

Driving Reproducibility at UW eScience



Jake VanderPlas
*NYU Reproducibility
Symposium*
May 3rd, 2016

Outline

- What is Reproducibility?
- Why Reproducibility?
- Barriers to Reproducibility
- Reproducibility efforts at UW

Defining our Terms

“Replication is the ultimate standard by which scientific claims are judged.”



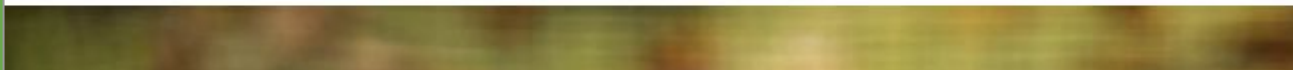
Sections 

The Washington Post

Monkey Cage

Does social science have a replication crisis?

By Joshua Tucker March 9 





Sections

Monkey Cage

Does so replicat

By Joshua Tucker Ma

Psychology's Replication Crisis Has a Silver Lining

It's an opportunity for the field to lead.

684



TEXT SIZE



PAUL BLOOM

FEB 19, 2016

SCIENCE

There is a crisis in psychology. It's not those rare cases of outright fraud, as when the social psychologist Diedrik Stapel simply made up the results of dozens of



Sections

Monkey Cage

Psychology's Replication Crisis

The replication crisis has engulfed economics

November 2, 2015 7:31pm EST

No two alike? Image sourced from Shutterstock.com

Email

Twitter

84

Facebook

68

LinkedIn

26

A sense of crisis is developing in economics after [two Federal Reserve economists](#) came to the alarming conclusion that economics research is usually not replicable.

The economists took 67 empirical papers from 13 reputable academic

t fraud, as when
of dozens of



Sections

Monkey Cage

The replication crisis has engulfed ec

November 2, 2015 7:31pm EST

No two alike? Image sourced from Shutterstock

Email

Twitter

Facebook

LinkedIn

84

68

26

A sense

econom

usually

The eco

Psychology's Replication Crisis

Cancer Research Is Broken

There's a replication crisis in biomedicine—and no one even knows how deep it runs.

By Daniel Engber



1.9k



604



84



as when

ens of



Sections



Big Science is broken



Pascal-Emmanuel Gobry



Facebook

68

usually

LinkedIn

26

The eco

as when

ens of





Sections



Big Science is broken



Pascal-Emmanuel Gobry

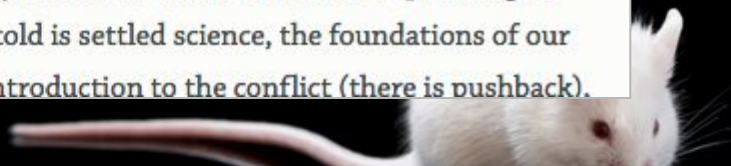
The replication crisis in science has just begun. It will be big.

[24 Replies](#)

Summary: After a decade of slow growth beneath public view, the replication crisis in science begins breaking into public view. First psychology and biomedical studies, now spreading to many other fields — overturning what we were told is settled science, the foundations of our personal behavior and public policy. Here is an introduction to the conflict (there is pushback).



as when
one of



Replication ~ *VS* ~ Reproducibility

"I define 'replication' as independent people going out and collecting new data and 'reproducibility' as independent people analyzing the same data."

Replicate:

Independently confirm results with
new data

Reproduce:

Independently confirm results with the
same data

Replicate:

Independently confirm results with
new data (and/or code)

Reproduce:

Independently confirm results with the
same data (and/or code)

In science . . .

“Replicable” = “Correct”

In science . . .

“Replicable” = “Correct”

* or about as close as we can hope to come to it . . . see Thomas Kuhn, etc.

In science . . .

“Replicable” = “Correct”

but . . .

**we have no control over the
replicability of our research!**

In science . . .

“Replicable” = “Correct”

but . . .

**we have no control over the
replicability of our research!**

* leaving aside low bars
like “*don’t commit fraud*”

In science . . .

“Reproducible” ≠ “Correct”

In science . . .

“Reproducible” ≠ “Correct”

* necessarily . . .

In science . . .

“Reproducible” ≠ “Correct”

but . . .

**we do have control over it
in our own research.**

“The standard of reproducibility
calls for the data and the computer
code used to analyze the data be
made available to others.”

[Side note: other definitions exist; e.g. Stodden (2013)]

- ***Reviewable Research***: sufficient detail for peer review & assessment.
- ***Replicable Research***: Tools are available to duplicate the author's results using their data.
- ***Confirmable Research***: Main conclusions can be attained independently without author's software.
- ***Auditable Research***: Process & tools archived such that it can be defended later if necessary.
- ***Open/Reproducible Research***: Auditable research made openly available.

Why Reproducibility?

The Philosophical Motive:

Replicability is the ultimate goal.

Reproducibility is the part we can control.

The Selfish Motive:

Reproducible work is Extensible work!

Step 1: make work reproducible

Step 2: get more citations

Step 3: ???

Step 4: Profit!!!

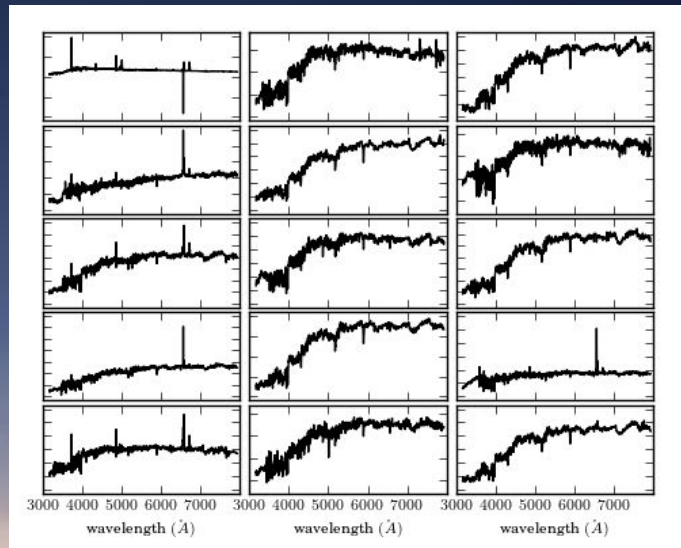
My Own Story . . .



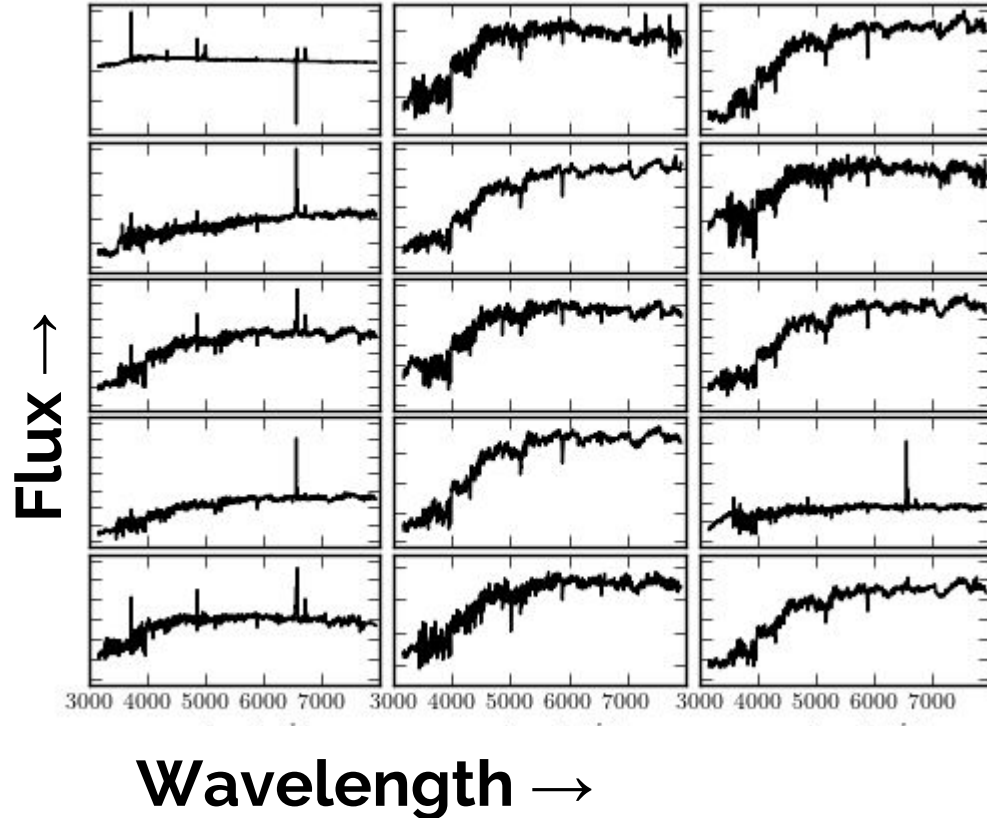
My Own Story . . .

The light from galaxies
reveals a *lot* about their
nature

The Sloan Digital Sky Survey

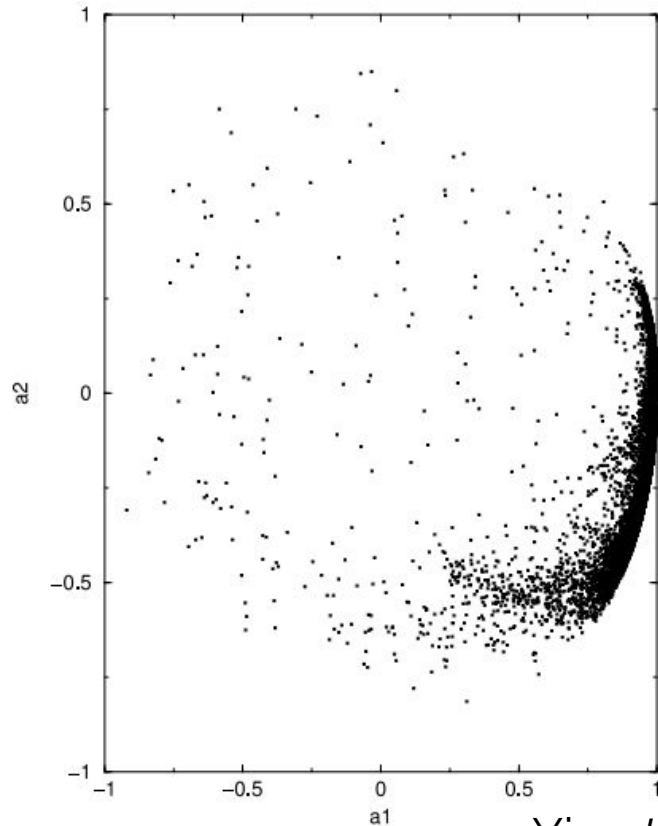
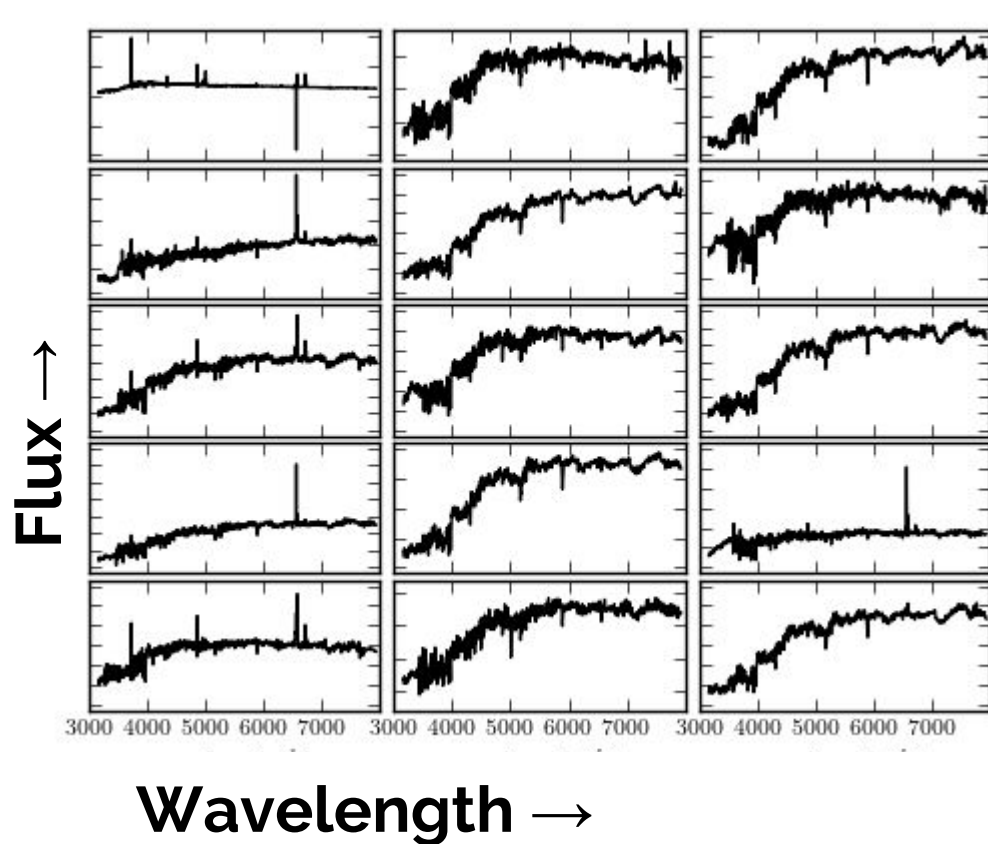


Galaxy Spectra



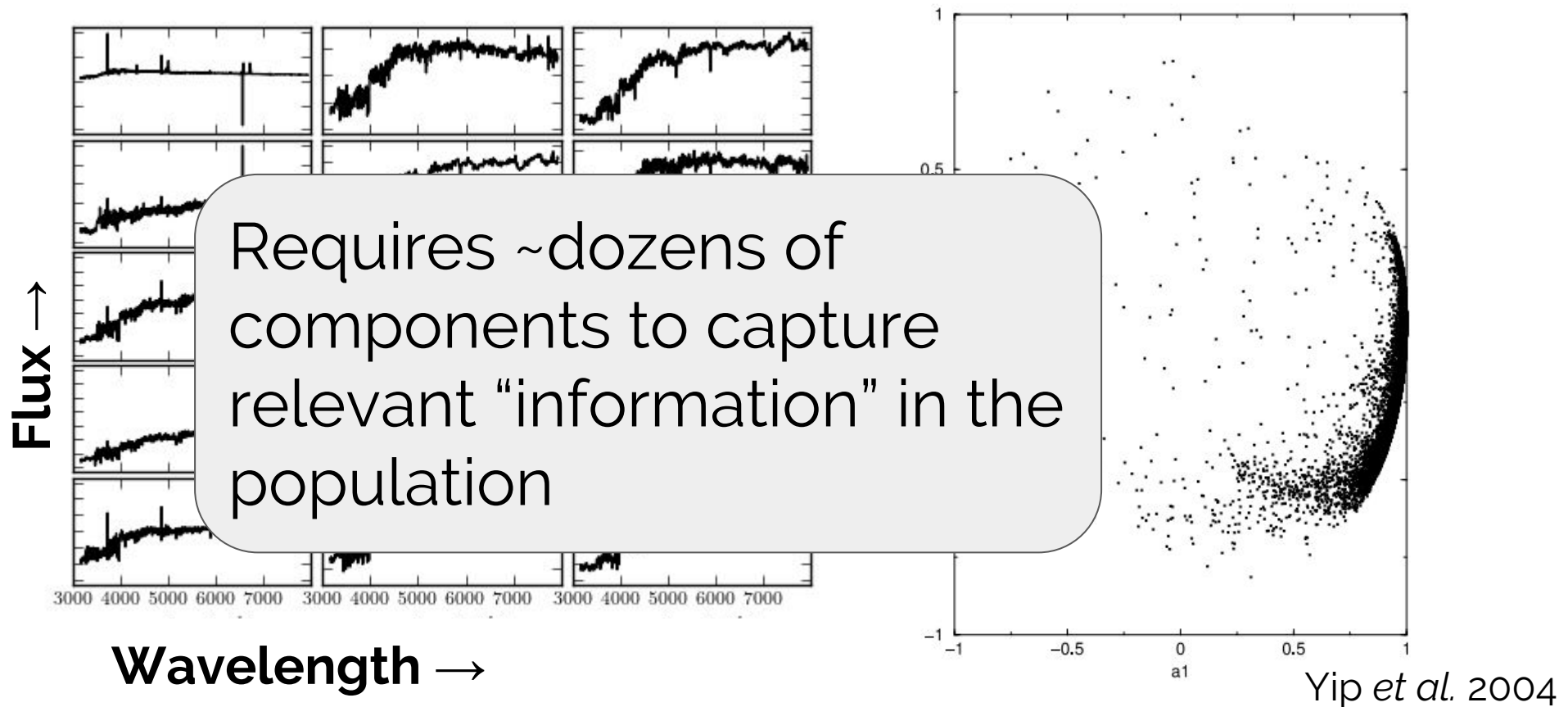
- 1000s of measurements per galaxy.
- Traces stellar population, activity, formation rate, dust content, etc.

Principal Component Analysis



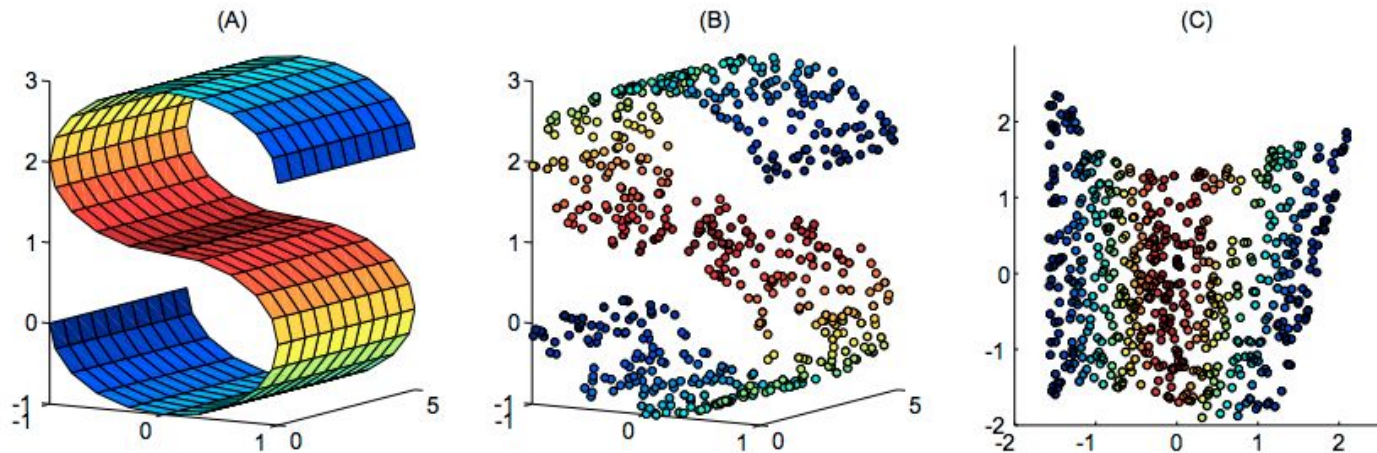
Yip *et al.* 2004

Principal Component Analysis



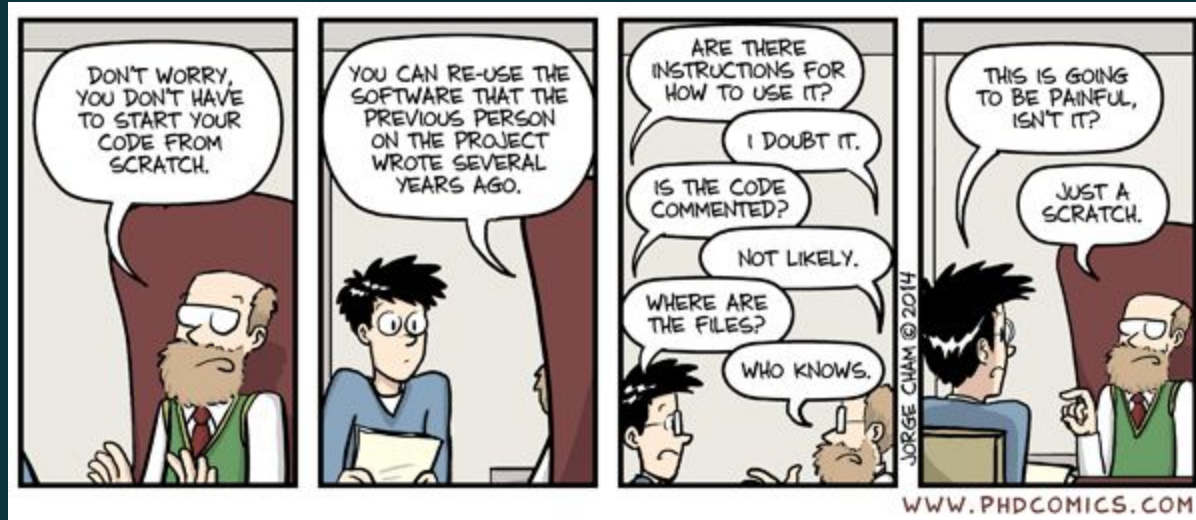
Our question:

Could a *nonlinear* projection more efficiently capture this information?



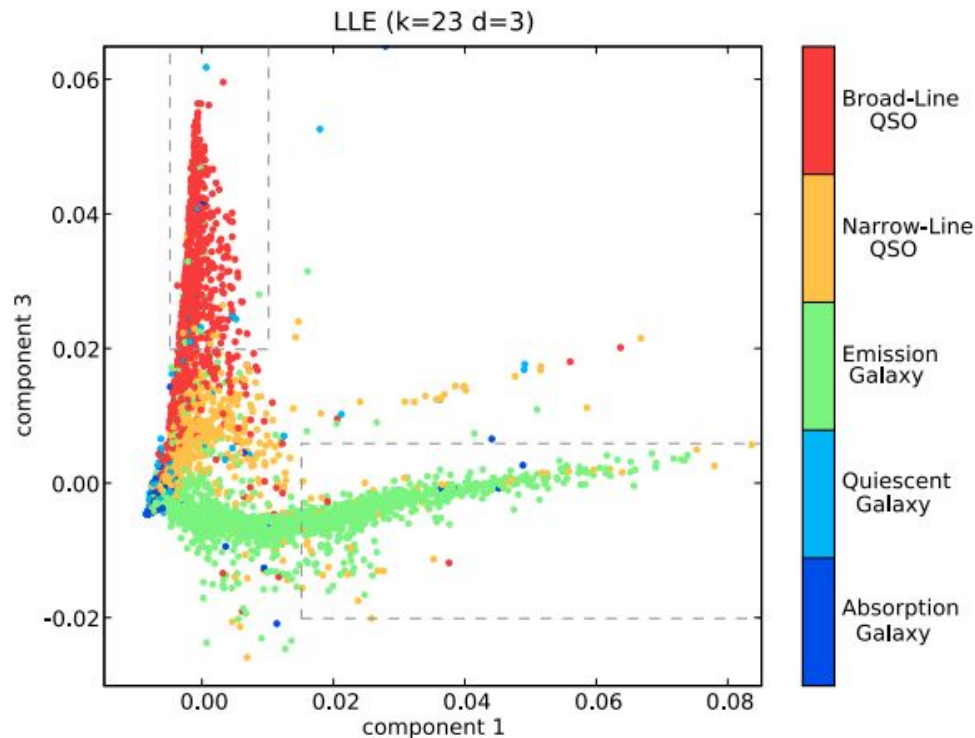
Locally Linear Embedding (Roweis & Saul 2000)

They even had a MatLab implementation!



After 4 months of pain . . .

Yes! LLE gives a much more compact embedding!



(VanderPlas & Connolly 2009)

I was left wondering . . .

Would the next poor grad student
have to re-implement LLE again?

“An article about computational results is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.”

Paraphrased from
[Claerbout and Karrenbach 1992](#);
[Buckheit & Donoho 1995](#)

I first did the tried-and-true
“Tarball on a Webpage” solution.



The screenshot shows the homepage of the SSG Software website. The header features a dark banner with the SSG logo on the left, which consists of the letters 'SSG' in a stylized, metallic font against a background of a starry space scene. To the right of the logo, the text 'University of Washington Astronomy' is in a purple serif font, and 'Survey Science Group: Research' is in a white serif font. Below the banner is a navigation bar with several buttons: 'Home', 'Research', 'LSST', 'eScience', 'People', 'Private', and 'AstroViz'. The main content area has a white background. It starts with the heading 'SSG Software' in a large, bold, black serif font. Below this is the subheading 'Locally Linear Embedding' in a smaller, italicized black serif font. Further down, it lists 'Author: Jake VanderPlas' and 'Download: DimReduce version 3.1 (tgz file)'. At the bottom, there is an 'About' section that describes DimReduce as a C++ package for performing nonlinear dimensionality reduction of very large datasets with Locally Linear Embedding (LLE) and its variants. The text is in a small black serif font.

SSG
University of Washington Astronomy
Survey Science Group: Research

Home Research LSST eScience People Private AstroViz

SSG Software

Locally Linear Embedding

Author: Jake VanderPlas

Download: DimReduce version 3.1 ([tgz file](#))

About: DimReduce is a C++ package for performing nonlinear dimensionality reduction of very large datasets with [Locally Linear Embedding \(LLE\)](#) and its variants. DimReduce

But I had the
nagging suspicion
it wasn't enough...

Then someone told me about
a (then) brand new project . . .



[Home](#) [Installation](#) [Documentation](#) [Examples](#)

Fork me on GitHub



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

... it turns out people like good tools!

Title	1–20	Cited by	Year
Scikit-learn: Machine learning in Python	F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, ... The Journal of Machine Learning Research 12, 2825-2830	2799	2011
Reducing the dimensionality of data: Locally linear embedding of sloan galaxy spectra	J VanderPlas, A Connolly The Astronomical Journal 138 (5), 1365	32	2009

(* not that I'm keeping track . . .)

Barriers to Reproducibility

The Incentive Problem – reproducibility takes time, and is not always valued by the academic reward structure.

The Pipeline Problem – reproducibility requires skills that are often not included in undergrad/grad curriculum!

Some “Incentive” Efforts at UW eScience

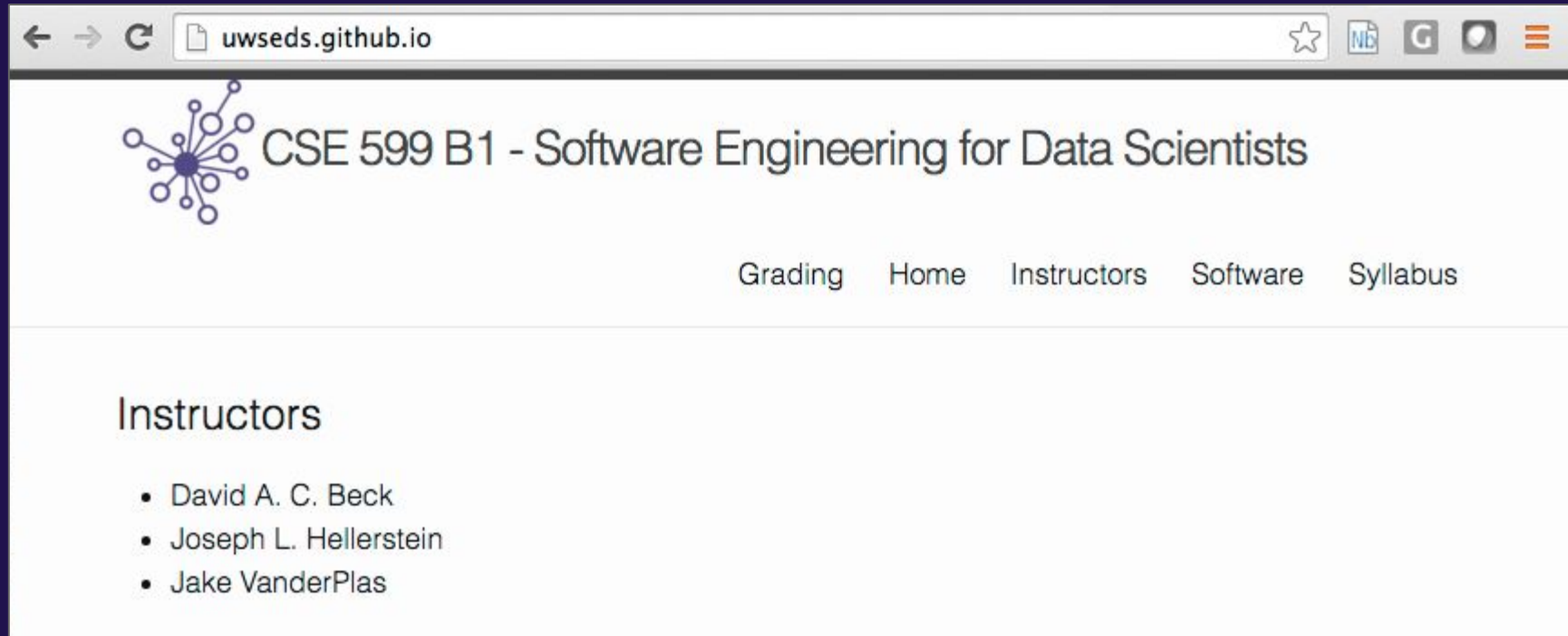
Addressing incentives is *hard*...

A few ideas in progress:

- Reproducibility & Open Science Badges
 - <https://github.com/uwescience/reproducible/issues/3>
- *Journal of Open Source Software* (soon!)
 - <http://joss.theoj.org/about>

Some “Pipeline” Efforts at UW eScience

Software Engineering for Data Science




Offered Winter 2016




Software Engineering for Data Science


- 30 grad students from dozens of departments
- Data cleaning & analysis, Version control, CI, unit testing, documentation, packaging . . .
- Every topic motivated by quarter-long example problem: open bicycle data.
- Final: apply these tools to a reproducible, collaborative group project.
- Teaching material available at <http://uwseds.github.io/>

Shablona: Python software template

 This repository

[Pull requests](#) [Issues](#) [Gist](#)

 [uwescience](#) / [shablona](#)


[Unwatch](#) 18 [Star](#) 217 [Fork](#) 41




[Code](#) [Issues 10](#) [Pull requests 6](#) [Wiki](#) [Pulse](#) [Graphs](#) [Settings](#)

A template for small scientific python projects — [Edit](#)

[79 commits](#) [1 branch](#) [0 releases](#) [4 contributors](#)

Branch: [master](#) [New pull request](#) [New file](#) [Upload files](#) [Find file](#) [HTTPS](#) <https://github.com/uwesci> [Download ZIP](#)

 [arokem](#) Merge pull request #26 from arokem/doc-reference-empty [...](#) Latest commit 48ffcba 12 days ago

 doc	Remove README file, that got clobbered every time you ran `make clean`	12 days ago
 scripts	Add content in the notebook.	11 months ago
 shablona	PEP8 fixes.	3 months ago

Research Incubator & Data Science for Social Good

Example – Fall 2014 Incubator: *Unlocking Kenya's Health Data*



Gregoire Lurton



Dan Halperin

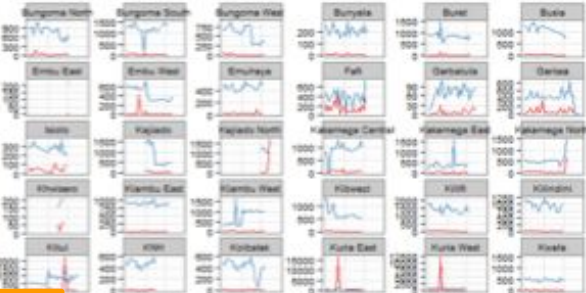


Abie Flaxman

"Much of the material remains unprocessed, or, if processed, unanalyzed, or, if analyzed, not read, or, if read, not used or acted upon"

Results

- Generalizable method to process HIS-like data
- Clean, integrate, and make available this data for research
- Preliminary analysis of HIS data – known spikes in malaria visible in the data (finally!)



Summer 2015

Data Science for Social Good



16 student interns matched with
6 data scientists working on data-oriented
social-good projects with **4 non-profit groups**

- *Assessing Community Well-Being Through Open Data*
- *King County Metro Paratransit*
- *Open Sidewalk Graph for Accessible Trip Planning*
- *Predictors of Permanent Housing for Homeless Families*



Leading by Example . . .

Leading by Example . . .

**Another case study:
Reproducibility in the Extreme Limit**

Coding in the open leads to clear benefits:

Eases Collaboration — Eases Sharing
Heightens Visibility — Encourages Reproducibility
Encourages Extensibility — Encourages Integrity
etc. — etc.

... why not do *all research* in the open?

Eases Collaboration — Eases Sharing
Heightens Visibility — Encourages Reproducibility
Encourages Extensibility — Encourages Integrity
etc. — etc.

“Extreme Openness”

We wrote this paper on GitHub *from day 1*

THE ASTROPHYSICAL JOURNAL, 812:18 (15pp), 2015 October 10

© 2015. The American Astronomical Society. All rights reserved.

doi:10.1088/0004-637X/812/1/18

PERIODOGRAMS FOR MULTIBAND ASTRONOMICAL TIME SERIES

JACOB T. VANDERPLAS¹ AND ŽELJKO IVEZIĆ²

¹ eScience Institute, University of Washington, Seattle, WA, USA

² Department of Astronomy, University of Washington, Seattle, WA, USA

Received 2015 February 5; accepted 2015 August 24; published 2015 October 6

ABSTRACT

This paper introduces the *multiband periodogram*, a general extension of the well-known Lomb–Scargle approach for detecting periodic signals in time-domain data. In addition to advantages of the Lomb–Scargle method such as treatment of non-uniform sampling and heteroscedastic errors, the multiband periodogram significantly improves period finding for randomly sampled multiband light curves (e.g., Pan-STARRS, DES, and LSST). The light curves in each band are modeled as arbitrary truncated Fourier series, with the period and phase shared across all bands. The key aspect is the use of Tikhonov regularization which drives most of the variability into the so-called base model common to all bands, while fits for individual bands describe residuals relative to the base model and typically require lower-order Fourier series. This decrease in the effective model complexity is the main reason for improved performance. After a pedagogical development of the formalism of least-squares spectral analysis, which motivates the essential features of the multiband model, we use simulated light curves and randomly subsampled SDSS Stripe 82 data to demonstrate the superiority of this method compared to other methods from the literature and find that this method will be able to efficiently determine the correct period in the majority of LSST's bright RR Lyrae stars with as little as six months of LSST data, a vast improvement over the years of data reported to be required by previous studies. A Python implementation of this method, along with code to fully reproduce the results reported here, is available on GitHub.

Key words: methods: data analysis – methods: statistical – surveys

https://github.com/jakevdp/multiband_LS/

Mutiband Lomb-Scargle Periodograms

This repository contains the source for our multiband periodogram paper. It makes use of the [gatspy](#) package, which has been developed concurrently. The paper has been submitted to the Astrophysical Journal, and a preprint is available on [arXiv](#). To see a current build of the paper from the master branch of this repository, refer to http://jakevdp.github.io/multiband_LS (powered by [gh-publisher](#)).

Feel free to submit comments or feedback via the Issues tab on this repository.

Reproducing the Paper

The LaTeX source of the paper, including all figure pdfs, is in the `writeup` directory. The code to reproduce the analysis and figures in the paper is in the `figures` directory.

To reproduce the figures, first install the following packages (Python 2 or 3):

- Standard Python scientific stack: ([IPython](#), [numpy](#), [scipy](#), [matplotlib](#), [scikit-learn](#), [pandas](#))
- [seaborn](#) for plot styles.
- [astroML](#) for general astronomy machine learning tools.
- [gatspy](#) for astronomical time-series analysis.
- [supersmoother](#) for the supersmoother algorithm used by [gatspy](#).

README.md

Mutiband Lomb-Scargle Periodograms

This repository contains the source for our multiband periodogram paper. It makes use of the [gatspy](#) package, which has been developed concurrently. The paper has been submitted to the Astrophysical Journal, and a preprint is available on [arXiv](#). To see a current build of the paper from the master branch of this repository, refer to http://jakevdp.github.io/multiband_LS powered by [gh-publisher](#).

Feel free to submit comments or feedback via the Issues tab on this repository.

Reproducing the Paper

The LaTeX source of the paper, including all figure pdfs, is in the `writeup` directory. The code to reproduce the analysis and figures in the paper is in the `figures` directory.

To reproduce the figures, first install the following packages (Python 2 or 3):

- Standard Python scientific stack: ([IPython](#), [numpy](#), [scipy](#), [matplotlib](#), [scikit-learn](#), [pandas](#))
- [seaborn](#) for plot styles.
- [astroML](#) for general astronomy machine learning tools.
- [gatspy](#) for astronomical time-series analysis.
- [supersmoother](#) for the supersmoother algorithm used by [gatspy](#).

Periodograms for Multiband Astronomical Time Series

Jake VanderPlas

eScience Institute, University of Washington

Zeljko Ivezić

Astronomy department, University of Washington

This site displays an automated build of the LaTeX source of the paper, which is hosted on GitHub. The arXiv url is <http://arxiv.org/abs/1502.01344>

View the Project on GitHub
jakevdp/multiband_LS

Send feedback

DRAFT VERSION NOVEMBER 2, 2015
Preprint typeset using L^AT_EX style emulateapj v. 01/23/15

PERIODOGRAMS FOR MULTIBAND ASTRONOMICAL TIME SERIES

JACOB T. VANDERPLAS¹ AND ŽELJKO IVEŽIĆ²
Draft version November 2, 2015

ABSTRACT

This paper introduces the *multiband periodogram*, a general extension of the well-known Lomb-Scargle approach for detecting periodic signals in time-domain data. In addition to advantages of the Lomb-Scargle method such as treatment of non-uniform sampling and heteroscedastic errors, the multiband periodogram significantly improves period finding for randomly sampled multiband light curves (e.g., Pan-STARRS, DES and LSST). The light curves in each band are modeled as arbitrary truncated Fourier series, with the period and phase shared across all bands. The key aspect is the use of Tikhonov regularization which drives most of the variability into the so-called base model common to all bands, while fits for individual bands describe residuals relative to the base model and typically require lower-order Fourier series. This decrease in the effective model complexity is the main reason for improved performance. After a pedagogical development of the formalism of least-squares spectral analysis which motivates the essential features of the multiband model, we use simulated light curves and randomly subsampled SDSS Stripe 82 data to demonstrate the superiority of this method compared to other methods from the literature, and find that this method will be able to efficiently determine the correct period in the majority of LSST's bright RR Lyrae stars with as little as six months of LSST data, a vast improvement over the years of data reported to be required by previous studies. A Python implementation of this method, along with code to fully reproduce the results reported here, is available on GitHub.

Subject headings: methods: data analysis — methods: statistical

1. INTRODUCTION

Many types of variable stars show periodic flux variability (Eyer & Mowlavi 2008). Periodic variable stars are important both for testing models of stellar evolution and for using such stars as distance indicators (e.g., Cepheids and RR Lyrae stars). One of the first and main goals of the analysis is to detect variability and to estimate the period and its uncertainty. A number of parametric and non-parametric methods have been proposed to estimate the period of an astronomical time series (e.g., Graham et al. 2013, and references therein).

The most popular non-parametric method is the phase dispersion minimization (PDM) introduced by Stellingwerf (1978). Dispersion per bin is computed for binned phased light curves evaluated for a grid of trial periods. The best period minimizes the dispersion per bin. A similar and related non-parametric method that has been recently gaining popularity is the Supersmoother routine (Reinman 1994). It uses a running mean or running linear regression on the data to fit the observations as a function of phase to a range of periods. The best period minimizes a figure-of-merit, adopted as weighted sum of absolute residuals around the running mean. Neither the Supersmoother algorithm nor the PDM method require a priori knowledge of the light curve shape.

The most popular parametric method is the Lomb-Scargle periodogram, which is discussed in detail in Section 2. The Lomb-Scargle periodogram is related to the

ing model of the Lomb-Scargle periodogram is nonlinear in frequency and so the likelihood surface in frequency is non-convex. This non-convexity is readily apparent in the many local maxima of the typical periodogram, which makes it difficult to find the maximum via standard numerical optimization routines. Thus in practice the global maximum of the periodogram is often found by a brute-force grid search (for details see, e.g. Ivezić et al. 2014).

A more general parametric method based on the use of continuous-time autoregressive moving average (CARMA) model was recently introduced by Kelly et al. (2014). CARMA models can also treat non-uniformly sampled time series with heteroscedastic measurement uncertainties, and can handle complex variability patterns.

A weakness of all these standard methods is that they require homogeneous measurements – for astronomy data, this means that successive measurements must be taken through a single photometric bandpass (filter). This has not been a major problem for past surveys because measurements are generally taken through a single photometric filter (e.g. LINEAR, Sesar et al. 2011), or nearly-simultaneously in all bands at each observation (e.g. SDSS, Sesar et al. 2010). For the case of simultaneously taken multiband measurements, Sivigaa et al. (2012) utilized the principal component method to optimally extract the best period. Their method is essen-

gh-publisher

Each new commit triggers a Travis CI process which builds the current paper PDF and pushes it to this website...

<https://github.com/ewanmellor/gh-publisher/>

Periodograms for Multiband Astronomical Time Series

Jake VanderPlas

eScience Institute, University of Washington

Zeljko Ivezić

Astronomy department, University of
Washington

This site displays an automated build of the
LaTeX source of the paper, which is hosted on
GitHub. The arXiv url is
<http://arxiv.org/abs/1502.01344>

View the Project on GitHub
[jakevdp/multiband_LS](https://github.com/jakevdp/multiband_LS)

Send feedback

DRAFT VERSION NOVEMBER 2, 2015
Preprint typeset using L^AT_EX style emulateapj v. 01/23/15

PERIODOGRAMS FOR MULTIBAND ASTRONOMICAL TIME SERIES

JACOB T. VANDERPLAS¹ AND ŽELJKO IVEZIĆ²
Draft version November 2, 2015

ABSTRACT

This paper introduces the *multiband periodogram*, a general extension of the well-known Lomb-Scargle approach for detecting periodic signals in time-domain data. In addition to advantages of the Lomb-Scargle method such as treatment of non-uniform sampling and heteroscedastic errors, the multiband periodogram significantly improves period finding for randomly sampled multiband light curves (e.g., Pan-STARRS, DES and LSST). The light curves in each band are modeled as arbitrary truncated Fourier series, with the period and phase shared across all bands. The key aspect is the use of Tikhonov regularization which drives most of the variability into the so-called base model common to all bands, while fits for individual bands describe residuals relative to the base model and typically require lower-order Fourier series. This decrease in the effective model complexity is the main reason for improved performance. After a pedagogical development of the formalism of least-squares spectral analysis which motivates the essential features of the multiband model, we use simulated light curves and randomly subsampled SDSS Stripe 82 data to demonstrate the superiority of this method compared to other methods from the literature, and find that this method will be able to efficiently determine the correct period in the majority of LSST's bright RR Lyrae stars with as little as six months of LSST data, a vast improvement over the years of data reported to be required by previous studies. A Python implementation of this method, along with code to fully reproduce the results reported here, is available on GitHub.

Subject headings: methods: data analysis — methods: statistical

1. INTRODUCTION

Many types of variable stars show periodic flux variability (Eyer & Mowlavi 2008). Periodic variable stars are important both for testing models of stellar evolution and for using such stars as distance indicators (e.g., Cepheids and RR Lyrae stars). One of the first and main goals of the analysis is to detect variability and to estimate the period and its uncertainty. A number of parametric and non-parametric methods have been proposed to estimate the period of an astronomical time series (e.g., Graham et al. 2013, and references therein).

The most popular non-parametric method is the phase dispersion minimization (PDM) introduced by Stellingwerf (1978). Dispersion per bin is computed for binned phased light curves evaluated for a grid of trial periods. The best period minimizes the dispersion per bin. A similar and related non-parametric method that has been recently gaining popularity is the Supersmoother routine (Reinman 1994). It uses a running mean or running linear regression on the data to fit the observations as a function of phase to a range of periods. The best period minimizes a figure-of-merit, adopted as weighted sum of absolute residuals around the running mean. Neither the Supersmoother algorithm nor the PDM method require a priori knowledge of the light curve shape.

The most popular parametric method is the Lomb-Scargle periodogram, which is discussed in detail in Section 2. The Lomb-Scargle periodogram is related to the

ing model of the Lomb-Scargle periodogram is nonlinear in frequency and so the likelihood surface in frequency is non-convex. This non-convexity is readily apparent in the many local maxima of the typical periodogram, which makes it difficult to find the maximum via standard numerical optimization routines. Thus in practice the global maximum of the periodogram is often found by a brute-force grid search (for details see, e.g. Ivezić et al. 2014).

A more general parametric method based on the use of continuous-time autoregressive moving average (CARMA) model was recently introduced by Kelly et al. (2014). CARMA models can also treat non-uniformly sampled time series with heteroscedastic measurement uncertainties, and can handle complex variability patterns.

A weakness of all these standard methods is that they require homogeneous measurements – for astronomy data, this means that successive measurements must be taken through a single photometric bandpass (filter). This has not been a major problem for past surveys because measurements are generally taken through a single photometric filter (e.g. LINEAR, Sesar et al. 2011), or nearly-simultaneously in all bands at each observation (e.g. SDSS, Sesar et al. 2010). For the case of simultaneously taken multiband measurements, Sivigaa et al. (2012) utilized the principal component method to optimally extract the best period. Their method is essen-



This repository Search

Pull requests Issues Gist



jakevdp / multiband_LS

Unwatch 6

Star 11

Fork 2

Code

Issues 1

Pull requests 0

Wiki

Pulse

Graphs

Settings

Filters

is:issue is:open

Labels

Milestones

New Issue

1 Open 2 Closed

Author

Labels

Milestones

Assignee

Sort

Scope of Jaynes analysis/Bayesian formalism

#1 opened on Feb 11, 2015 by jscargle

4

ProTip! Adding `no:label` will show everything without a label.



This repository Search

Pull requests Issues Gist



jakevdp / multiband_LS

Unwatch 6

Star 11

Fork 2

Code

Issues 1

Pull requests 0

Wiki

Pulse

Graphs

Settings

Filters

is:issue is:open

Labels

Milestones

New Issue

1 Open 2 Closed

Author

Labels

Milestones

Assignee

Sort

Scope of Jaynes analysis/Bayesian formalism

4

#1 opened on Feb 11, 2015 by jscargile

ProTip! Adding `no:label` will show everything without a label.

Scope of Jaynes analysis/Bayesian formalism #1



jscargle opened this issue on Feb 11, 2015 · 4 comments



jscargle commented on Feb 11, 2015



Your reference to Jaynes (1987) implies that he discussed the Lomb-Scargle periodogram. I don't think he knew about this particular algorithm, but was remarking more generally on least-squares fits to sinusoids.

Also, I wonder if you considered and rejected a fully Bayesian formalism for the problem you address in this paper? In <http://arxiv.org/abs/math/0111127> I tried to demonstrate a generic link between frequentist statistics such as the periodogram and the Bayesian posterior for a corresponding quantity (in the spirit of Larry Bretthorst's power spectrum analysis: Bretthorst, G. Larry, 1988, Bayesian Spectrum Analysis and Parameter Estimation, Lecture Notes in Statistics, Springer-Verlag, No. 48; <http://bayes.wustl.edu/>). I don't know how universal it is, but a connection of the form

$$P(a) \sim \exp[-C(a) / \text{error variance}]$$

seems to pop up a lot. $C(a)$ can be an auto- or cross-correlation function; or a power spectrum or a cross-power spectrum, corresponding to a being a time lag or the frequency of a sinusoidal component. Then the expedient of simply multiplying posteriors (for independent quantities) might be useful in the multi band context. If nothing else this might lead to a more rigorous "Occam factor" regularization.



jakevdp commented on Feb 12, 2015

Owner



Thanks @jscargle - yes, the sentence about the Jaynes paper is a bit misleading. We'll correct that

Scope of Jaynes analysis/Bayesian formalism #1



jscargle opened this issue on Feb 11, 2015 · 4 comments

Jeff Scargle!!!
— as in “Lomb-Scargle
Periodogram”!!!!!!



jscargle commented on Feb 11, 2015

Your reference to Jaynes (1987) implies that he discussed the Lomb-Scargle periodogram. I don't think he knew about this particular algorithm, but was remarking more generally on least-squares fits to sinusoids.

Also, I wonder if you considered and rejected a fully Bayesian formalism for the problem you address in this paper? In <http://arxiv.org/abs/math/0111127> I tried to demonstrate a generic link between frequentist statistics such as the periodogram and the Bayesian posterior for a corresponding quantity (in the spirit of Larry Bretthorst's power spectrum analysis: Bretthorst, G. Larry, 1988, Bayesian Spectrum Analysis and Parameter Estimation, Lecture Notes in Statistics, Springer-Verlag, No. 48; <http://bayes.wustl.edu/>). I don't know how universal it is, but a connection of the form

$$P(a) \sim \exp[-C(a) / \text{error variance}]$$

seems to pop up a lot. $C(a)$ can be an auto- or cross-correlation function; or a power spectrum or a cross-power spectrum, corresponding to a being a time lag or the frequency of a sinusoidal component. Then the expedient of simply multiplying posteriors (for independent quantities) might be useful in the multi band context. If nothing else this might lead to a more rigorous "Occam factor" regularization.



jakevdp commented on Feb 12, 2015

Owner



Thanks @jakevdp — yes, the sentence about the Jaynes paper is a bit misleading. We'll correct that

and expect to simply multiplying posterior (or independent quantities) might be useful in the multi-band context. If nothing else this might lead to a more rigorous "Occam factor" regularization.



jakevdp commented on Feb 12, 2015

Owner



Thanks [@jscargle](#) – yes, the sentence about the Jaynes paper is a bit misleading. We'll correct that.

I'd thought briefly about the Bayesian point-of-view for the multiband method, but I avoided discussing it for a couple reasons:

1. As your paper makes clear, the Bayesian and frequentist results for periodograms are usually related by a simple monotonic function, so going from one to the other is easy.
2. I think a fully-Bayesian treatment of the periodogram would involve marginalization over nuisance parameters in the model, and in some cases you want the nuisance parameter to be the period! Unfortunately, since the posterior is so highly multi-modal with varying period, I don't know of any Bayesian approach which can properly handle the problem in higher than a couple dimensions (perhaps nested sampling, but that's still a long-shot). I thought that rather than doing a half-Bayesian job which amounts to not much more than proposing some improper priors that lead to the exponent of the frequentist result, I'd stick to classical statistics here.

That said, a Bayesian approach to the multiband sinusoid-fitting problem would likely end up with a couple features:

- Priors would probably take the form of some constraint on the deviation of amplitudes and/or phases within each band. Mathematically, the result would be equivalent to a nonlinear regularization of the model, and likely be similar in spirit to the ad hoc nonlinear regularization proposed in Long (2014).

more satisfying than the statistically motivated (but admittedly still ad hoc) regularization we use.

For a truly Bayesian approach to period finding, I think a CARMA or similar model is a much better avenue (see e.g. [Kelly et al 2014](#))



jscargle commented on Feb 16, 2015



Good comments.

In the restricted context of a single sinusoid don't you think Larry Bretthorst's treatment is pretty compelling? Sure, you have to choose what parameters you take as "nuisance" and there are many possibilities, but Larry treats the most useful case(s) IMHO.

Yes, in this context Bayes and frequentist are related one-to-one in a monotonic fashion. But, if only cosmetically, the exponentiation can emphasize the "correct" mode at the expense of other (smaller) local maxima. More important if you have a good handle on the observational errors the Bayesian expression gives the full posterior probability (as opposed to a single global maximum).

You make excellent points regarding uncertainty in priors making the marginalized distribution ad hoc in much the same way as regularizations. (This is especially clear for regularization via the approximate BIC - Bayesian Information Criterion. I wonder what the count of regularization schemes is? AIC, BIC, MDL, ...) Alas, well-justified priors are rare; but when then are available the so-called Occam factor in the marginal posterior avoids ad hocery.



jscargle commented on Feb 16, 2015



on a different topic. In the discussion of Fig. 3 you state:



jscargle commented on Feb 16, 2015



... on a different topic. In the discussion of Fig. 3 you state:

"Second, notice that as the number of terms is increased, the general \background" level of the periodogram increases. This is due to the fact that the periodogram power is inversely related to the χ^2 at each frequency. A more flexible higher-order model can better fit the data at all periods, not just the true period. Thus in general the observed power of a higher-order Fourier model will be everywhere higher than the power of a lower-order Fourier model."

This seems counterintuitive to me. Adding harmonic components to the model in the manner of eq. (14) makes the frequency, ω , represent both ω itself and its harmonics $n\omega$, $n > 1$. You can see this in the right-hand panels of Fig.3: as n goes from 1 to 2 to 3, the peak power at the true fundamental P_0 increases -- power from the harmonics is incorporated via eq. (14) into the fundamental. How can one interpret the rise (from total insignificance) of the power at the first harmonic, $2P_0$? And why doesn't a better-fitting higher order model move power from the background continuum into the harmonics? -- the reverse of what you state. I am sure you are correct, but I guess I don't understand your "inversely related to ..." argument.



jakevdp commented on Feb 16, 2015

Owner



Jeff – thanks for all the comments! I really appreciate your close read of the paper and taking the time to discuss.

A couple responses:

I think you're right that Bretthorst's treatment is compelling in terms of fitting a single sinusoid to data. But the problem is that $P(\omega | \text{data}, M)$ with $M = (\text{a single sinusoid fits our data})$ is not exactly what we're

SCIENCE!



“But I might get scooped!”

Merriam-Webster SINCE 1828 MENU Dictionary publish

Full Definition of PUBLISH

transitive verb

- a** : to make generally known
b : to make **public** announcement of
- a** : to disseminate to the public
b : to produce or release for distribution; *specifically* : **PRINT** 2c
c : to issue the work of (an author)

f
t
g+
♥
CITE

. . . putting work on Github
is publication!

Merriam-Webster SINCE 1828 MENU Dictionary publish

Full Definition of PUBLISH

transitive verb

- a** : to make generally known
b : to make **public** announcement of
- a** : to disseminate to the public
b : to produce or release for distribution; *specifically* : **PRINT** 2c
c : to issue the work of (an author)

f
t
g+
♥
CITE

... putting work on Github
is publication!
“~~Scooping~~” → “Plagiarism”

**“But Jake. . . would you do
this if there were any *real*
competition?”**

[Periodograms for Multiband Astronomical Time Series](#)

[JT VanderPlas](#), [Ž Ivezić](#) - [The Astrophysical Journal](#), 2015 - [iopscience.iop.org](#)

Abstract This paper introduces the multiband periodogram, a general extension of the well-known Lomb–Scargle approach for detecting periodic signals in time-domain data. In addition to advantages of the Lomb–Scargle method such as treatment of non-uniform ...

Cited by 6 Related articles All 5 versions Web of Science: 2 Cite Save

[Estimating a Common Period for a Set of Irregularly Sampled Functions with Applications to Periodic Variable Star Data](#)

[JP Long](#), [EC Chi](#), [RG Baraniuk](#) - [arXiv preprint arXiv:1412.6520](#), 2014 - [arxiv.org](#)

Abstract: We consider the estimation of a common period for a set of functions sampled at irregular intervals. The problem arises in astronomy, where the functions represent a star's brightness observed over time through different photometric filters. While current methods ...

Cited by 3 Related articles All 3 versions Cite Save

[A Multiband Generalization of the Analysis of Variance Period Estimation Algorithm and the Effect of Inter-band Observing Cadence on Period Recovery Rate](#)

[N Mondrik](#), [JP Long](#), [JL Marshall](#) - [arXiv preprint arXiv:1508.04772](#), 2015 - [arxiv.org](#)

Abstract: We present a new method of extending the single band Analysis of Variance period estimation algorithm to multiple bands. We use SDSS Stripe 82 RR Lyrae to show that in the case of low number of observations per band and non-simultaneous ...

Cited by 1 Related articles Cite Save

Three *very similar* approaches all published within a few months last summer ...

So...

how did it turn out?

[Periodograms for Multiband Astronomical Time Series](#)

[JT VanderPlas](#), [Ž Ivezić](#) - [The Astrophysical Journal](#), 2015 - [iopscience.iop.org](#)

Abstract This paper introduces the multiband periodogram, a general extension of the well-known Lomb–Scargle approach for detecting periodic signals in time-domain data. In addition to advantages of the Lomb–Scargle method such as treatment of non-uniform ...

[Cited by 6](#) [Related articles](#) [All 5 versions](#) [Web of Science: 2](#) [Cite](#) [Save](#)

[Estimating a Common Period for a Set of Irregularly Sampled Functions with Applications to Periodic Variable Star Data](#)

[JP Long](#), [EC Chi](#), [RG Baraniuk](#) - [arXiv preprint arXiv:1412.6520](#), 2014 - [arxiv.org](#)

Abstract: We consider the estimation of a common period for a set of functions sampled at irregular intervals. The problem arises in astronomy, where the functions represent a star's brightness observed over time through different photometric filters. While current methods ...

[Cited by 3](#) [Related articles](#) [All 3 versions](#) [Cite](#) [Save](#)

[A Multiband Generalization of the Analysis of Variance Period Estimation Algorithm and the Effect of Inter-band Observing Cadence on Period Recovery Rate](#)

[N Mondrik](#), [JP Long](#), [JL Marshall](#) - [arXiv preprint arXiv:1508.04772](#), 2015 - [arxiv.org](#)

Abstract: We present a new method of extending the single band Analysis of Variance period estimation algorithm to multiple bands. We use SDSS Stripe 82 RR Lyrae to show that in the case of low number of observations per band and non-simultaneous ...

[Cited by 1](#) [Related articles](#) [Cite](#) [Save](#)

[Periodograms for Multiband Astronomical Time Series](#)

[JT VanderPlas](#), [Ž Ivezić](#) - [The Astrophysical Journal](#), 2015 - [iopscience.iop.org](#)

Abstract This paper introduces the multiband periodogram, a general extension of the well-known Lomb–Scargle approach for detecting periodic signals in time-domain data. In addition to advantages of the Lomb–Scargle method such as treatment of non-uniform ...

[Cited by 6](#) [Related articles](#) [All 5 versions](#) [Web of Science: 2](#) [Cite](#) [Save](#)

[Estimating a Common Period for a Set of Irregularly Sampled Functions with Applications to Periodic Variable Star Data](#)

[JP Long](#), [EC Chi](#), [RG Baraniuk](#) - [arXiv preprint arXiv:1412.6520](#), 2014 - [arxiv.org](#)

Abstract: We consider the estimation of a common period for a set of functions sampled at irregular intervals. The problem arises in astronomy, where the functions represent a star's brightness observed over time through different photometric filters. While current methods ...

[Cited by 3](#) [Related articles](#) [All 3 versions](#) [Cite](#) [Save](#)

[A Multiband Generalization of the Analysis of Variance Period Estimation Algorithm and the Effect of Inter-band Observing Cadence on Period Recovery Rate](#)

[N Mondrik](#), [JP Long](#), [JL Marshall](#) - [arXiv preprint arXiv:1508.04772](#), 2015 - [arxiv.org](#)

Abstract: We present a new method of extending the single band Analysis of Variance period estimation algorithm to multiple bands. We use SDSS Stripe 82 RR Lyrae to show that in the case of low number of observations per band and non-simultaneous ...

[Cited by 1](#) [Related articles](#) [Cite](#) [Save](#)

Periodograms for Multiband Astronomical Time Series

[JT VanderPlas](#), [Ž Ivezić](#) - [The Astrophysical Journal](#), 2015 - [iopscience.iop.org](#)

Abstract This paper introduces the multiband periodogram, a general extension of the well-known Lomb-Scargle approach for detecting periodic signals in time-domain data. In addition to the Scargle method such as treatment of non-uniform sampling, the multiband periodogram includes the following extensions: Web of Science: 2 Cite Save

Cited by 6

Estimating a Common Period for a Set of Irregularly Sampled Functions with Applications to Periodic Variable Star Data

[JP Long](#), [EC Chi](#), [RG Baraniuk](#) - [arXiv preprint arXiv:1412.6520](#), 2014 - [arxiv.org](#)

Abstract: We consider the estimation of a common period for a set of functions sampled at irregular intervals. This problem arises in astronomy, where the functions represent a star's light curve observed through different photometric filters. While current methods ... Cite Save

Cited by 3

A Multiband Generalization of the Analysis of Variance Period Estimation Algorithm and the Effect of Inter-band Observing Cadence on Period Recovery Rate

[N Mondrik](#), [JP Long](#), [JL Marshall](#) - [arXiv preprint arXiv:1508.04772](#), 2015 - [arxiv.org](#)

Abstract: We present a new method of extending the single band Analysis of Variance period estimation algorithm to multiple bands. We use SDSS Stripe 82 RR Lyrae to show that the period recovery rate is significantly improved when observations per band and non-simultaneous ... Cite Save

Cited by 1

(* not that I'm keeping track . . .)

~ Thank You! ~



Email: jakevdp@uw.edu



Twitter: [@jakevdp](https://twitter.com/jakevdp)



Github: [jakevdp](https://github.com/jakevdp)



Web: <http://vanderplas.com/>



Blog: <http://jakevdp.github.io/>

Driving Reproducibility at UW

Reproducibility of research has clear benefits, but there are many practical challenges with the approach in today's scientific world. After discussing just *what* we mean by “reproducible”, I'll dive into some of our efforts at UW to encourage reproducible research on campus, and offer a few case studies drawn from my own experiences.