Using R to Make Sense of NMR Datasets

Prof. Bryan A. Hanson
Dept. of Chemistry & Biochemistry
DePauw University, Greencastle Indiana

5th Annual
Practical
Applications of NMR
in Industry Conference
SONESTA RESORT HILTON HEAD ISLAND • HILTON HEAD ISLAND, SC
PANIC
FEBRUARY 20-23, 2017

Presentation available at github.com/bryanhanson/PANIC2017
Additional references & resources on last slide

DEPAUW UNIVERSITY

What Exactly is R?

- R is a free software "environment" for statistical computing and graphics
- The Ecosystem:
  - Base R via "R-Core"
  - Add-on packages from many authors
    - Comprehensive R Archival Network (aka CRAN) (>10,000 packages)
    - Bioconductor (>1,300 packages)
    - Unofficial repositories: Github, Gitlab, SourceForge etc.
  - Support forums
  - User guides galore!

The Ecosystem: Support Resources

- Official Documentation
- Focused, Topical Task Views
- R-Bloggers: over 600 R-oriented bloggers
- Stack Overflow: over 160K questions on use of R
- Hundreds of "Intro to R" documents on the web
- Dozens of free R books on the web
- Many packages have a "vignette" or user guide.
- More resources on last slide.

Features of R

- Written by statisticians
- " . . . a rather unlikely linguistic cocktail . . . "[1]
- Cross-Platform: Windows, Linux, Mac OS
- Infrastructure: ready integration, interactive options
- Interfaces to many other languages, programs
  - SAS, SPSS, python, JavaScript, MATLAB, C++ etc.
- Several ways of running in parallel, using multiple cores
- Command line, or several GUI options

[1] Structure of the R Language

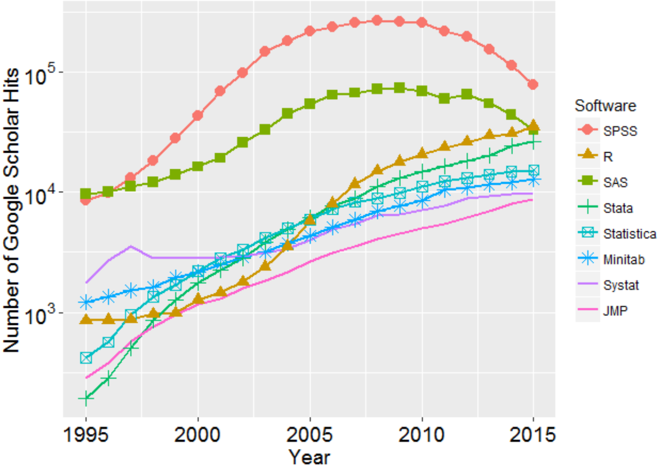
R is Open Source

- Free!
- Transparent: All code readily available for inspection
- "Given enough eyeballs, all bugs are shallow" – Linus Torvalds
- Many parts of the ecosystem are community driven

"Open source means everyone can see my stupid mistakes. Version control means everyone can see every stupid mistake I've ever made."

Karl Broman

Do People Use R?[2]



[2] Bob Muenchen r4stats.com/articles/popularity/

Who Uses R?

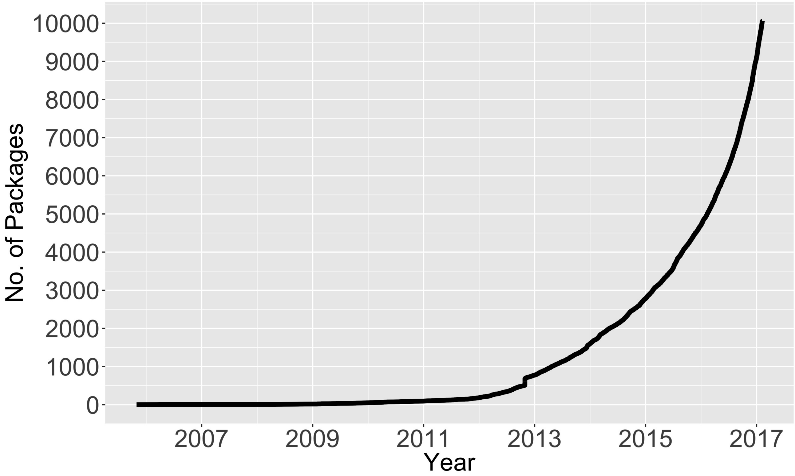
- AirBnB
- Zillow[3]
- Etsy
- NYT
- Twitter
- Facebook

Zillow, Python, R, and GIS

Mike Babb, GIS Analyst

[3] Data Science at Zillow

User Contributed Packages[4]



[4] Script by Gergely Daróczi

Reproducible Research with R

- Automation of Workflow:

  $\text{data} \rightarrow \underline{\text{analysis code} + \text{explanatory text}} \rightarrow \text{figures} + \text{tables} + \text{text} = \text{report}$

- Many resources for reproducible research
- Several possible input formats
- Typical output formats are pdf files and web pages
- This presentation written with LaTeX and R via the knitr package.

What is ChemoSpec?

- ChemoSpec = Chemometrics + Spectroscopy
- Tools for exploratory data analysis
- No attempt to duplicate functions available on the spectrometer

ChemoSpec: Design Goals

- User friendly design
- Helpful error messages
- Reliable results
- High quality plots
- Consistent plot appearance
- Provide access to a wide range of chemometric operations
- Extensibility
- Developed with metabolomics and IR, NMR & Raman in mind

What Can ChemoSpec Do?

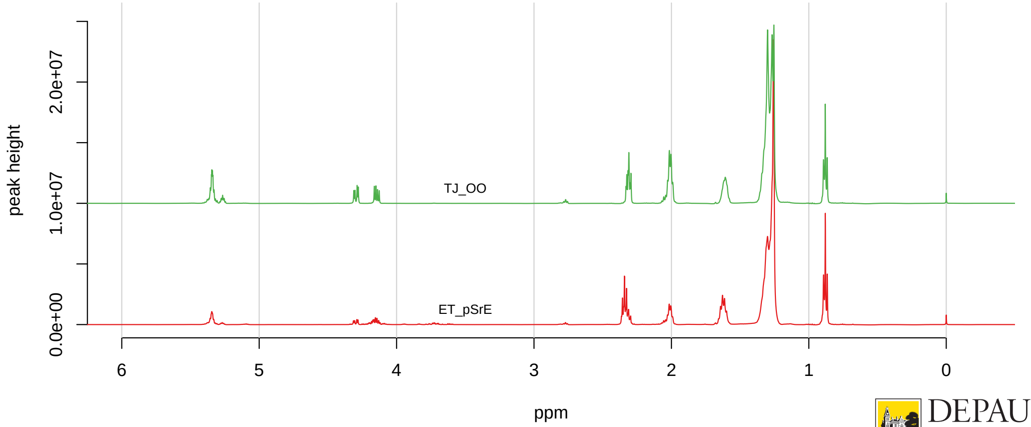
- Data Cleaning & Prep
  - Import data
  - Remove samples
  - Drop frequency ranges
  - Baseline correction
  - Signal alignment
  - Normalization
  - Savitzky-Golay filters

- Exploratory Data Analysis
  - Plotting & surveying
  - Hierarchical cluster analysis (HCA)
  - Principal component analysis (PCA)
  - PCA diagnostics
  - Score & loading plots
  - ANOVA-PCA
  - Empirical clustering

Demonstration Data Set: Saw Palmetto Caps

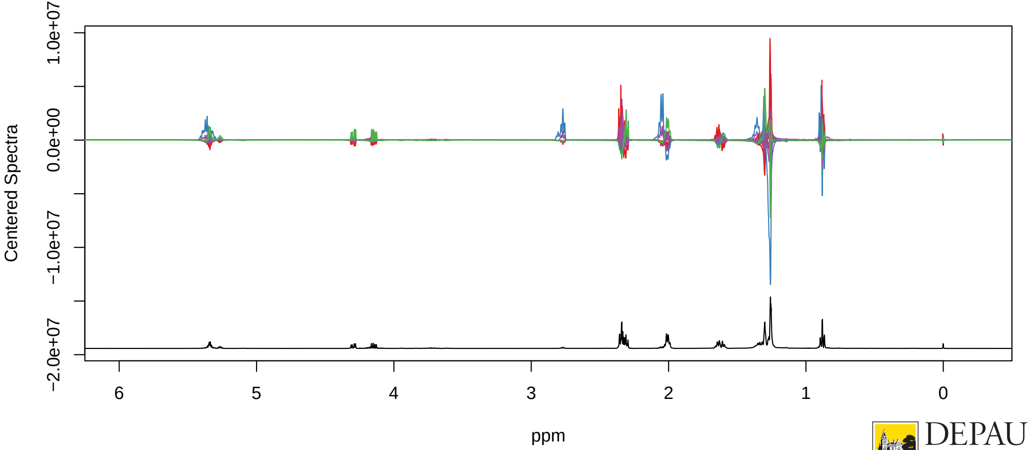
- Retail samples of *Serenoa repens* gel caps
- 500 MHz ^{1}H NMR in $CDCl_3$
- 4 samples were pure according to the label
- 10 samples have another oil present per label
- 2 outliers: olive oil, and evening primrose oil
- *Serenoa repens* extracts mainly fatty acids
- Outliers mainly triglycerides

Representative ^{1}H NMR Spectra

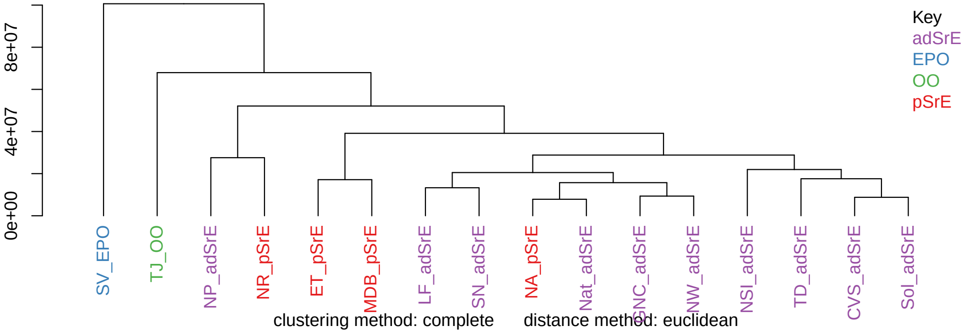


DEPAUW UNIVERSITY

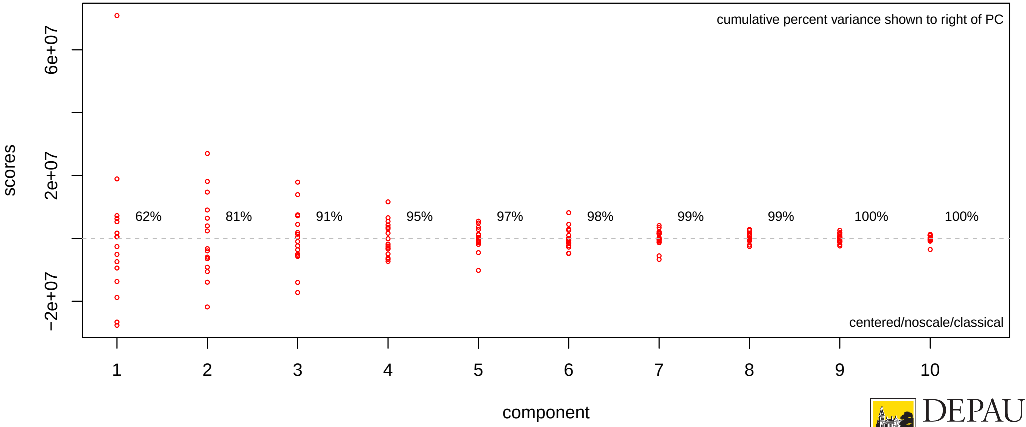
Where is the Variation in the ^{1}H NMR Spectra?



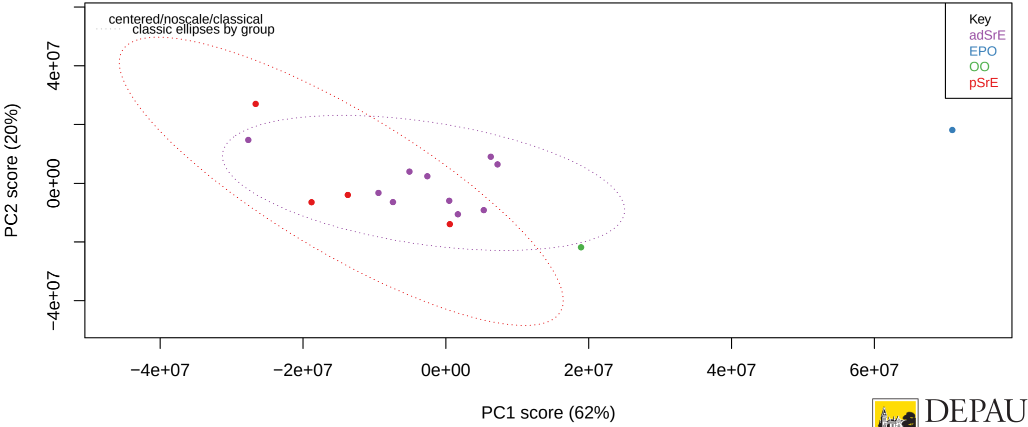
Hierarchical Clustering



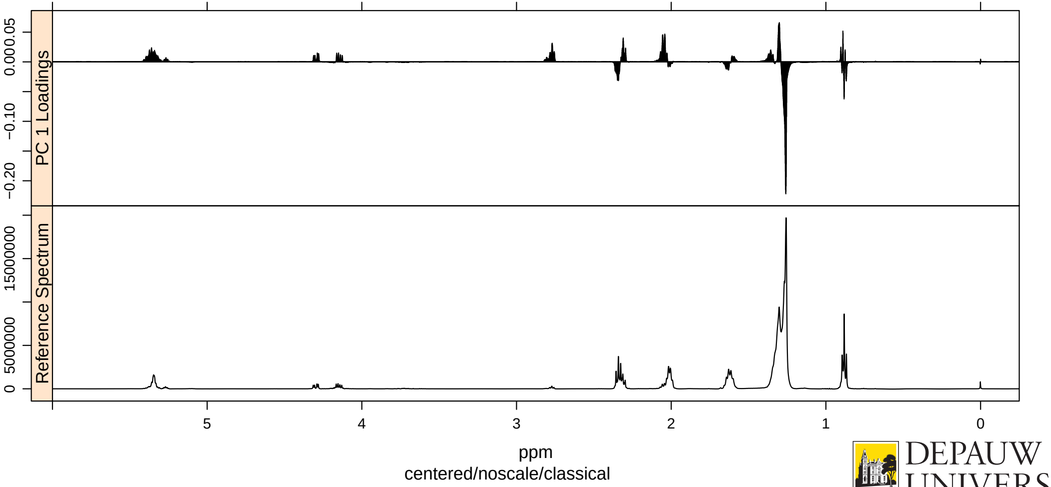
Principal Component Analysis: Scree Plot



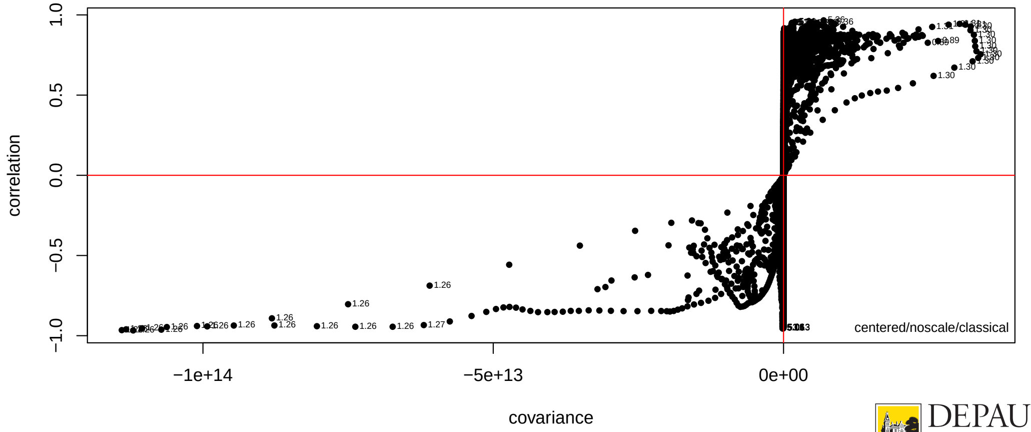
Principal Component Analysis: Score Plot



Principal Component Analysis: Loadings Plot



Principal Component Analysis: "S" Plot



Acknowledgements

- Thanks for your attention!
- Kristie Adams for the invite
- Sabbatical Support, DePauw University

DEPAUW UNIVERSITY

Additional References & Resources

- R Project Home Page
- Selected Topical Task Views
  - Chemometrics & Computational Physics
  - Clinical Trials
  - Experimental Design
  - Pharmacokinetics
  - Machine Learning
  - Reproducible Research
- Bioconductor Home Page