



Conducting Reproducible Research using R Packages

Gabriel J. Odom, PhD, ThD

http://rpubs.com/gabrielodom/FIU_Stempel_20190418

April 16, 2019

Overview

Overview

- About Me
- Reproducible Data Science and Software
- Example 1: Decentralised Water Quality Monitoring
- Example 2: Detecting Pathways that Drive Cancer
- Conclusion

About Me

Academic Interests

Research

- High-Dimensional Statistics ($p \gg n$)
- Computational Statistics
- Bayesian Statistics
- Spatial Statistics
- Data Science
- “Big” EHR Data

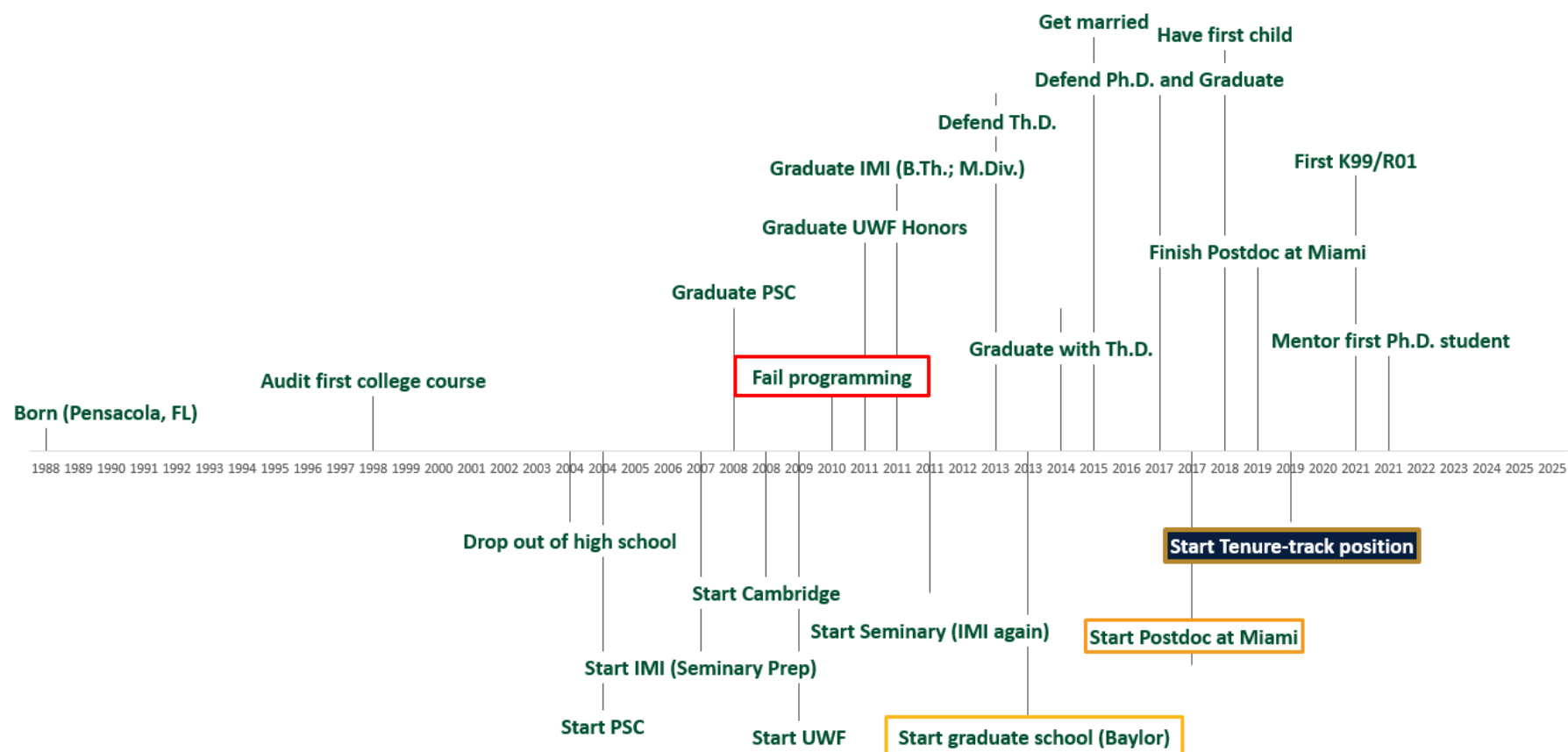
Curiosities

- Pedagogy
- Health and Science Ethics
- Leadership Development
- Health Economics
- Infectious Disease Modelling
- Apologetics and Patristics
- Film Music

Vitae: http://rpubs.com/gabrielodom/curriculum_vitae

GitHub: <https://github.com/gabrielodom>

Academic Timeline



Teaching

Courses Taught

- Computational Statistics
- High-Dimensional Statistics
- R for Data Science and Development
- Special Topics in Research
- Statistics for Business I & II
- Statistics for Health Sciences

Courses I Can Teach

- Bayesian Methods
- Bayesian Theory
- Spatial Statistics
- Multivariate Statistics
- Applied Regression
- Introduction to Machine Learning

Statistics vs Data Science

Six Divisions of “Greater” Data Science (Donaho, 2017):

1. Data Exploration and Preparation: **exploratory data analysis** and data cleaning
2. Data Representation and Transformation: **mathematical transformations**, querying databases, and data formatting
3. Computing with Data: programming languages and code packaging
4. Data Modeling: **statistical modelling** and machine learning
5. Data Visualization and Presentation: from **plots** to interactive websites
6. Science about Data Science: **meta-analysis** about the utility of statistical and computational tools

* *Traditional “statistical” topics in **bold**.*

My Philosophy

“When you use something, leave it better than when you found it.”

– My mother

Investing in people and for people drives my work:

1. My students should be better scientists, better collaborators, and better people after my classes / mentoring
2. The PIs I collaborate with and their staff should feel that I added lasting value
3. My research should be easy for other scientists to **replicate, use, and build upon**

Reproducible Data Science and Software

NATURE | NEWS

nature
International journal of science

First results from psychology's largest
reproducibility test

Crowd-sourced effort raises nuanced questions about what counts as replication.

Comment | Published: 28 March 2012

SCIENCE

Drug development

Raise standards
research

C. Glenn Begley & Lee M. Ellis

BBC

Sign in

Psychology's Replication Crisis Is Running Out of Excuses

Another big project has found that only half of studies can be repeated. And this time, the usual explanations fall flat.

NEWS

Home | Video | World | US & Canada | UK | Business

Science's 'Replication Crisis' Has Reached Even The Most Respectable Journals, Report Shows

MIKE MCRAE 27 AUG 2018

Most scientists 'can't replicate studies by their peers'

NATURE | NEWS FEATURE

By T
Scie

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

ture
NEWS DRUG
DISCOVERY

ence | Published: 31 August 2011

...ve it or not: how much can we rely on
published data on potential drug targets?

Florian Prinz, Thomas Schlange & Khusru Asadullah

The Reproducibility Crisis

Reproducibility: if we repeat a study, the repetition should agree with the original results, or—at minimum—not refute the study's conclusions.

Published bio-science is largely not reproducible:

- Oncology: 53 published articles tested, six successes (**11%**) (Nature, 2012)
- Psychology:
 - 100 published articles, **39** successes (Nature News, 2015)
 - 71 published articles tested, 92 replication attempts, 35 successes (**38%**).
The PsychFileDrawer project is ongoing.
- Pharmacology: 67 published models tested, 14 successes (**21%**) (Nature Reviews, 2011)

The Ioannidis crusade

John Ioannidis, physician scientist, speaks harshly against the lack of replicability in science:

- "Replication validity of genetic association studies" (2001)
- "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research" (2005)
- "Why Most Published Research Findings Are False" (2005)
- "Why Most Clinical Research Is Not Useful" (2016)

Reproducible Data Science

“Before reproducibility must come preproducibility.”

“Instead of arguing about whether results hold up, let’s push to provide enough information for others to repeat the experiments ... In computational science, ‘reproducible’ often means that **enough information is provided to allow a dedicated reader to repeat the calculations** in the paper for herself.”

– Philip B. Stark, Professor of Statistics, UC Berkeley

Forensic Bioinformatics

Baggerly and Coombes (2009) founded the field of Forensic Bioinformatics. They wrote this paper while waging war against fabricated data:

- Dr. Potti's falsified research had made it to the clinical trial phase before it was stopped.
- Dr. Potti was found to have "engaged in research misconduct by including false research data".
- He was fired from the Duke University School of Medicine.
- Links: Dr. Baggerly's slides on their process; HHS Office of Research Integrity statement on Anil Potti; *Fostering Integrity in Research*, Appendix D

Potti's mistakes could have been detected much sooner—and corrected—if his team had built a software package to document their data science.

Software Packages

Software Package: a self-contained suite of programs necessary to accomplish a set of related tasks.

Software packages for data science often contain the following components:

- functions and scripts written in one or more programming languages
- example data
- code and data documentation
- metadata about the package development process, team, and timeline
- a users' guide interweaving motivation, documentation, code, output, and analysis, known as a *vignette*

Software packages published in accompaniment to a published paper document and organize the code necessary to replicate **every aspect of the data analysis** shown in the paper.

Why Packages

Software packages:

1. make your entire paper **reproducible**,
2. follow the spirit of the *Literate Programming* principle ([Knuth, 1984](#)),
3. steer your research process towards higher **ethics** and **transparency**, and
4. enable the next generation of scientists to “stand on your shoulders”.

Reproducible Data Science means that we publish the software package necessary to transmute our raw data into our published results.

Example 1: Monitoring Water Quality with **mvMonitoring**

Motivating Example

- Water conservation is a growing concern around the world (and currently in the western U.S.)
- Lack of sanitation affects 35% of the world's population (65% of East Asia, 33% South Asia, 31% Sub-Saharan Africa) (CDC, 2016; WHO, 2008):
 - Over 1900 children die *each day* due to sanitation-preventable diarrheal diseases
 - “Almost one tenth of the global disease burden could be prevented by improving water supply, sanitation, hygiene and management of water resources”
 - Communities without sanitation are less likely to provide education to female students after puberty
- Decentralized wastewater reclamation is a modern strategy to provide access to potable water and curb water waste

Decentralized Water Treatment

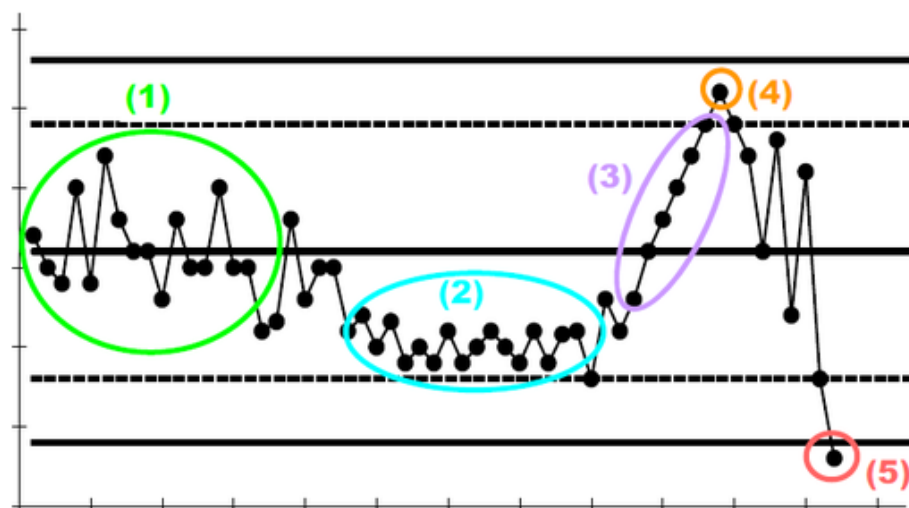
- Most global communities do not have the infrastructure for centralized wastewater treatment or reclamation
- Decentralized wastewater treatment processes are cheaper to build and maintain, and require little human interaction
- Decentralized treatment requires sophisticated automatic process monitoring
- These processes:
 - are adaptive (exhibiting change over weeks, months, and years relative to the local population),
 - are dynamic (exhibiting strong temporal dependence),
 - have multiple steady-states, and
 - use redundant sensors (multiple highly-correlated features)

Example Problem

- We care about monitoring the state of the system to ensure that the water quality stays within proper limits and that the system does not collapse.
- A *sequencing-batch membrane bioreactor* is a tightly-controlled hybrid biological-mechanical purification system.
- The internal biological ecosystem is very sensitive; if it crashes, the entire treatment facility can be inoperable for *months*
- One such bio-system crash shut down a decentralized plant for **four months**

Methods

Control charts are often employed to analyze dynamic processes (Shewhart, 1926). Example:



We created a new control chart that addresses the four process-monitoring concerns listed previously.

Multi-state, Adaptive, Dynamic PCA

Control charts monitor independent, univariate sensors. We need to monitor many dependent sensors simultaneously. We need PCA (*principal components analysis*):

- Orthogonal linear combinations of the original features
- maximize the amount of information explained.
- **Benefits:**
 - reduced number of charts to monitor
 - charts are now independent (under mild assumptions)
 - account for correlated sensors.

Multi-state, Adaptive, Dynamic PCA combines the following three algorithmic components: Dynamic PCA, Adaptive PCA, and Multi-state PCA.

PCA Formulation

- Let $\mathbf{X} \in \mathbb{R}_{(N-\ell) \times p}$ be the observed process data (including up to ℓ previous time points as predictors).
- Let $\mathbf{P}_d \in \mathbb{R}_{p \times d}$, for $d < p$, be the projection matrix of the d eigenvectors corresponding to the largest d eigenvalues of \mathbf{X} .
- The principal components matrix, $\mathbf{Y} = \mathbf{X}\mathbf{P}_d \in \mathbb{R}_{(N-\ell) \times d}$, is the transformation of the original p features into a d -dimensional orthogonal subspace.
- Multi-state, adaptive PCA estimates a different \mathbf{P}_d for each state and updates these estimates regularly.
- Instead of monitoring a new observation \mathbf{x}_{new} , we monitor $\mathbf{y}_{\text{new}} = \mathbf{x}_{\text{new}}\mathbf{P}_d$.

Adaptive and Dynamic PCA

Dynamic PCA

- Include any relevant sensor data from previous time points
- **Benefits:** how much water people used at 6PM yesterday is a great predictor for how much they will use at 6PM today

Adaptive PCA

- Change the projection over time:
 - each hour, day, week, etc., “learn” the most recent observations and “forget” the oldest observations
 - re-estimate the principal components for the next time period.
- **Benefits:** account for the fact that people and communities change unpredictably over time

Multi-state PCA

- Different steps in water-treatment process \implies different relationships between correlated sensors
- Estimate the principal components for each state in the process independently
- **Benefits:** within-state variance is smaller than between-state variance

Monitoring Statistics: Hotelling's T^2

Hotelling's $T^2 = \mathbf{y}_{\text{new}} \Lambda_d^{-1} \mathbf{y}_{\text{new}}^T$, where \mathbf{y}_{new} is a new observation to check, and Λ_d is the diagonal matrix of the first d eigenvectors.

- T^2 is the Mahalanobis distance of the mapped value \mathbf{y}_{new} from the original p -space into the d -dimensional PCA subspace.
- T^2 measures deviations from expectation in the lower subspace.

Monitoring Statistics: Squared Prediction Error

First, let $\mathbf{e}_{\text{new}} := \mathbf{x}_{\text{new}} - \mathbf{y}_{\text{new}} \mathbf{P}_d^T$. Then, $SPE = \mathbf{e}_{\text{new}} \mathbf{e}_{\text{new}}^T$.

- SPE is the squared distance between the original sensor vector and reduced-dimension approximation of this vector.
- SPE measures the goodness-of-fit of the d -dimensional model.

Multiple Correlated and Redundant
Sensors



(MSAD-PCA)

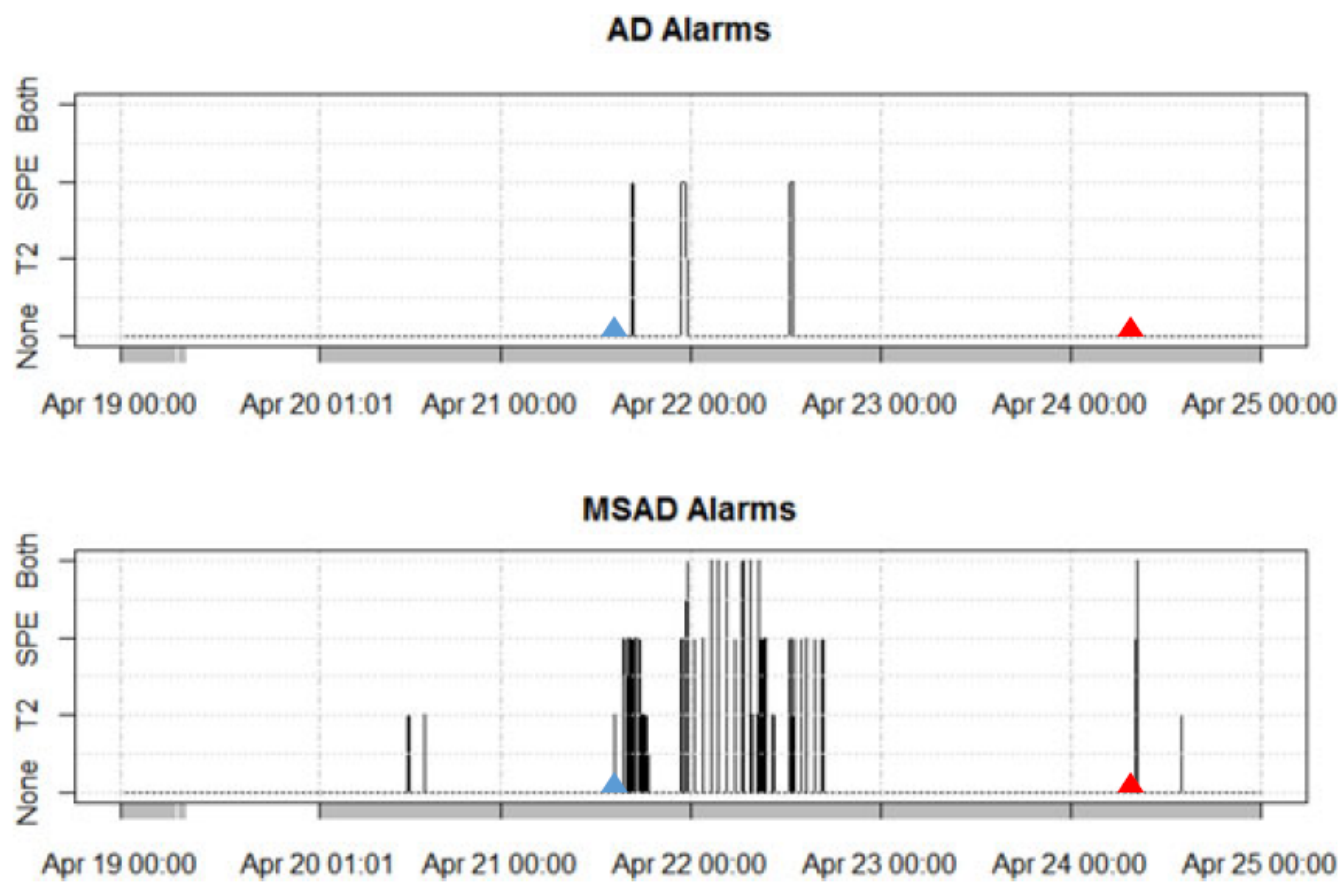
Two Dozen Independent
Composite Sensors (PCs)



(Hotelling's T^2 and SPE)

Two Control Charts

Performance




Publication



SPECIAL ISSUE PAPER |  Full Access

Multistate multivariate statistical process control

Gabriel J. Odom, Kathryn B. Newhart, Tzahi Y. Cath, Amanda S. Hering 

Multi-State Adaptive-Dynamic Process Monitoring

build **passing** downloads 2792 CRAN 0.1.0

Overview

We create this package, `mvMonitoring`, from the foundation laid by Kazor et al (2016). This package is designed to make simulation of multi-state multivariate process monitoring statistics easy and straightforward, as well as streamlining the online process monitoring component.

Installation from CRAN

As of 18 October, 2017, we have submitted this package to CRAN. Shortly thereafter, you will be able to install the stable version of the package via

```
install.packages("mvMonitoring")
```

Study Replication

- Our code and examples are available at <https://gabrielodom.github.io/mvMonitoring/index.html>
- The simulation study and data analysis are completely documented and repeatable through the users' guide on this page.
- This method and corresponding software are currently in use at the Mines Park Decentralized Wastewater Treatment facility in Golden, CO.
- This project is part of a pilot program for “sustainable clean water and sanitation”.

Example 2: Interrogating Biological Pathways with **pathwayPCA**

Motivation

- Each year, 18 Million new cancer cases are diagnosed, and nearly 10 Million people die from cancer ([WHO, 2018](#))
- A person dies of cancer every **3.3 seconds**.
- Cancer is the second leading cause of death in the US ([CDC, 2017](#))
- Different cancers cause disparities in mortality for ([NCI, 2019](#)):
 - women
 - minorities
 - the indigent
 - the elderly
- Mortality is affected by society, but incidence is driven by genetics

Genetics and Cancer

- The Central Dogma of molecular biology states that DNA (genes) encode RNA, RNA encode proteins, and proteins govern the behavior of the cell (thereby governing the tissue) (Clancy et al., 2008)
- Cancers are primarily caused by *multiple mutations* in genes (Knudson hypothesis) belonging to certain biological processes, such as apoptosis (programmed cell death) or proliferation (ACS, 2014)
- Many cancers are caused by multiple mutations of *multiple genes*, all working in concert to advance the disease state (Sugimura et al., 1992)

Challenges

While discovering single-gene cancer drivers is important, such as TP53 ([NCBI, 2011](#)), this approach has a few challenges:

- Cancers are often caused by concurrent abnormalities in multiple genes
- Gene knockdown experiments often find redundant cancer-driving genes
- Single-gene testing of 20,000 genes has very low statistical power after controlling for the false discovery rate

Solutions

To overcome these challenges:

1. Group genes by their biological pathways ([NIH NHGRI, 2015](#)).
 - Depending on the grouping, there are anywhere from 50-5000 pathways to consider.
 - In cancer research, we usually care about
 - The C2 Canonical Pathways collection ([Broad Institute](#)) in the Molecular Signatures Database (1,329 pathways), or
 - The WikiPathways collection (approximately 500 pathways) ([Slenter et al., 2018](#)).
2. For each of the pathways selected, test a summary of the pathway for the presence of a statistically-significant relationship with some outcome (survival time, tumor size, or cancer subtype)

Methods to Summarize Pathways

Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes ^{FREE}

Xi Chen ✉, Lily Wang, Jonathan D. Smith, Bing Zhang [Author Notes](#)

Bioinformatics, Volume 24, Issue 21, 1 November 2008, Pages 2474–2481,



[Original Article](#) | [Full Access](#)

Pathway-based analysis for genome-wide association studies using supervised principal components

Xi Chen ✉, Lily Wang, Bo Hu, Mingsheng Guo, John Barnard, Xiaofeng Zhu



[Stat Appl Genet Mol Biol](#). 2011 Jan 1; 10(1): 48.

PMCID: PMC3215429

Published online 2011 Oct 24. doi: [10.2202/1544-6115.1697](#)

PMID: [23089825](#)

Adaptive Elastic-Net Sparse Principal Component Analysis for Pathway Association Testing

[Xi Chen](#)

SuperPCA and AES-PCA

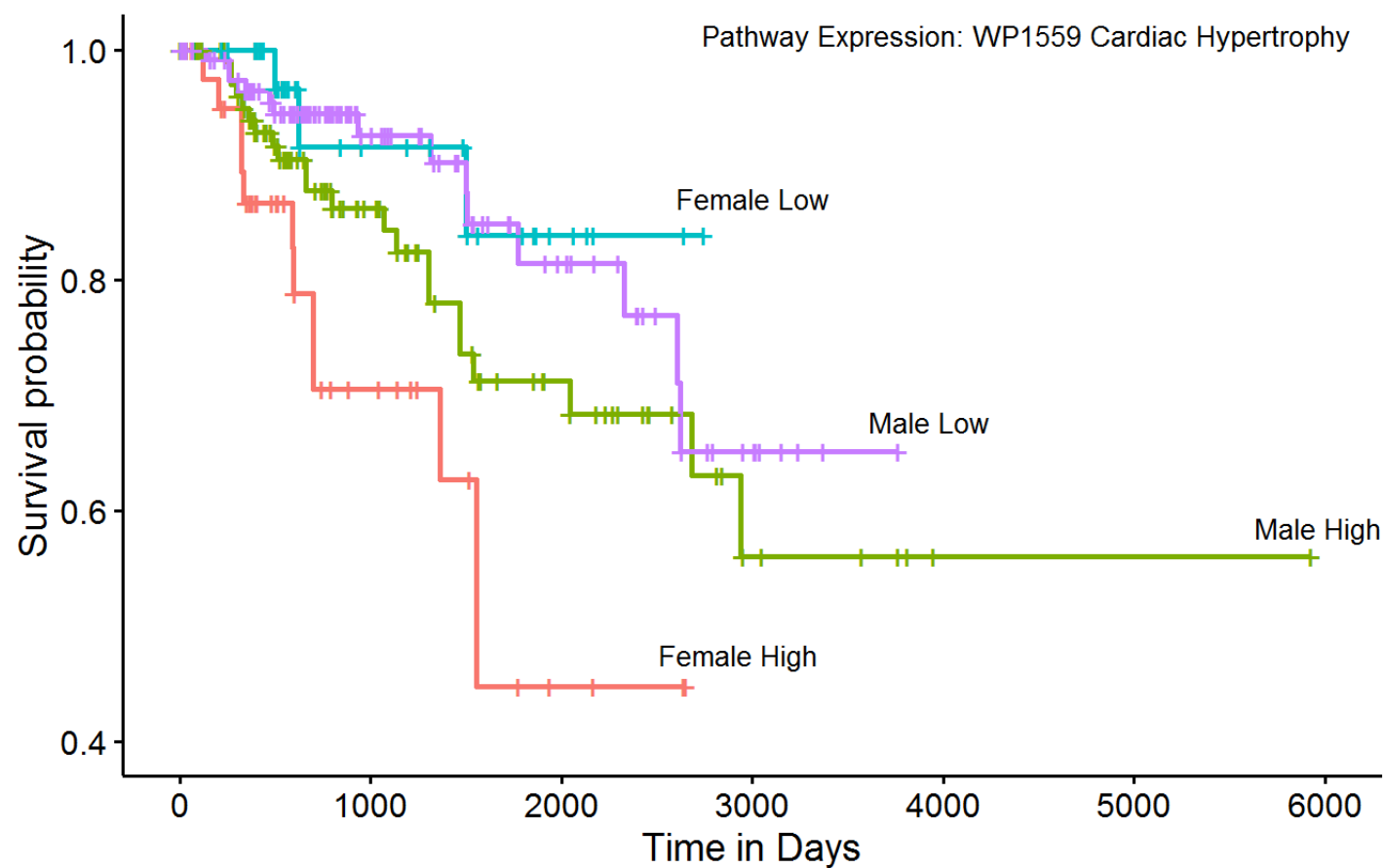
- Let $\mathbf{X}_i \in \mathbb{R}_{n \times p_i}$ be the expression matrix for features in pathway i .
- Supervised PCA (SuperPCA; [Chen et al., 2008](#); [Chen et al., 2010](#)):
 - ranks each feature in pathway i by its univariate relationship with the outcome of interest (survival time, tumor size, cancer subtype, etc.), then
 - extracts principal components from the most relevant features
- Adaptive, Elastic-net, Sparse PCA (AES-PCA; [Chen, 2011](#)) combines into a single objective function the following methods:
 - Elastic-Net ([Zou and Hastie, 2005](#))
 - Adaptive Lasso ([Zou, 2012](#))
 - Sparse Principal Component Analysis ([Zou, Hastie, and Tibshirani, 2012](#))
- AES-PCA then extracts principal components from pathway i which minimize this composite objective function

Pathway Associations with Cancer Survival

- Many cancers have pronounced survival disparities to gender (Dorak and Karpuzoglu, 2012)
- Renal cancers have a known gender effect (ACS, 2017)
- We found a potential association between survival outcomes and the interaction of gender and the first principal component of pathway WP1559.
- This pathway measures transcription factors related to *cardiac hypertrophy* (thickening of the heart muscle).
- A recent paper in *Cardiorenal Medicine* shows a strong relationship between kidney diseases and cardiac hypertrophy (De Lullo et al., 2015).
- Our Cox Proportional Hazards model was

$$h(t) = h_0(t) \exp[\beta_1 \text{PC}_1 + \beta_2 \text{male} + \beta_3 (\text{PC}_1 \times \text{male})]$$

Kidney Cancer Survival



Execution

pathwayPCA 0.99.5

Reference

User Guides ▾

Version Logs

pathwayPCA: an R package for integrative pathway analysis with modern PCA methodology and gene selection

Initial Date: 2017-10-19

Introduction

With the advance in high-throughput technology for molecular assays, multi-omics datasets have become increasingly available. However, most currently available pathway analysis software do not provide estimates on sample-specific pathway activities, and provide little or no functionalities for analyzing multiple types of omics data simultaneously. To address these challenges, we present pathwayPCA, a unique integrative pathway analysis software that utilizes modern statistical methodology on principal component analysis (PCA) and gene selection.

The main features of pathwayPCA include:

1. Performing pathway analysis for datasets with binary, continuous, or survival outcomes with computational efficiency.
2. Extracting relevant genes from pathways using the SuperPCA and AESPCA approaches.
3. Computing PCs based on the selected genes. These estimated latent variables represent pathway activity for individual subjects, which can be used to perform integrative pathway analysis, such as multi-omics analysis, or predicting survival time.
4. Can be used to analyze studies with complex experimental designs that include multiple covariates and/or interaction effects. For example, testing whether pathway associations with clinical phenotype are different between male and female subjects.
5. Performing analyses with enhanced computational efficiency with parallel computing and enhanced data safety with S4-class data objects.

Links

Download from BIOC at
<https://www.bioconductor.org/packages/pathwayPCA>

Browse source code at
<https://github.com/gabrielodom/pathwayPCA>

Report a bug at
<https://github.com/gabrielodom/pathwayPCA/issues>

License

GPL-3

Developers

Gabriel Odom
Author, maintainer

James Ban
Author

Lizhong Liu
Author

Lily Wang
Author

Steven Chen
Author

pathwayPCA

platforms **all** rank **unknown** posts **0** in Bioc **devel only**
build **ok** updated **< 1 month**

DOI: [10.18129/B9.bioc.pathwayPCA](https://doi.org/10.18129/B9.bioc.pathwayPCA)  

This is the **development** version of pathwayPCA; to use it, please install the [devel version](#) of Bioconductor.

Integrative Pathway Analysis with Modern PCA Methodology and Gene Selection

Bioconductor version: Development (3.9)

Apply the Supervised PCA and Adaptive, Elastic-Net, Sparse PCA methods to extract principal components from each pathway. Use these pathway- specific principal components as the design matrix relating the response to each pathway. Return the model fit statistic p-values, and adjust these values for False Discovery Rate. Return a data frame of the pathways sorted by their adjusted p-values. This package has corresponding vignettes hosted in the "User Guides" page of , and the website for the development information is hosted at .

Author: Gabriel Odom [aut, cre], James Ban [aut], Lizhong Liu [aut], Lily Wang [aut], Steven Chen [aut]

Maintainer: Gabriel Odom <gabriel.odom at med.miami.edu>

Citation (from within R, enter `citation("pathwayPCA")`):

Odom G, Ban J, Liu L, Wang L, Chen S (2019). *pathwayPCA: Integrative Pathway Analysis with Modern PCA Methodology and Gene Selection*. R package version 0.99.5,
<https://gabrielodom.github.io/pathwayPCA/>; <https://github.com/gabrielodom/pathwayPCA>.

Conclusion

Review

- Published research in biomedical and social sciences is largely irreproducible.
- Lack of documentation during the data cleaning, visualizing, modelling, and analyzing steps is partly to blame.
- Software packages are valuable, self-contained, research apparatuses that greatly increase the chance that published research is replicable by the scientific community.

Be a good steward of your science! When you publish methodological or data analytical research, build a software package to share your code, data, and reports.

Next Steps

Tools for reproducible data science, bioinformatics, and biostatistics I am currently collaborating on or recently completed:

- **Moonlight**: A multi-omics tool to interpret pathways and indict cancer-driver genes (with Antonio Colaprico, Steven Chen, et al.).
- **coMethDMR**: An unsupervised approach for identifying differentially-methylated regions in epigenome-wide association studies (with Lissette Gomez & Lily Wang).
- **DMRcompare**: An evaluation of supervised methods for identifying differentially-methylated regions in Illumina methylation arrays (with Saurav Mallik, Lily Wang, & Steven Chen).
- **regionPredictR**: Predict clinical outcomes using CpGs from differentially-methylated regions of the genome (with Lizhong Liu & Lily Wang).
- **rnaEditR**: An unsupervised approach to cluster regions of co-edited RNA (with Jenny Zhang & Lily Wang).

Acknowledgements

- **Chen and Wang Translational Bio Lab:** Steven Chen, Lily Wang, Lizhong Liu, Antonio Colaprico, James Ban, Jenny Zhang, Zhen Gao, Lissette Gomez, and Shirley Sun
- **Hering and Cath Engineering Lab:** Amanda Hering, Tzahi Cath, Kate Newhart, Ben Barnard, Karen Kazor, and Melissa Johnson
- **My mentors:** Dean Young, Amanda Hering, James Stamey, Steven Chen, and Lily Wang

Thank You!
Questions?