A Brief Tutorial on the `mixtools` and `tolerance` Packages for `R`

Derek S. Young

Dr. Bing Zhang Department of Statistics
University of Kentucky

Quality and Productivity Research Conference (QPRC) 2021
July 28^{th}, 2021

College of Arts and Sciences
Dr. Bing Zhang Department of Statistics

Funding Acknowledgment

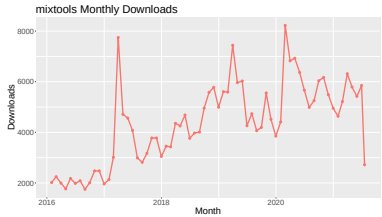
 This work is supported by the Chan Zuckerberg Initiative: Essential Open Source Software for Science (Grant Number 2020-225193).

Before We Begin...

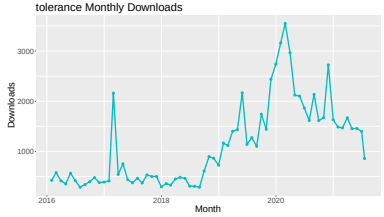
- You can download this presentation file and the corresponding R scripts from my GitHub repo: https://github.com/dsy109/Supplemental
- If you plan to run the R examples concurrently, make sure you have the most recent versions installed for all of the packages listed at the top of the .R file
- This tutorial was developed and tested using R version 3.6.2

Outline of Topics

- First release: 2006
- The package includes: functions for estimating classic mixture models (e.g., Gaussian mixture models and mixtures-of-regressions models), non/semiparametric procedures for mixture analysis, visualization tools, bootstrapping routines for estimating standard errors, and model selection routines
- The figure below shows a very strong increasing trend over time; the average number of monthly downloads per year is approximately 2080, 3819, 4436, 5151, and 5942, in 2016, 2017, 2018, 2019, and 2020, respectively



- First release: 2009
- The package includes: tolerance interval procedures for numerous parametric distributions, nonparametric settings, regression models, and some multivariate settings, as well as visualizations
- Used by NASA, 3M, EcoLab, PepsiCo, and NIST, among others
- The figure below also shows an increasing trend over the years as the average number of monthly downloads per year is approximately 418, 592, 557, 1717, and 2203, in 2016, 2017, 2018, 2019, and 2020, respectively



Development and Maintenance

- For `mixtools`, I am the original developer and the maintainer; for `tolerance`, I am the sole author
- Both packages were updated semi-annually in the past, but have been relegated to annual updates over the past 5 years simply due to time constraints
- Version control was informal and bugs were addressed through email communications with the end-users
  - Both packages predate `GitHub`
- Plotting capabilities of both packages were designed using `base R`
  - `mixtools` predates `ggplot2`, while `mixtools` and `tolerance` both predate `plotly`
- Many updates to the packages, starting back at the beginning of 2021, are being coordinated with Dr. Kedai Cheng (UNC-Asheville)

CZI: Essential Open Source for Software Program

- CZI notes that despite its importance, even the most widely-used research software often lacks dedicated funding for maintenance, growth, development, and community engagement
- The focus of this program is to provide software projects with resources to support their tools and the communities behind them, such as improving documentation, addressing usability, improving compatibility, onboarding contributors, or convening a community
- The program aims to support tools that are essential to biomedical research – for which `mixtools` and `tolerance` have both been extensively used – but the scope also more generally includes foundational tools and infrastructure that enable a wide variety of downstream software across several domains of science and computational research

Goals of Tutorial

- First and foremost, provide a general overview of the `mixtools` and `tolerance` packages
  - Briefly define finite mixture models and statistical tolerance regions
  - Demonstrate capabilities of both packages with analysis of real datasets
- Highlight some of the recent and planned additions to both packages
- Seek feedback from end-users about improvements that will benefit both researchers and practitioners alike
- Expand the respective user communities


mixtools

Motivation: Finite Mixture Models

- **Finite mixture models** are used to model data where the observations are sampled from a population that consists of several homogeneous subpopulations (called the **components** of the population), but to which subpopulation each observation belongs is unknown
- Estimation of mixture components is an unsupervised learning task
- Make a *soft* probabilistic classification of each observation to a component, whereas cluster analysis performs a *hard* classification to a cluster
  - Finite mixture models are, thus, naturally used as the underlying models for model-based clustering
- Can be used for density estimation and viewed as a kind of kernel method
- Used for numerous interesting problems in areas like pattern recognition, finance, psychology, and astronomy

Finite Mixture Models

Definition

The random vector $\mathbf{X} \in \mathbb{R}^p$ follows a parametric **mixture distribution** if it has the mixture density

$$f(\mathbf{x}; \boldsymbol{\psi}) = \sum_{j=1}^{k} \lambda_j g_j(\mathbf{x}; \boldsymbol{\theta}_j), \tag{1}$$

where $k \in \mathbb{N}$ is the number of components, which ideally should be known *a priori*, the λ_js are **mixing proportions** that comprise the standard simplex

$$\left\{ \boldsymbol{\lambda} \in \mathbb{R}^k : \sum_{j=1}^{k} \lambda_j = 1, \lambda_j > 0, j = 1, \ldots, k \right\},$$

and the g_j are component-specific densities from a parametric family with parameter $\boldsymbol{\theta}_j \in \Theta_j \subseteq \mathbb{R}^q$, such that Θ_j is open in $\mathbb{R}^q$. Usually, $g_j \equiv g$. The mixture density f in (1) is, thus, parameterized by $\boldsymbol{\psi} = (\lambda_1, \ldots, \lambda_{k-1}, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_k^\top)^\top \in \boldsymbol{\Psi} \subset \mathbb{R}^{k(q+1)-1}$.

- When $\mathbf{X}$ is a set of covariates for a univariate response Y, then the preceding framework is easily extended to model $Y|\mathbf{X}$ using a mixture structure, giving us what are called **mixtures-of-regressions models**

General Capabilities of `mixtools`

Parametric Procedures

- Classic parametric distributions
  - (Multivariate) normal [`normalmixEM`, `mvnormalmixEM`]
  - Gamma [`gammamixEM`]
  - Multinomial [`multmixEM`]
- Parametric regression settings
  - (Piecewise) linear regression [`regmixEM`, `regmixMH`, `segregmixEM`]
  - Logistic regression [`logisregmixEM`]
  - Poisson regression [`poisregmixEM`]

Nonparametric Procedures

- Individual repeated measures [`npEM`, `npMSL`]
- Conditionally independent multivariate repeated measures [`mvnpEM`]

Semiparametric Procedures

- Univariate symmetric location densities [`spEMsymloc`, `spEMsymlocN01`]
- Linear regressions with unspecified error structure [`spregmix`]
- Scaled mixture of censored data [`spRMM_SEM`]

General Capabilities of `mixtools` (ctd.)

Secondary Procedures

- Bootstrapping for standard errors [`boot.se`]
- Bootstrapping to determine number of components [`boot.comp`]
- Model selection routines [`regmixmodel.sel`, `multmixmodel.sel`]

Visualizations

- Mixture density estimates [`plot.mixEM`, `plot.mvnpEM`]
- Mixturegram *[`mixturegram`]
- FDR estimates from EM-like strategies [`plotFDR`]

*D. S. Young, C. Ke, and X. Zeng (2018), "The Mixturegram: A Visualization Tool for Assessing the Number of Components in Finite Mixture Models." *Journal of Computational and Graphical Statistics*, 27(3):565–575.

Format of Typical `mixtools` Function

```
foo_EM(data, k, theta, conv.crit)
```

Function Inputs

- `data`: Often a vector, matrix, or data frame
- `k`: The number of components to fit for the mixture; an integer greater than 1
- `theta`: Optional starting values for the parameters in the model, otherwise a starting value strategy is employed
- `conv.crit`: Controls for determining convergence of the EM algorithm

Function Outputs

- Final parameter estimates; i.e., the maximum likelihood estimates
- Loglikelihood values to monitor convergence or use in model selection criteria calculations (e.g., AIC, BIC)
- Posterior membership probabilities of each observation (subject); can be used for model-based clustering

Example #1: Quasars Data

- Quasars are extremely luminous objects in the Universe that arise from the accretion of gas onto supermassive black holes in the center of a galaxy
- The analysis of absorption lines in quasar spectra aid in the study of metal-enriched environments
- We consider the normalized intensity of the quasar light for the 3-times-ionized silicon line Si IV 1394 for the $z = 0.653411$ absorption system
- This dataset consists of $n = 104$ measurements
- We will model with a mixture of normals, and employ model selection criteria and the mixturegram
- Source: J. C. Charlton et al. (2013), "High-Resolution STIS/Hubble Space Telescope and HIRES/Keck Spectra of Three Weak Mg II Absorbers Toward PG 1634 + 706." *The Astrophysical Journal*; 589(1):111–125.

Example #2: Diffuse Large B-Cell Lymphoma

- Three biomarkers (CD3, CD5, and CD19) were measured on cells derived from the lymph nodes of patients diagnosed with Diffuse Large B-Cell Lymphoma (DLBCL)
- This dataset is for one subject with 8183 biomarkers
- Analyzed a subset of $n = 500$ observations to identify possible clusters of biomarkers
- We will model using a two-component mixture of bivariate normals and explore various visualizations
- Source: N. Aghaeepour et al. (2013), "Critical Assessment of Automated Flow Cytometry Data Analysis Techniques." *Nature Methods*; 10(3):228–238.

Example #3: Aphids Data

- An experiment was ran where a number of green peach aphids (winged insects able to transmit plant viruses) were released at various times over $n = 51$ small tobacco plants (used as surrogates for potato plants)
- Recorded the number of infected plants was recorded after each release; this is the response
- Modeled the number of infected plants as a function of the number of aphids released in each batch
- Two groups appear to emerge: one with overall lower rates of transmission and one with higher rates
- The hypothesis is that some batches of the aphids may have passed their maiden phase, which indicates lower transmission levels of the virus
- We will model the data with a mixture of linear regressions and a semiparametric mixture of regressions
- Source: T. R. Turner (2000), "Estimating the Propagation Rate of a Viral Infection of Potato Plants via Mixtures of Regressions." *Applied Statistics*; 49(3):371–384.

Some Other Applications Using `mixtools`

- Used to classify different seed types in maize production to understand their oil content. [Melchinger et al. (2015), "Oil Content is Superior to Oil Mass for Identification of Haploid Seeds in Maize Produced with High-Oil Inducers." *Crop Science*; 55(1):188–195.]
- Used to cluster patients with acute myeloid leukemia to aid in identifying mutations in genes in their DNA. [Kroeze et al. (2014), "Characterization of Acute Myeloid Leukemia Based on Levels of Global Hydroxymethylation." *Blood*; 124(7):1110—1118.]
- Used to evaluate the distribution of fluorodeoxyglucose (FDG) uptake in the dorsal and ventral streams in the prefrontal white matter in an effort to understand that feature in people with autism spectrum disorder and schizophrenia. [Mitelman et al. (2018), "Increased White Matter Metabolic Rates in Autism Spectrum Disorder and Schizophrenia." *Brain Imaging and Behavior*; 12(5):1290–1305.]
- Used to cluster color morph variants on manta rays based on long-term photo identification catalogs. [Venables et al. (2019), "It's Not All Black and White: Investigating Colour Polymorphism in Manta Rays Across Indo-Pacific Populations." *Proceedings of the Royal Society B*; 286(1912):1–10.]

Recent and Forthcoming Additions to `mixtools`

Recent Additions

- Launched `GitHub` repo (https://github.com/dsy109/mixtools)
- Added the `mixturegram` function
- Heavily-revised `gammamixEM` function

Forthcoming Additions

- `Shiny` app (https://mixtools.as.uky.edu/)
- Complete overhaul of graphics capabilities using `ggplot2` and `plotly`
- Expanding available `S3` methods, such as for `predict` to predict likely component membership of a new observation
- Recording a training module; will make available on my website


tolerance

Motivation: Tolerance Regions

- Based on a random sample, we have three primary statistical regions that can be calculated:
  - **Confidence regions** $\rightarrow$ provide regions for an unknown population parameter (e.g., mean vector, variance-covariance matrix)
  - **Prediction regions** $\rightarrow$ provide regions for one or more future observations from the sampled population
  - **Tolerance regions** $\rightarrow$ provide regions that are expected to contain at least a specified proportion of the sampled population
- Typical applications of tolerance regions (or **tolerance intervals** for the univariate setting) are found in clinical and industrial studies, statistical quality control, environmental monitoring, and setting statistically-based engineering design limits
- Used in regulations published by the Environmental Protection Agency (EPA), the Food and Drug Administration (FDA), the International Atomic Energy Agency (IAEA), and standard 16269-6 of the International Organization for Standardization (ISO)

Statistical Tolerance Sets

Definition

Let $\mathcal{F}$ be the class of all Borel measurable distributions in $\mathbb{R}^p$, $p \in \mathbb{N}$. Let $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be an iid random sample of size $n > p$ drawn from $F \in \mathcal{F}$, and $\mathbf{X}$ be a random vector that also follows F, independently of $\mathcal{X}$. Let $T(\mathcal{X})$ be a random subset of $\mathbb{R}$ and define $C_F(T(\mathcal{X}))$ to be the probability content under F of the set $T(\mathcal{X})$. If

$$\inf_{F \in \mathcal{F}} \Pr\{C_F(T(\mathcal{X})) \geq P\} = 1 - \alpha,$$

then $T(\mathcal{X})$ is a $(1-\alpha, P)$ **tolerance set**, where $P, \alpha \in (0,1)$ are the (probability) content level and confidence level, respectively. If

$$\mathrm{E}[C_F(T(\mathcal{X}))] = \beta,$$

then $T(\mathcal{X})$ is a β-**expectation tolerance set**, where $\beta \in (0,1)$ is the desired average content level of the tolerance set.

- β-expectation tolerance sets are just prediction sets

General Capabilities of `tolerance`

Univariate Tolerance Intervals

- Continuous distributions
  - Normal [`normtol.int`, `bayesnormtol.int`, `norm.ss`]
  - Extreme value [`exttol.int`]
  - Gamma [`gamtol.int`]
  - Nonparametric based on order statistics [`nptol.int`]
- Discrete distributions
  - Poisson [`poistol.int`, `fidpoistol.int`]
  - (Negative) binomial [`bintol.int`, `fidbintol.int`, `negbintol.int`, `fidnegbintol.int`]
  - (Negative) hypergeometric [`hypertol.int`, `neghypertol.int`]

Multivariate Tolerance Regions

- Multivariate normal [`mvtol.region`]
- Nonparametric hyperrectangular tolerance regions [`npmvtol.region`]

General Capabilities of `tolerance` (ctd.)

(Pointwise) Regression Tolerance Intervals/Regions

- Linear regression [`regtol.int`]
- Nonparametric regression [`npregtol.int`]
- Nonlinear regression [`nlregtol.int`]
- Multivariate linear regression [`mvregtol.region`]

Visualizations

- Histograms and control charts with tolerance limits [`plottol`]
- Scatterplots with tolerance regions [`plottol`]
- OC-type curves for sample size determination [`norm.OC`]

Format of Typical `tolerance` Function

```
foo_tol.int(data, alpha, P, side, method)
```

Function Inputs

- `data`: Often a vector or matrix
- `alpha`: The level chosen such that $(1-\alpha)$ is the confidence level
- `P`: The content level, P, of the tolerance interval/region
- `side`: Whether one-sided limits or a two-sided interval is required; unnecessary for multivariate settings
- `method`: Approximation method to use, often with respect to the k-factor

Function Outputs

- The $(1-\alpha, P)$ tolerance limits or intervals, which are formatted a certain way for regression tolerance intervals

Example #4: Monitoring Wells Data

- Upgradient monitoring wells show the background concentrations of constituents in groundwater; required by the EPA for monitoring these concentrations
- The data we will analyze are vinyl chloride concentration measurements collected from $n = 34$ clean upgradient monitoring wells
- We will compute $(0.95, 0.95)$ one-sided gamma tolerance intervals, which can be used to show if the vinyl chloride concentration is beyond a certain guideline, such as one used by the EPA
- Source: D. K. Bhaumik and R. D. Gibbons (2006), "One-Sided Approximate Prediction Intervals for at Least p of m Observations from a Gamma Population at Each of r Locations." *Technometrics*; 48(1):112–119.

Example #5: Hospital Infections Data

- Hospitals frequently study factors that are likely related to infections patients gain while hospitalized
- We will analyze data from a study of $n = 113$ US hospitals
- Numerous variables are available to study, but a parsimonious multiple linear regression model is one that has infection risk as the response, and average length of patient stay and number of x-rays given by the hospital as predictors
- We will compute $(0.90, 0.90)$ pointwise tolerance intervals
- Source: M. Kutner, C. Nachtsheim, and J. Neter (2004), *Applied Linear Regression Models, 4^{th} edn.* McGraw-Hill.

Example #6: Adolescent Kidney Function Reference Regions

- Kidney function laboratory tests include a urinalysis to screen for the presence of protein and blood in the urine, a blood urea nitrogen (BUN) test to check for waste product in the urine, and a test to obtain the estimated glomerular filtration rate (eGFR), which is used to detect the presence and cause of kidney disease
- Little information is available regarding normal reference values for kidney function in adolescents, which impacts how physicians diagnose and manage diabetes in this population
- The reference population studied is healthy US adolescents between 12 and 17 years of age, with a number of criteria used to determine "healthy"
- We will construct nonparametric $(0.95, 0.95)$ semi-space rectangular tolerance regions to represent the reference regions of normal adolescent kidney function; we will look at the males in this sample, yielding $n = 2529$ subjects in the reference sample
- Source: D. S. Young and T. Mathew (2020), "Nonparametric Hyperrectangular Tolerance and Prediction Regions for Setting Multivariate Reference Regions in Laboratory Medicine." *Statistical Methods in Medical Research*; 29(12):3569–3585.

Some Other Applications Using `tolerance`

- Used to construct tolerance intervals for establishing a pass/fail criterion of radiation portal monitors. [Burr and Gavron (2012), "Pass/Fail Criterion for a Simple Radiation Portal Monitor Test ." *Modern Instrumentation*; 1(3):27–33.]
- Used to help the design verification process of the Vagus Nerve Stimulation (VNS) Therapy® system, which is FDA-approved for the treatment of refractory epilepsy and treatment-resistant depression. [Young et al. (2016), "Sample Size Determination Strategies for Normal Tolerance Intervals Using Historical Data." *Quality Engineering*; 28(3):337–351.]
- Used in molecular lymph node analysis to aid in the understanding of prostate cancer. [Heck et al. (2018), "Molecular Lymph Node Status for Prognostic Stratification of Prostate Cancer Patients Undergoing Radical Prostatectomy with Extended Pelvic Lymph Node Dissection." *Clinical Cancer Research*; 24(10):2342–2349.]
- Used to develop nonparametric and normal-based tolerance intervals on GH-2000 to detect doping in competitive sports. [Liu et al. (2021), "Comparison of Normal Distribution-Based and Nonparametric Decision Limits on the GH-2000 Score for Detecting Growth Hormone Misuse (Doping) in Sport." *Biometrical Journal*; 63(1):187–200.]

Recent and Forthcoming Additions to `tolerance`

Recent Additions

- Launched `GitHub` repo (https://github.com/dsy109/tolerance)
- Launched `Shiny` app (https://tolerance.as.uky.edu/)
- Added new graphics capabilities based on `ggplot2` and `plotly`

Forthcoming Additions

- Improve efficiency of exact normal k-factor calculations using `Rcpp`
- Expanding the suite of regression-based tolerance intervals; e.g., simultaneous tolerance intervals and spatial regression tolerance regions
- Recording a training module; make available on my website

Final Comments

- `mixtools` and `tolerance` have thus far been significant components of my career, and have served as excellent supplemental outlets for the computational routines I've developed
- I hope you got a high-level overview of the capabilities of both packages and the direction I see them heading
- Please reach out to me via any mode of communication to express what would be helpful for your research; feel free to also contribute via their respective `GitHub` repos
- Last but not least, a big "thank you" to Profs. Bradley and Chicken for giving me a slot to deliver this tutorial!

References

mixtools

T. Benaglia, D. Chauveau, D. R. Hunter, and D. S. Young (2009). "mixtools: An R Package for Analyzing Mixture Models." *Journal of Statistical Software*, **32**(6), 1–29. https://www.jstatsoft.org/index.php/jss/article/view/v032i06/v32i06.pdf

D. Chauveau (2012). "New Mixture Models and Algorithms in the mixtools Package." *hal-00717545*, 1–3. https://hal.archives-ouvertes.fr/hal-00717545/document

tolerance

D. S. Young (2010). "tolerance: An R Package for Estimating Tolerance Intervals." *Journal of Statistical Software*, **36**(5), 1–39. https://www.jstatsoft.org/index.php/jss/article/view/v036i05/v036i05.pdf

D. S. Young (2014). "Computing Tolerance Intervals and Regions Using R." In M. B. Rao and C. R. Rao, editors, *Handbook of Statistics, Volume 32: Computational Statistics with R*, 309–338. North-Holland: Amsterdam, Netherlands.

Contact Information


derek.young@uky.edu


http://young.as.uky.edu


https://github.com/dsy109