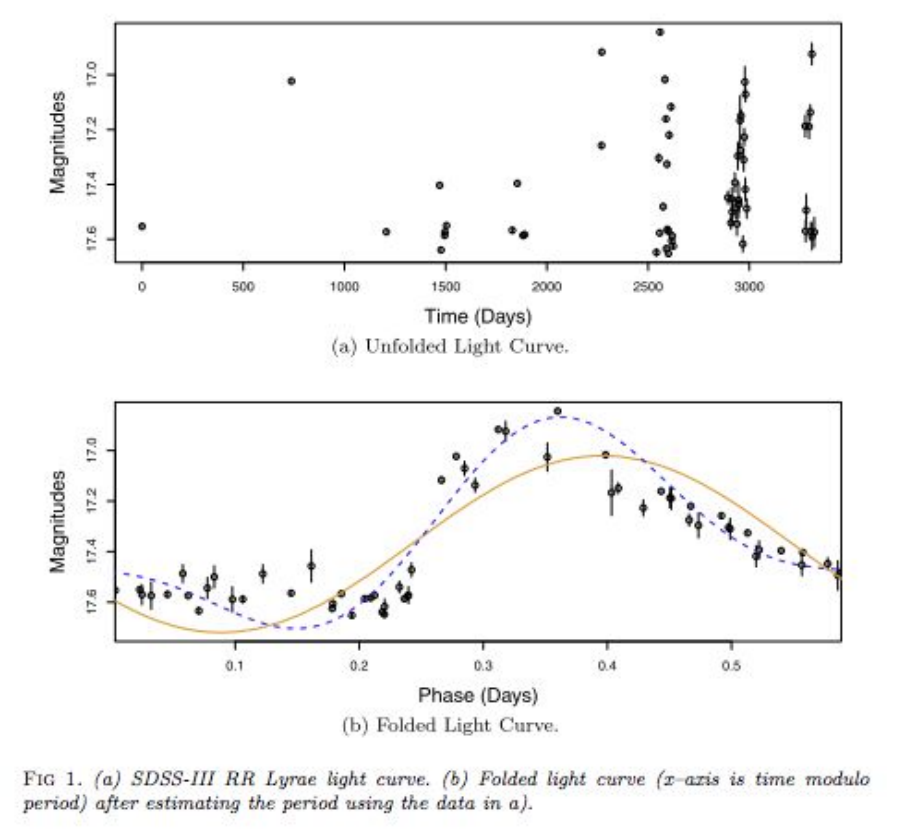


FIG 1. *(a) SDSS-III RR Lyrae light curve. (b) Folded light curve (x–axis is time modulo period) after estimating the period using the data in a).*

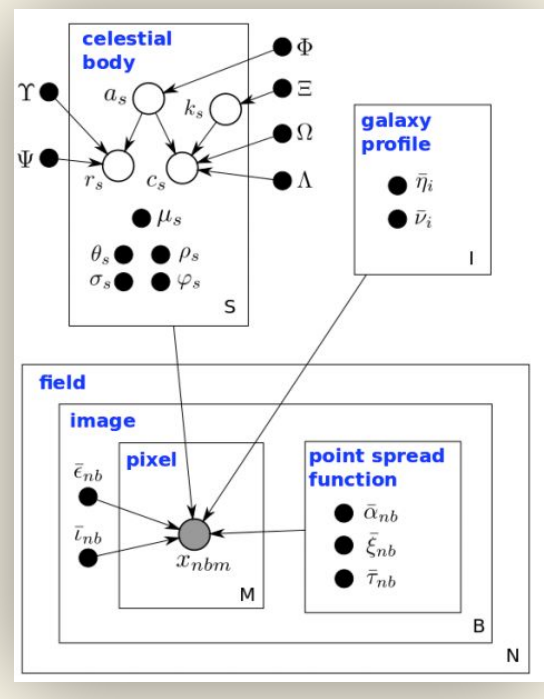
Models for light curves **misspecified**, but period estimates remain accurate

Broad Potential: Accept that models are incorrect, bias can be reduced through error quantification

Example: Variational Inference

Reiger, et al. (2015): Multiparameter, hierarchical **generative model** for pixel intensity from a star/galaxy

Variational inference approximates posterior when using models of such complexity



Instead, we use optimization to find a distribution that best approximates the posterior. For any distribution q over Θ,

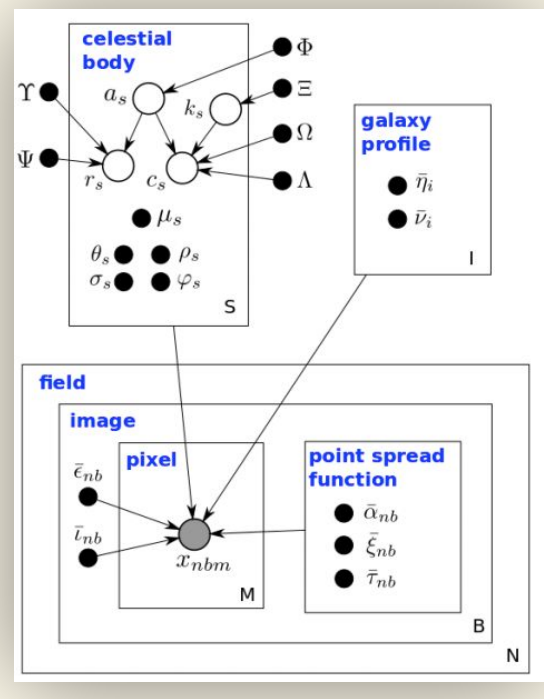
$$\log p(x) \geq \mathbb{E}_q[\log p(x, \Theta)] - \mathbb{E}_q[\log q(\Theta)] \quad (21)$$
$$=: \mathcal{L}(q). \quad (22)$$

Here $\mathbb{E}_q$ is expectation with respect to q. We call $\mathcal{L}$ the evidence lower bound (ELBO). To find a distribution q^* that approximates the exact posterior, we maximize the ELBO over a set $\mathcal{Q}$ of candidate q's. We restrict $\mathcal{Q}$ to distributions of the factored form

$$q(\Theta) = \prod_{s=1}^{S} q(a_s) q(r_s|a_s) q(k_s|a_s) \prod_{b=1}^{B-1} q(c_{sb}|a_s). \quad (23)$$

Example: Variational Inference

Reiger, et al. (2015): Multiparameter, hierarchical **generative model** for pixel intensity from a star/galaxy



Variational inference approximates posterior when using models of such complexity

Broad Potential: Variational inference could enable more realistic model fits

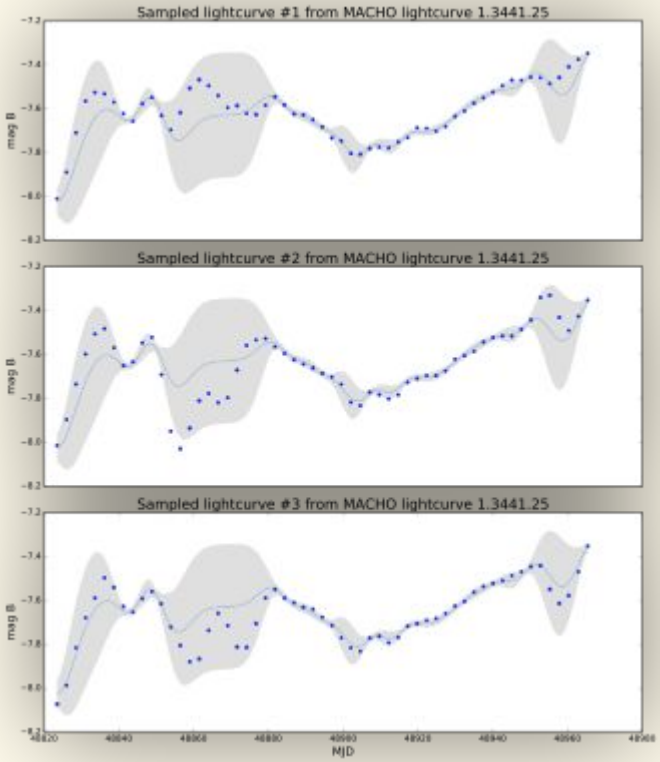
Example: Light Curve Classification

Castro, et al. (2017): “Classification of Variable Stars Handling Observational Gaps and Noise”

Bootstrap sampling of models fit to light curves in order to assess stability of features

Probabilistic classification based on feature uncertainty

Broad Potential: Quantification of uncertainty crucial in classification



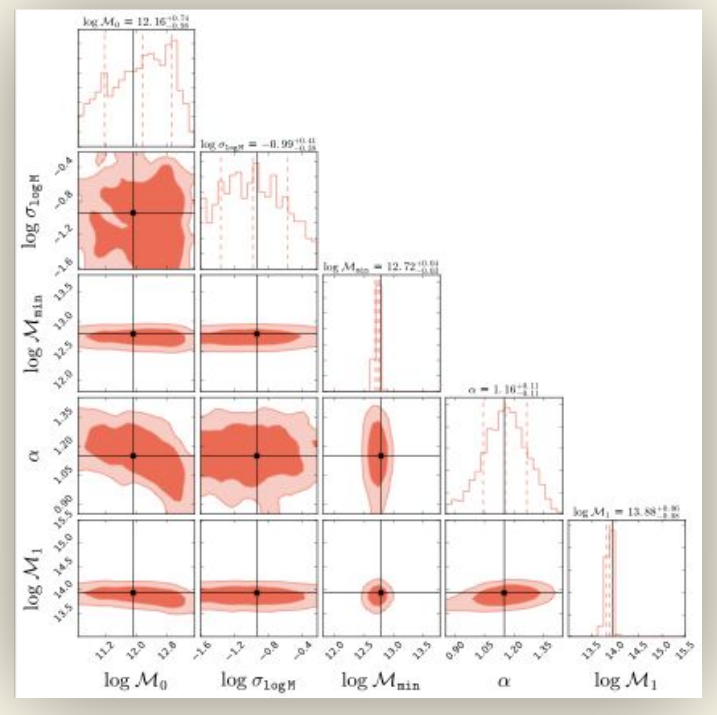
Example: Likelihood Free Inference

Hahn, et al. (2017): “Approximate Bayesian computation in large-scale structure.”

Complexity of LSS makes impossible writing likeilhood function of sufficient accuracy.

ABC offers approach which circumvents likelihood

Broad Potential: Errors in likelihood are magnified with large samples, esp. comparing simulations and data



The Examples

What do the examples have in common?

The shift from the **variance-dominated era** to the **bias-dominated era** requires careful thought about methods, not just the scaling of existing methods.

The gap from method to application is larger, and requires

- adaptation of existing methods
- deep knowledge of application
- **expertise on both data science and astronomy sides**

Barriers to Progress

Complexities

Communication

Computations

Compensation

Why am I here?

Working to **close the gap**

Increased role in this discussion

Data analysis tools as a **crucial community resource**

Start of work towards a community white paper

Massive, rich data sets are an huge opportunity, but reaching full potential requires **avoiding overcompression** and **adequate modelling**

References

Castro, et al. (2017) *Astronomical Journal*, Vol. 155

Chen, et al. (2017) *MNRAS*, Vol. 466

Graham, et al. (2017) arXiv 1706.09507

Hahn, et al. (2017) MNRAS, Vol. 469

Long (2017) *Elec. J. of Statistics*, Vol. 11

Reiger, et al. (2015) *Proceedings of ICML*