elementary?

Teaching advanced methodology to astronomers

astrostatistics and astroinformatics

Eric Feigelson

Center for Astrostatistics

Penn State University

BigSkyEarth COST conference

Sorrento IT October 2016

about Eric Feigelson ….

Professor of Astronomy & Astrophysics and of Statistics, Penn State
Associate Director, Center for Astrostatistics, Penn State
Statistical Scientific Editor, AAS Journals (Astrophys J, Astron J)
President, IAU Commission B3, Astroinformatics & Astrostatistics
Councils, ISI/IAA, AAS/WGAA, LSST/ISSC
Co-editor, Astrostatistics & Astroinformatics Portal (http://asaip.psu.edu)
Lead author, *MSMA* textbook

also an X-ray astronomer who studies star formation

Statistical questions for a scientist

- How reliable are my results? (P=0.05 vs. P=0.003)
- Have I used the most effective procedures for achieving my results? (KS vs. AD test)
- Do my results depend on uncertain models? (power law?)
- Do my results depend on my chosen analysis path?

A suggested path:

- Nonparametric data analysis (including hypothesis tests, local regression, Fourier or wavelet transforms, etc)
- Clustering & classification (if dataset is inhomogeneous)
- Parametric analysis: maximum likelihood with model selection
- Bayesian analysis: weight likelihood with prior knowledge
- Discuss: What have I learned from each step along the path?

Astronomy involves many branches of modern statistics …

Cosmology		**Statistics**
Star/galaxy discrimination	⟷	Image processing & classification
Galaxy clustering	⟷	Spatial point processes, clustering
Galaxy morphology	⟷	Regression, mixture models
Galaxy luminosity fn	⟷	Gamma distribution
Power law relationships	⟷	Pareto distriution
Weak lensing morphology	⟷	Geostatistics, density estimation
Strong lensing morphology	⟷	Shape statistics
Strong lensing timing	⟷	Time series with lag
Faint source detection	⟷	False Discovery Rate
SN Ia & QSO lightcurves	⟷	Nonstationary time series
CMB spatial analysis	⟷	Markov fields, ICA, etc
ΛCDM parameters	⟷	Bayesian inference & model selection
Comparing data & simulation	⟷	Uncertainty quantification

The education gap

Statistics

Astronomers usually take 0-1 courses in statistical methods during their university/graduate education, in contrast to extensive training in physics and associated mathematics. Students are thirsty for statistics training.

Computation

Survey of ~1100 astronomers worldwide (Momcheva & Tollerud 2015) shows that 90% write software but only 8% receive substantial training in software development and 43% receive no training.

Preferred languages are Python (67%), IDL (44%), C/C++ (37%), Fortran (28%), …, Matlab (8%), R (6%, 3% in USA), … julia (1%).

“Considering the wide-spread use of R in other scientific fields, its popularity among astronomers is strikingly low.”

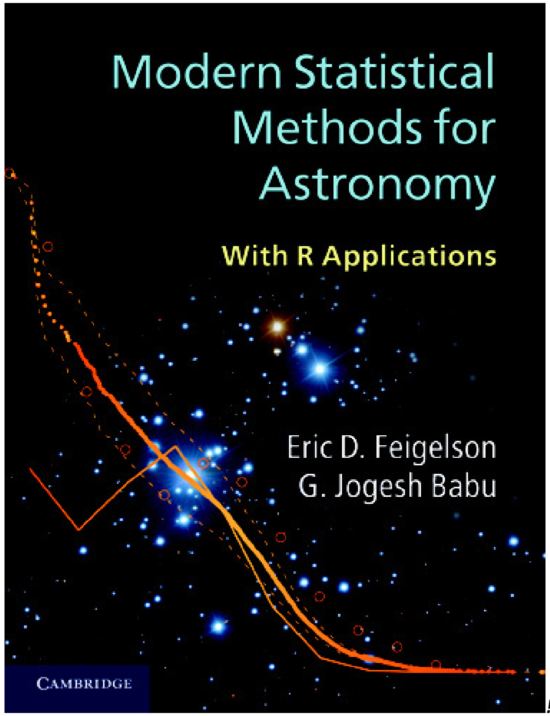
The result of weak training is
a widespread ignorance
of the myriad advances in modern
statistical & computational methodology.

Under-utilized methodology:

- modeling (MLE, EM Algorithm, BIC, bootstrap)
- multivariate classification (LDA, SVM, CART, RFs)
- time series (autoregressive models, state space models)
- spatial point processes (Ripley's K, kriging)
- nondetections (survival analysis)
- image analysis (computer vision methods, False Detection Rate)
- statistical computing (R)

Advertisement …

Modern Statistical Methods for Astronomy with R Applications (MSMA)
E. D. Feigelson & G. J. Babu,
Cambridge Univ Press, 2012



" #$$%&' ()' !*+, -. !/0 1&2!34&!
5%67!/67&4$48 9!: !; 468 4<4=9!544>!

MSMA: A curriculum in astrostatistics

Introduction

- Role & history of statistics in astronomy

Probability

- Axioms, Bayes Theorem, random variables, distributions, limit theorems

Statistical inference

- Astronomical context
- Point estimation: least squares, maximum likelihood
- Hypothesis testing
- Confidence intervals & bootstrap resampling
- Model selection & goodness-of-fit
- Bayesian inference

Probability distribution functions

- Binomial, multinomial, Poisson
- Normal, lognormal
- Pareto & gamma distributions

Nonparametric statistics

- Astronomical context & statistical concepts
- Univariate (KS & AD tests, …)
- Hypothesis tests (2-sample, …)
- Contingency tables
- Multivariate tests

Density estimation or data smoothing

- Astronomical context & statistical concepts
- Histograms & kernel density estimators
- Nonparametric local regression

Regression

- Astronomical context & statistical concepts
- Least squares (weighted, robust symmetric, quantile, MLE)
- Measurement error models
- Nonlinear models
- Model validation, selecition & misspecification

Multivariate analysis

- Astronomical context & statistical concepts
- Hypothesis tests
- Regression, principal components, outliers, nonlinear methods

Clustering and classification

- Astronomical context & statistical concepts
- Nonparametric clustering & mixture models
- Supervised classification (decision trees, k-NN, ANN, LDA, SVM)
- Classifier validation & improvement

Nondetections: censored & truncated data

- Astronomical context & statistical concepts
- Univariate data (Kaplan-Meier, 2-sample tests)
- Multivariate data (correlation, regression)
- Truncation (LBW)

Time series analysis

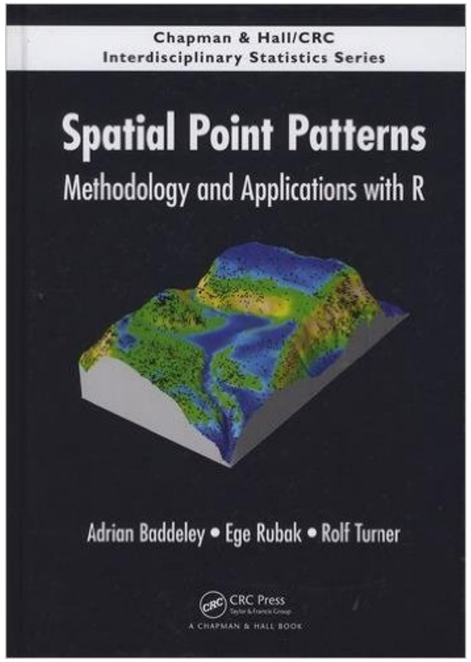
- Astronomical context & statistical concepts
- Time domain analysis (interpolation, ACF, ARMA, unevenly spaced data)
- Frequency domain analysis (Fourier, LS, improvement & significance)
- Nonstationary & long-memory (1/f) behaviors

Spatial point processes

- Astronomical context & statistical concepts
- Tests for uniformity, spatial autocorrelation & interpolation
- Global clustering, model-based analysis, tessellations
- Points on a circle or sphere

spatstat: A large CRAN package useful for astronomy

~1500 statistical functions
~800 book describing methods w/ recipes



Spatial Point Patterns:
Methodology and Applications with R
Adrian Baddeley et al. 2015

2D/3D, 1M pts for plotting, 100K exploratory, 10K modeling (on laptop)
Smoothing locations and mark variables
Global spatial autocorrelation functions (PCF, K, F, G, J, L*)
Tests for inhomogeneity & anisotropy
Clustering & hypothesis tests for multitype data
MLE and Bayesian modeling, model residuals & validation
Poisson, Gibbs, user-supplied models
Tessellations

The R statistical computing environment

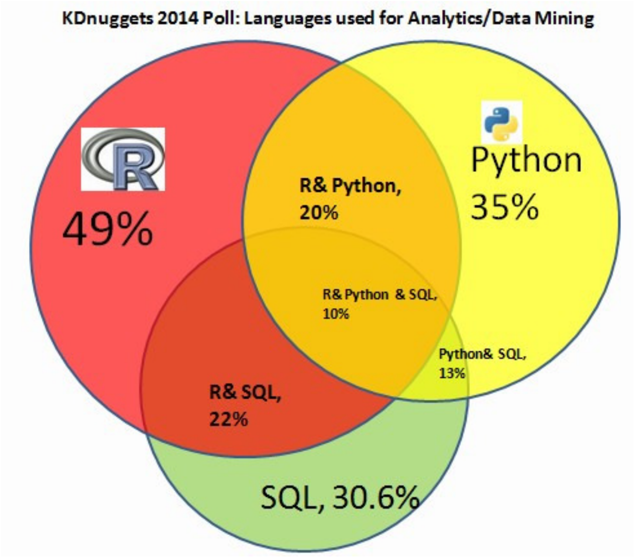
- R integrates data manipulation, graphics and extensive statistical analysis. Uniform documentation and coding standards.

- Fully programmable C-like language, similar to IDL. Specializes in vector/matrix inputs. Same compiler speed as IDL, Python, Matlab.

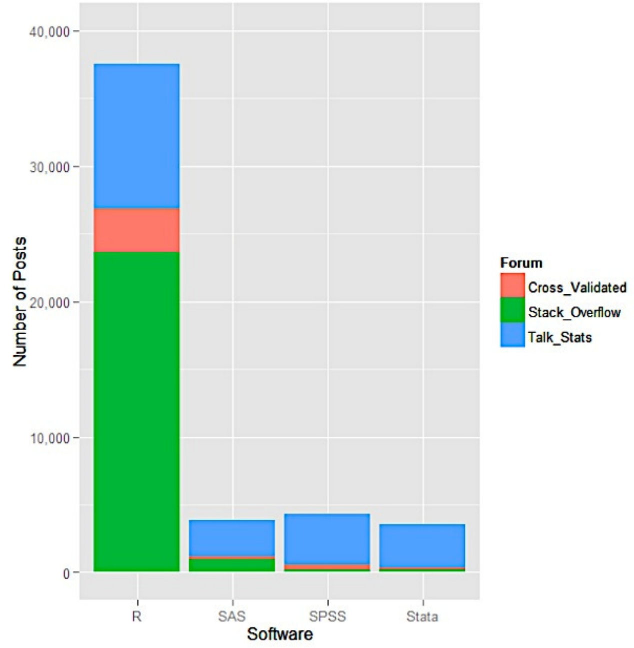
- Easy binary download for Windows, Mac or linux. On-the-fly installation of CRAN packages. Quick 2-way communication with C, Fortran, Python. Emulator of Matlab.

- 9000+ user-provided add-on **CRAN** packages including ~20 for astronomy (FITSio, astrolibR, …)

R's growing importance in data science



Kdnuggests 2014 poll data analytics/mining



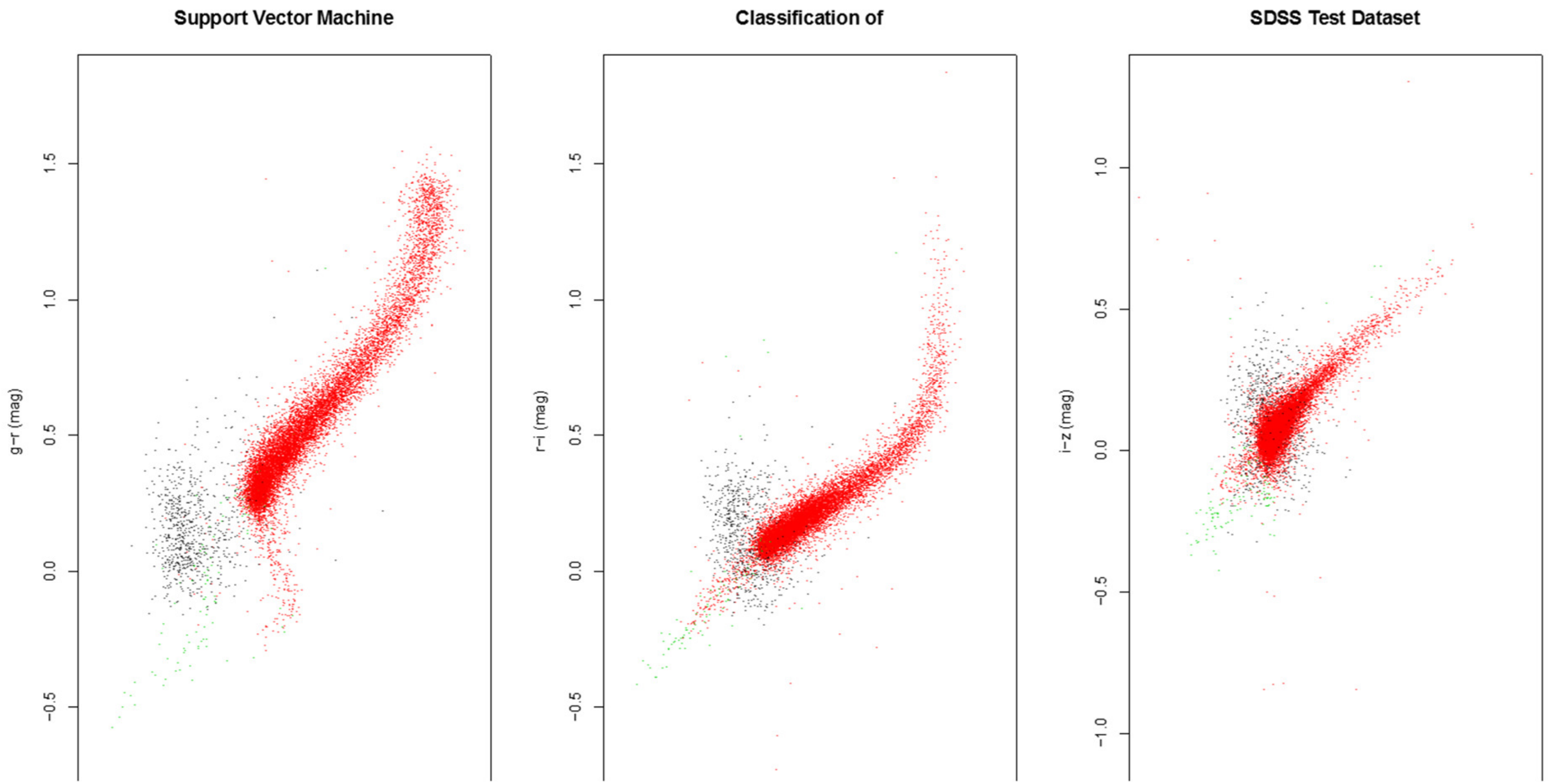
Posts on software forums 2013

Job trends from Indeed.com

Classification in R

 # Support Vector Machine model, prediction and validation

```
install.packages('e1071') ; library(e1071)
SDSS_svm_mod <- svm(SDSS_train[,5] ~.,data=SDSS_train[,1:4],cost=100, gamma=1)
summary(SDSS_svm_mod) ; str(SDSS_svm_mod)
SDSS_svm_test_pred <- predict(SDSS_svm_mod, SDSS_test)
plot(SDSS_test[,1], SDSS_test[,2], xlim=c(-0.7,3), col=round(SDSS_svm_test_pred),
 ylim=c(-0.7,1.8), pch=20, cex=0.5, main='Support Vector Machine', xlab='u-g (mag)’)
dev.copy2pdf(file='SDSS_SVM_class.pdf')
```



www2.astro.psu.edu/users/edf/Santiago_2016

Closing remarks

- Astronomy graduate curriculum needs two semesters for statistical and computational methodology. Textbook in astroinformatics needed. Some astronomers should get M.S. in statistics or computer science. All astronomers should know methodology at Wikipedia level.

- Astrostatistics and astroinformatics can become a well-funded, cross-disciplinary research field involving a few percent of astronomers (cf. astrochemists) promulgating modern methodology and pushing its frontiers.

- Astronomers can benefit from regular use of many methods coded in R & CRAN. R propels implementation of modern procedures with little effort. Experts should investigate utility of R/CRAN in HPC environments.