Current Environmental Health Reports (2026) 13:14
https://doi.org/10.1007/s40572-026-00532-y

REVIEW

A Review of R Packages for Bayesian Model-based Clustering of High-dimensional Multivariate Environmental Exposures

Briana J. K. Stephenson[1] · **Yiran Fu**[2]

Received: 2 January 2026 / Accepted: 25 February 2026

Abstract

Purpose of Review The analysis of a collective set of environmental health exposures in large populations can be complex due to its dimensionality, correlation between exposures, and skewed data structures. Bayesian model-based clustering has gained traction for its computational advantages and model flexibility for the analysis of big data. This review aims to highlight R packages available to derive environmental exposure patterns in heterogeneous populations using a Bayesian mixture model framework.

Recent Findings Twenty-one packages have been developed since 2006 with the capability of analyzing multivariate exposures, with more than half being released in the last ten years. Packages are available to identify both cross-sectional and longitudinal exposure patterns from a joint set of continuous, binary, categorical, directional, and ranked data types. Bayesian nonparametric priors are available to determine the appropriate number of latent clusters to fit. Variable selection is also available in some packages to focus primarily on exposures that may drive underlying patterns present.

Summary Bayesian mixture models are an attractive solution to analyze joint environmental exposures and its subsequent impact on population health. While software packages have been developed to provide researchers with advanced tools and models suitable for data complexities seen in environmental exposure data, their applications in the literature is significantly low. Further awareness of software packages and their capabilities could increase utilization and improve analysis of environmental exposures and population health.

Keywords High-dimensional data · Mixture models · Bayesian model-based clustering · Environmental exposures · R packages

Introduction

There has been a growing interest in the field for jointly analyzing the impact of multiple coexisting exposures on population health. This interest is in response to a growing awareness of how exposures manifest in the real world. It is rare that a single exposure circulates in isolation. Instead, a myriad of exposures is highly correlated and interacting with one another. In environmental health, these often are analyzed as sets or classes of exposure that share similar chemical, social or structural properties. The combination upon which many of the individual exposures within each set interact could vary widely from person to person based on a myriad of factors. In nutrition, diet exposure, examined as the set of foods consumed over a fixed period, could vary based on one's geography, accessibility, and culture. Air pollution, examined as the set of different air particles circulating within a given area, could vary based on where one resides and interacts daily. Exposure to classes of chemicals, like PFAS or phthalates, which are analyzed as a combination of different individual substances are present in everyday products which could vary based on personal usage. Overall lived environment, which could refer to the collection of resources, services, and structures available within a community, could vary across different municipalities based on policy priorities and needs of its constituents.

Clustering assumes that there are individuals who share similar combination of the same sets of exposures. In large

✉ Briana J. K. Stephenson
bstephenson@hsph.harvard.edu

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[2] Massachusetts General Hospital Biostatistics, Boston, MA, USA

Published online: 18 March 2026

populations, Identifying subgroups of individuals with shared exposure patterns can increase our understanding of the exposure-disease relationship and explain why risks impacts a certain groups disproportionately. This makes it a popular dimension reduction technique to jointly analyze a large set of coexisting exposures.

Distance-based clustering techniques, like k-means, are commonly used in multivariate exposure analysis cases [1–4]. However, they fail to efficiently detect and cluster outliers that deviate from patterns shared by the population majority. They also are limited to data structures that are interpretable with a proximal distance measure. Alternatively, model-based clustering uses a mixture model to provide a probabilistic framework to group together observations or individuals that share similar patterns and characterize what those patterns may look like. With the flexibility to fit any set of distributional assumptions, mixture models can analyze a wide variety of data structures. This generates both reliable and easily interpretable results [5, 6].

Mixture models are generated from two probabilistic components: (1) the probability that an observation belongs in a given subgroup and (2) the shared behavior of that subgroup modeled via a probability distribution. The probability distribution offers a broad level of versatility to analyze and visualize a diverse collection of environmental exposure patterns. Figure 1 illustrates the basic components of a mixture model, and examples of plots visualizing the environmental exposure patterns derived from these models.

Groups of chemical exposures such as phthalates, per- and polyfluoroalkyl substances (PFAS), air pollutants, are measured as continuous. The distributions of these chemicals are typically skewed and must undergo data transformations to fit with a normal or t-distribution. However, continuous exposures that are bounded may reflect a uniform or beta distribution. Some continuous exposures that do not experience a significant loss of information when imposing a threshold level of detection or significant exposure, may be viewed as binary (Bernoulli distribution). The Bernoulli distribution may also be used in an analysis of the presence or absence of a set of exposures. Counts of exposed cases may warrant a Poisson distribution. Other discrete exposures may require additional flexibility assuming a negative binomial, Poisson lognormal, or extensions of common distributions as data warrants (e.g. Zero-inflated Poisson). There are also alternative data structures such as ranked data which may be used for measuring quality of

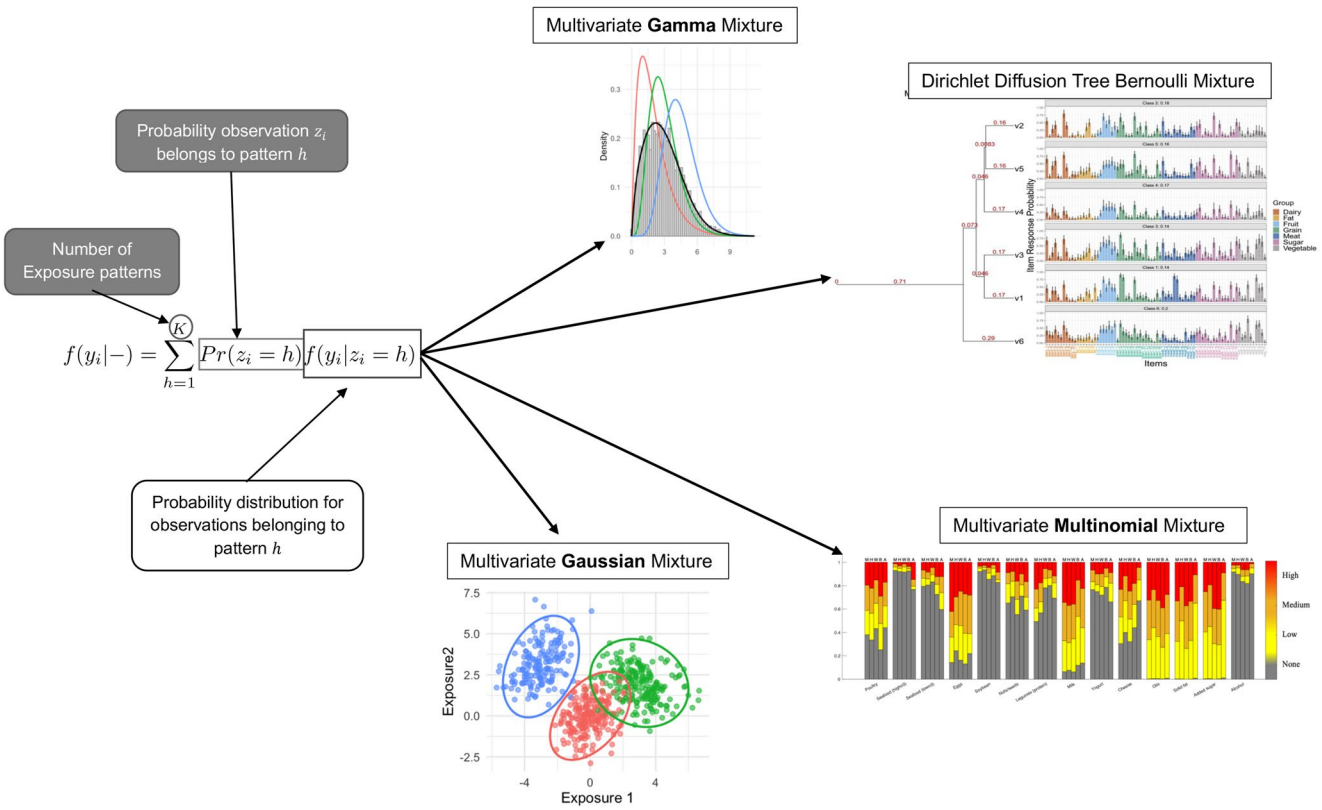

Fig. 1 Basic components of the mixture model equation are presented with illustrative examples of different distributional assumptions for exposure data. Multivariate Gamma and multivariate Gaussian mixture models were generated from simulations using ggplot2 R package [7]. Multivariate multinomial mixture model is abstracted from Stephenson et al. analyzing dietary intake patterns amongst 5 different subgroups [8]. Dirichlet diffusion tree Bernoulli mixture model figure is obtained from Li et al. and can be reproduced using the ddtlcm package [9, 10]

an environment or a comparison of standardized indices, as well as angular/directional data which may be present in evaluating the exposure of wind, migration patterns, or chemial spills [11, 12].

The mixture model assumes there exists a finite set of K latent clusters, where each cluster describes a set of shared behaviors by a set of exposures. Let $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denote the set of p exposures observed from an individual source $i \in (1,2,\ldots,n)$. The two mixture model components are characterized by: (1) π_h, describing the probability that an individual source reflects the shared behaviors of latent cluster h; (2) $f(\boldsymbol{X}_i|\boldsymbol{\theta}_h)$, the joint probability distribution for the set of p exposures in latent cluster h, where $\boldsymbol{\theta}_h$, denote the set of parameters describing the cluster-specific distribution.

$$f(\boldsymbol{X}_i) = \sum_{h=1}^{K} \pi_h f(\boldsymbol{X_i}|\boldsymbol{\theta}_h)$$

Typically, the number of latent clusters is not known *a priori*. Researchers may account for this by either fitting multiple models varying the number of K clusters and evaluating model fitness based on a criterion. More data-driven approaches may overfit the model with an infinitely large K^+ components and assume the number of latent clusters, K, will fill during estimation.

To date, over 130 R packages are available for mixture models and more than half of these packages ($n=73$) are limited to univariate distributions that do not allow the analysis of multiple exposures. Of the remaining packages, many of the mixture models are fit using an expectation-maximization (EM) algorithm or an extension thereof. In the analysis of high-dimensional multivariate exposures, the EM algorithm can be restrictive with slow convergence, sensitive to initial values, and compromise parameter stability, particularly for the Hessian matrix when working within the precision limitations of the R software [13, 14]. R packages have circumvented some of the computational limitations by either adapting the EM algorithm for hybrid approaches with other sampling approaches or porting code into C/++ to run more efficiently and effectively [15–18]. Bayesian estimation offers an attractive alternative for high-dimensional exposures due to its flexibility in parameter estimation, model assumptions and stability. Through the solicitation of priors, one can incorporate prior knowledge into the estimation as well as allow for complete naivety with noninformative priors, providing data-driven inference. Further, Bayesian analysis of sparse multivariate exposures can produce stable estimates within a probabilistic framework, as data dimensionality and complexity increase.

This review aims to provide an overview of currently available R packages available on Comprehensive R Archive Network (CRAN) that offer mixture models for multivariate exposure data, under a Bayesian framework. Section 2 briefly details the search criteria. In Sect. 3, the packages currently available based on their data structures are described. In Sect. 4, we discuss current applications of these packages for the analysis of environmental exposures. Section 5 we discuss how these packages could be more easily adopted by environmental researchers. Section 6 concludes with final thoughts on future research directions.

Search Criteria

A CRAN task view has already been completed on packages available that offer clustering or mixture models [19]. This was later expanded to the full list of CRAN packages, using the following keywords: “model-based clustering”, “mixture models”, “mixtures” and “latent class”. This resulted in a list of 182 CRAN R packages. Each package was then examined to determine if the models and functions available in the package fell within the scope of model-based clustering. This reduced the list of R packages to 133. Of that list, 61 packages contained functions to analyze multivariate exposures Packages that did not allow more than one exposure to be clustered at a time ($n=72$) and packages that only implemented a frequentist estimation procedure (e.g. Expectation-Maximization) were excluded ($n=33$). Packages that were removed from CRAN at the time of submission were also excluded. After applying all inclusion and exclusion criteria, a total of 21 packages remained for discussion.

Bayesian Model-based Clustering R Packages

Table 1 provides an overview of the packages discussed in this review. Seven packages were developed to implement mixture models already established in the literature. The remaining packages are developed in conjunction with new extensions or estimation techniques to the mixture model framework. Most of the packages utilize an unsupervised clustering approach, except for two (*PreMiuM* and *lcra*). The packages are split between normalized and discretized measures. Eight packages provide clustering exclusively for continuous data assuming a multivariate Gaussian distribution. Six packages provide clustering exclusively for discretized data assuming either a Bernoulli or multinomial (categorical) distribution. Two packages provide clustering

Table 1 Brief summary of R packages offering Bayesian mixture models

Package	Descriptor	Cluster Count Determination	Additional features	Applications
Angular/Directional				
BAMBI	Bivariate angular mixture models	2	Bivariate wrapped normal or von mises distributions	Sensorimotor adaptation, motion direction in human vision cortex [20, 21]
Binary				
BayesLCA	An R package for Bayesian latent class analysis	AICM, BICM, DIC		COVID seroprevalence [22], metabolic syndrome components [23]
BayesBinMix	Bayesian Estimation of mixtures of multivariate Bernoulli distribution	Allocation sampler		Allergenic molecules [24]
ddtlcm	Latent class analysis with Dirichlet diffusion tree process prior	WAIC	Best for small-sized dataset ($n < 500$)	Dietary intake [9]
slcm	Sparse latent class model for cognitive diagnosis	Fixed; likelihood criterion	Sparsity induced via variable selection	Cognitive diagnosis [25]
Categorical				
lcra	Bayesian joint latent class and regression models	Variational Bayes	Profile regression; Variable selection available	Acculturation [26]
VICatMix	Variational mixture models for clustering categorical data	Variational Bayes		
Continuous				
IMIFA	Infinite mixtures of infinite factor analyzers and related models	Pitman-Yor	Varied number of factors per cluster via multiplicative gamma process	
Dirichletprocess	Build Dirichlet process objects for Bayesian modeling	Dirichlet process		Biomass in forests, water quality [27, 28]
Fabmix	Overfitting Bayesian mixtures of factor analyzers with parsimonious covariance and unknown number of components	Overfitted Dirichlet	Fixed number of factors per cluster	
BNPmix	Bayesian nonparametric mixture models	Pitman-Yor		
bpgmm	Bayesian model selection approach for parsimonious Gaussian mixture models	RJ-MCMC		
dppmix	Determinantal point process mixture models	Determinantal point process		
Batchmix	Semi-supervised Bayesian mixture models incorporating batch correction	Observed likelihood, complete likelihood, BIC	Adjustment for sample collection variability	Viral surveillance [29]
COMIX	Coarsened mixtures of hierarchical skew kernels	Dirichlet process	Skewed data	Flow cytometry [30]
Continuous/discrete (mixed)				
Mixtools	Tools for analyzing finite mixture models	MFM-Uniform	Mixture of regression only Bayesian estimation option	Air pollutant, volatile organic compounds (VOCs) [31, 32]
PReMiuM	Dirichlet process Bayesian clustering profile regression	Dirichlet process	Profile regression	Air pollution, food choice, social determinants, PFAS [33–36]
Telescope	Bayesian mixtures with an unknown number of components	Telescoping sampler		Imaging data, pain symptoms [37, 38]
Longitudinal exposures				
mixAK	Multivariate normal mixture models and mixtures of generalized linear mixed models including model-based clustering	Fixed; AIC, BIC, penalized expected devance	Can accommodate censored data, missing values allowed	Clinical biomarkers, microbial resistance monitoring [39, 40]
BCCLong	Bayesian consensus clustering for multiple longitudinal features	fixed	Breaks global clustering assumption, irregular time points allowed, mixed data types allowed	Patient subtyping [41]

Table 1 (continued)

Package	Descriptor	Cluster Count Determination	Additional features	Applications
Partial ranked data				
PLMIX	Bayesian analysis of finite mixtures of Plackett-Luce models	DIC1,DIC2,AICM, BPIC1, BPIC2, BICM1, BICM2		Clinical pathology [42]

Acronyms: AICM, Akaike Information Criterion for Markov Chain Monte Carlo; BICM, Bayesian information criterion for Markov chain Monte Carlo; DIC, Deviance information criterion; WAIC, Watanabe-Akaike information criterion; RJ-MCMC, Reversible jump Markov chain Monte Carlo; MFM, Mixture of finite mixtures; BPIC, Bayesian predictive information criterion; PFAS, per- and polyfluroalkyl substances.

for longitudinal exposures of mixed data via a generalized linear mixed effects model. Three packages provide clustering for either type (normal or discrete). One package provides clustering for partial ranked data, and one package provides clustering for directional or angular data.

Normalized Exposure Measures

Bayesian model-based clustering R packages that handle multivariate continuous data are limited to a multivariate Gaussian distribution. This means any continuous measures collected for analysis will need to be transformed accordingly to fit a normal distribution. Once normalized any of the ten packages are available to suit the needs of your set of continuous exposures.

The *batchmix* package fits exposures to a finite mixture model with an additional correction feature to adjust for variation between different batch samples that may have been obtained during data collection. These differences could be due to different processing procedures, reagents used, or other lab-variations. If one finds that the distribution has heavier tails compared to what you see with a normal distribution, the package also offers a multivariate t-distribution. The package does not allow a mix of normal and t-distributed data. It is assumed the full set of exposures assume the same multivariate distribution. Another limitation of the package is that under the finite case, if the number of components is not known in advance, multiple models will need to be run and evaluated *post hoc* using an information criterion (e.g. BIC).

To avoid the burden of running multiple models to determine best fit, the remaining packages rely on Bayesian nonparametric priors to determine the number of components and clusters best suited for the data. *COMIX* and *dirichletprocess* estimate the number of clusters using a Dirichlet process [43]. This prior assumes an infinite number of clusters and allows clusters to fill as more data points are introduced. The number of clusters identified in a mixture model with a Dirichlet process prior grows logarithmically with the data, which can be overwhelming with large, high-dimensional data. *COMIX* offers additional flexibility allowing for adjustments if the exposure data happen to be skewed. *Fabmix* package uses an overfitted finite mixture, where the number of components is set to an exceedingly large number and is asymptotically equivalent to the Dirichlet process [44]. The *BNPmix* and *IMIFA* use a Pitman-Yor prior to determine the number of components [45, 46]. In some scenarios, this is a more attractive alternative to the Dirichlet process when smaller-sized clusters are expected in your data, and a less conservative rate of cluster growth is needed as the size of data increases.

Pitman-Yor and Dirichlet processes are the two most commonly used priors for Bayesian nonparametric mixture models. Yet, inconsistencies have been noted in properly estimating the true number of clusters [47]. This can often be mitigated by imposing a threshold size for cluster importance or introducing a hyperprior on the Dirichlet process concentration parameter controlling the rate of cluster growth in the model [48, 49]. Another alternative to determine the number of clusters is the reversible jump Markov chain monte carlo (RJ-MCMC), implemented in the *bpgmm* package [50]. Here, the model proposes to increase or decrease the number of current clusters at each iteration of the sampling algorithm. The decision to accept or reject this proposal is determined by a probability ratio derived from the likelihood of the current and proposed model fit.

Cluster separation can be challenging with continuous exposures where the variation of exposure variables can overlap between the latent subgroups in practice [51, 52]. The *dppmix* package utilizes a Determinantal Point Process (DPP) prior to induce repulsion between two clusters that may have too much noise or overlap in data coverage, resulting in well-separated clusters which improves identifiability [51].

Discrete Exposure Measures

Discrete exposures are typically seen via self-report in questionnaires or survey studies. Four packages are tailored exclusively for binary data. *BayesLCA* relies on the standard finite mixture model setup, where the number of clusters is preset and the exposure data is binary [53]. Appropriate number of clusters are determined through *post hoc* testing, where an information criterion is applied. *BayesBinMix*

adds a prior on the number of clusters, utilizing an allocation sampler to determine the appropriate number of clusters dynamically [54, 55]. The remaining two packages offer various extensions to the standard latent class model. The *ddtlcm* package is best tailored for small-sized sample populations. It fits a Dirichlet Diffusion tree prior over the number of classes to distinguish exposure patterns that may share similar expressions for a subset of the exposure variables analyzed [10]. The *slcm* package implements a sparse latent class analysis for binary exposure data [25]. Sparsity is induced by coefficients of a generalized linear model, where only exposure variables that are important in characterizing the latent class pattern are given nonzero estimates.

VICatMix and *lcra* packages exclusively handle categorical exposures. The *VICatMix* package improves computational efficiency during estimation through the implementation of variational Bayes inference. Variational Bayes is an alternative estimation approach to the typical MCMC, where instead of estimating parameters directly from a complex posterior distribution, it instead estimates parameters from a distribution that closely resembles the target posterior but is easier to draw and make optimizations [56]. The package also offers an option for variable selection to identify which variables are important in best characterizing exposure patterns. Like the *slcm* package, *lcra* integrates a generalized linear model (GLM) into the mixture model framework, offering users a supervised clustering technique. Here, a joint latent class model is estimated where the latent cluster assignment is a covariate in the GLM, still accounting for cluster assignment uncertainty derived from the probabilistic nature of the mixture model [26].

A few packages include mixture models that can accommodate both discrete and continuous exposures in the same model. We refer to them here as mixed-data mixture models. The 'telescope' package, which has the flexibility of analyzing a mixture of finite mixtures (MFM) of multivariate binary, Gaussian, categorical, or latent class models [57]. Users have the option of sampling from a static mixture of finite mixtures, where the prior on the number of components is fixed, or a dynamic MFM, where the prior on the number of components is inversely proportional to the number currently fit in the model at a given iteration. This package may be useful for those who find the process of parameter tuning of the RJ-MCMC sampler burdensome with each expansion of the model and parameter space. Alternatively, the telescoping sampler works within a fixed space, where both the number of clusters K^+ and components K update simultaneously within that space [57].

The *mixtools* package offers a diverse toolkit of distributions and options for finite mixture models, but for the purposes of this review, we focus primarily on the mixture of a regression option, as it was the only model in the package estimated under a Bayesian construction [58]. As a regression model, both continuous and discrete exposures can be included in the model. The number of components is again assumed to be unknown and determined with a randomized uniform prior on the number of components. The *PreMiuM* package offers the best of both worlds, with options to fit both an unsupervised and supervised Dirichlet Process mixture model [59]. In the unsupervised setting, exposure data can be continuous, discrete, or a combination of both. In the supervised setting, cluster membership is treated as a covariate in a regression model (e.g. profile regression). Outcome response can handle binary, categorical, count (Poisson), normal, and time-to-event (survival) data. Most recently, the package has been updated to also account for longitudinal responses. A variable selection option is also available to determine which exposure variables drive pattern characterization. Users can select among three different sampling strategies for the Dirichlet process: (1) a truncated sampler, where the infinite mixture model is truncated to a maximum of C components that exceeds the true number of clusters (or filled components), similar to an overfitted mixture model; (2) slice dependent sampler (Walker, 2007), where the slices are dependent on the current component weights, and (3) slice-independent sampler, based on Kalli et al. [60], where slices are independent of current weight estimates. The latter option is often preferred for its improved mixing and computational efficiency.

Longitudinal Exposures

Recently, more models have been developed to analyze multiple exposures jointly over time. Mixture models are ideal for this analysis as they can account for correlation that exists within and across different exposures. To date, there are three packages available to analyze multivariate longitudinal exposures. *BClustLong* clusters across coefficients of a linear mixed effects model [61]. The number of clusters is determined by a Dirichlet process prior, and high-dimensionality and correlation of regression coefficients are controlled for using an infinite sparse factor analysis model. The package does require exposures to be normalized with no missing data points. Due to dimensionality and scalability concerns, only a few time points are currently feasible. *BCCLong* implements Bayesian consensus clustering for longitudinal data. Bayesian consensus clustering is a technique that fits a mixture model from multiple sources. Here, each individual source assumes its own local clustering structure. An overall clustering structure is then defined based on similarities shared across each of those individual local clusters [41]. Translated to the longitudinal setting, each longitudinal exposure is fit to a generalized linear mixed effects model. This generalized model allows the inclusion of a mixture of data types (i.e.,

continuous, discrete, count). The number of clusters is finite, requiring multiple models to be run and evaluated for based on a statistical criterion, which can be computationally burdensome, depending on the size of the data. Like *BCCLong*, *MixAK* also models each exposure with a generalized linear mixed effects model, allowing mixed data types [62]. Time points are not required to be equally spaced. This model is also finite, requiring the number of appropriate clusters to be determined with statistical criterion (e.g. penalized expected deviance). Scalability is still a major concern in the analysis of joint longitudinal exposures. Too many exposures can lead to convergence issues and extensive computational burden. Still, analysis of a few variables or a small set of time points can be quite effective.

Other Exposure Types

BAMBI package offers a mixture model to determine patterns from angular or directional data. Data are fit with either a bivariate wrapped normal or a bivariate von mises distribution [63]. While limited to only two exposures, it is suitable for data found in meteorology, geoscience, and bioinformatics. The *PLMIX* package offers a mixture model for partial ranked data. Both packages use a finite mixture model framework requiring multiple testing to determine the appropriate number of clusters.

Applications in Environmental Health Research

Each of these packages offer unique features and extensions of the standard mixture model that are ideal methods to derive underlying patterns of multivariate exposures shared by subgroups in a population. Yet, the application of these packages is grossly underutilized in the environmental health literature. Only eight packages have been utilized to examine environmental exposures in the literature. Two packages (*ddtlcm* and *lcra*) were introduced in the literature with an environmental health context. The *ddtlcm* relied on a new method to improve identification and characterization of dietary exposure patterns in small-sized populations [9]. *Lcra* package was introduced to identify acculturation behavior patterns as predictors of depression [26]. The remaining six packages (*BayesBinMix*, *BayesLCA*, *dirichletprocess*, *mixAK*, *mixtools*, and *PreMiuM*) have been cited beyond their introductory papers with a wide variety of environmental health applications including microbial monitoring, chemical mixtures, biomarkers, social determinants, and measures of air and water quality [22, 24, 28, 31–33, 35, 36, 64, 65]. Five packages (*BAMBI*, *Batchmix*, *BCCLong*, *COMIX slcm*) have been applied in real data settings, but none within the scope of environmental health research. Seven packages (*BNPmix*, *bpgmm*, *dppmix*, *PLmix*, *fabmix*, *IMIFA*, *VICatMix)* were introduced in literature using benchmark data only. The underlying methods have motivated the development of more advanced mixture models, but those methods have not yet been made available on CRAN, limiting its application to environmental health and other research areas [66, 67].

Discussion

Questionnaires and surveys can assess the presence, absence or frequency of a wide variety of environmental exposures (diet intake, occupational hazards, product use, multimorbidity, air quality, etc.). Identifying underlying exposure patterns within these observed data can be both complex and muddled. This review provided an overview of the existing packages currently available on CRAN to cluster a large set of observations and exposures under a Bayesian framework. Bayesian mixture models provide researchers with the necessary tools to balance the demands of large data, model flexibility, and computational efficiency.

Most of the packages discussed in this review were introduced on CRAN during the last 10 years, but few have found wide applicability in real data settings. Greater awareness of the existence of these packages as well as an understanding of how best to leverage these packages for the analysis of high-dimensional environmental exposures could increase utilization of model-based clustering approaches and improve understanding of the complex relationship of the environment and population health.

Vignette tutorials outside of the standard CRAN documentation are available for 18 of the 21 packages (See Supplementary Table 1 for list of direct links to available tutorials). Most of these vignettes are made available on the developer's GitHub repository. This can be a challenge for broad adoption of the methods as the onus for promotion and awareness to the public falls solely on the developer. Vignettes made available through open access software journals have the advantage of providing developers with useful feedback to increase usability, validity, and reproducibility through the peer-review process. They also offer developers a broader audience through the journal's readership which could increase citation impact and motivate utility across different disciplines. For example, *BayesBinMix*, *BayesLCA*, *dirichletprocess*, *mixAK*, and *PreMiuM* have published tutorials and vignettes in peer-reviewed journals and have been widely cited with applications in environmental health. *Mixtools* is the oldest and most cited R package for mixture models, boasting over 1000 citations and several environmental health applications.

Newer packages, released in the last five years, have potential to hopefully gain popularity with time, but without publications in software or interdisciplinary journals, they risk being underutilized. *Batchmix*, released in 2022 has no published applications outside of its introductory paper. The package was designed to adjust for variation in batches of genomic samples. This could be of great benefit in environmental studies where labs collecting and analyzing chemical exposures (e.g. phthalates, PFAS, VOCs) need to account for differences between collection procedures, times, and sources. *COMIX* and *BCCLong*, were released in 2022 and 2025 respectively. They were introduced for the analysis of biological markers for disease and cell subtyping but could easily be used to analyze a set of exposures that are highly skewed or contain repeated measures. Other packages, such as *ddtlcm* could have broad applicability for any set of binary exposures, where derived exposure patterns may be weakly identifiable, due to small sample size [9].

As the field continues to grow and expand with the increase in data availability and accessibility, research has shifted towards the understanding of the intersection of multiple sources of exposure sets. This type of analysis will rely on a diverse set of data collection sources (e.g. questionnaires/surveys, lab measures, external monitor readings). Packages, such as *mixtools*, *telescope*,and *PReMiuM* are well-equipped to handle mixed data types and handle this type of analysis.

Some methods have already been applied to environmental science research, but not under a Bayesian framework. For example, mixture models using factor analyzers have been used for the analysis of air pollutant levels under a frequentist setting using the widely popular EM estimation approach [68]. IMIFA whose package vignettes are housed on a GitHub repository could see greater use to researchers who are interested in applying a model where the number of factors and mixture model components can vary with the data, providing more insight and increased data flexibility.

Analysis under a Bayesian framework may still be unfamiliar to non-statistical researchers. Implementation of these packages requires a deeper understanding of the mechanics for parameter estimation, prior specification and appropriate distribution assumptions. Guidance on how best to use these packages could increase usability and provide researchers with clear, interpretable results that improve pattern characterization and reduce bias.

Conclusion and Future Directions

This review focused on packages currently available on CRAN. More packages have been developed and are available on public repositories like GitHub. However, given our focus on available CRAN packages, discussion of those extensions are beyond the scope of this review. While no model is perfect, many of the packages discussed here are quite useful and could benefit from broader use in the environmental health field. Scalability continues to be a limitation with many of these methods, as the number of potential exposures to analyze in a single study increases over time and requires more resources. With a greater shift towards big data and efficiently managing the computational burden of analysis, these limitations can be diminished over time. Greater awareness of these packages and increased interdisciplinary collaborations could help with the dissemination and promotion of these software tools across new fields.

Key References

- Grün B. CRAN Task View: Cluster Analysis & Finite Mixture Models [Internet]. Comprehensive R Archive Network (CRAN); 2025 [cited 2026 Jan 1]. https://CRAN.R-project.org/view=Cluster. Accessed 1 Jan 2026.
  - List of CRAN R packages available for model-based clustering.
- Rousseau J, Mengersen K. Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models. J R Stat Soc Ser B Stat Methodol [Internet]. 2011 [cited 2026 Jan 1]; 73:689–710. https://doi.org/10.1111/j.1467-9868.2011.00781.x.
  - Paper describing the overfitted Dirichlet prior to determine the number of clusters in a mixture model.
- Frühwirth-Schnatter S, Malsiner-Walli G, Grün B. Generalized Mixtures of Finite Mixtures and Telescoping Sampling. Bayesian Anal [Internet]. 2021 [cited 2026 Jan 1]; 16. https://doi.org/10.1214/21-BA1294.
  - Paper describing the mixture of finite mixture model approach and telescoping sampler algorithm to determine number of components in a mixture model.
- Müller P. Bayesian Nonparametric Mixture Models. Handbook of Mixture Analysis. Chapman and Hall/CRC; 2019.
  - Book overviewing commonly used Bayesian nonparametric mixture models (e.g. Dirichlet process, Pitman-Yor) and estimation approach.

- Castillo I. Bayesian Nonparametric Statistics: École d'Été de Probabilités de Saint-Flour LI - 2023 [Internet]. Cham: Springer Nature Switzerland; 2024 [cited 2026 Jan 1]. https://doi.org/10.1007/978-3-031-74035-0.
    - Book overviewing Bayesian estimation approaches (e.g. MCMC, Variational Bayes) for a mixture model of multivariate data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40572-026-00532-y.

Acknowledgements This research was supported by National Heart Lung and Blood Institute (K01HL166442).

Author Contributions Briana Stephenson conceptualized the idea for the article. Yiran Fu performed the initial search of the literature. Briana Stephenson finalized and reviewed the literature search and drafted the manuscript. All authors have read and approved the manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing Interests The authors declare no competing interests.

References

1. Sabbarese C, Feola ML, Ambrosino F, Roca V, D'Onofrio A, La Verde G, et al. A preliminary study of the characteristics of radon data from indoor environments and building materials in the Campania Region using PCA and K-means statistical analyses. Environments. 2022;9:82. https://doi.org/10.3390/environments9070082
2. Beddows DCS, Dall'Osto M, Harrison RM. Cluster Analysis of Rural, Urban, and Curbside Atmospheric Particle Size Data. Environ Sci Technol [Internet]. 2009;43:4694–700. https://doi.org/10.1021/es803121t. [cited 2026 Jan 1].
3. Shi W, Zeng W. Application of k-means clustering to environmental risk zoning of the chemical industrial area. Front Environ Sci Eng [Internet]. 2014;8:117–27. https://doi.org/10.1007/s11783-013-0581-5. [cited 2026 Jan 1].
4. Riches NO, Gouripeddi R, Payan-Medina A, Facelli JC. K-means cluster analysis of cooperative effects of CO, NO2, O3, PM2.5, PM10, and SO2 on incidence of type 2 diabetes mellitus in the US. Environ Res. 2022;212:113259. https://doi.org/10.1016/j.envres.2022.113259.
5. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–31. https://doi.org/10.1198/016214502760047131
6. Greve B, Pigeot I, Huybrechts I, Pala V, Börnhorst C. A comparison of heuristic and model-based clustering methods for dietary pattern analysis. Public Health Nutr [Internet]. 2016;19:255–64. https://doi.org/10.1017/S1368980014003243. [cited 2026 Jan 1].
7. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C et al. ggplot2: create elegant data visualisations using the grammar of graphics [Internet]. 2026 [cited 2026 Feb 24]. https://cran.r-project.org/web/packages/ggplot2/index.html. Accessed 24 Feb 2026.
8. Stephenson BJK, Willett WC. Racial and ethnic heterogeneity in diets of low-income adult females in the United States: results from National Health and Nutrition Examination Surveys from 2011 to 2018. Am J Clin Nutr. 2023;117:625–34. https://doi.org/10.1016/j.ajcnut.2023.01.008.
9. Li M, Stephenson B, Wu Z. Tree-regularized Bayesian latent class analysis for improving weakly separated dietary pattern subtyping in small-sized subpopulations. Annals Appl Stat. [Internet]. Institute of Mathematical Statistics; 2025 [cited 2026 Jan 1];19:3003–22. https://doi.org/10.1214/25-AOAS2067
10. Li M, Wu B, Stephenson B, Wu Z. ddtlcm: An R package for overcoming weak separation in Bayesian latent class analysis via tree-regularization. J Open Source Softw [Internet]. 2024;9:6220. https://doi.org/10.21105/joss.06220. [cited 2026 Jan 1].
11. Swanson MB, Davis GA, Kincaid LE, Schultz TW, Bartmess JE, Jones SL, et al. A screening method for ranking and scoring chemicals by potential human health and environmental impacts. Environ Toxicol Chem [Internet]. 1997;16:372–83. https://doi.org/10.1002/etc.5620160237. [cited 2026 Jan 1].
12. French McCay DP, Isaji T. Evaluation of the consequences of chemical spills using modeling: chemicals used in deepwater oil and gas operations. Environ Model Softw. 2004;19:629–44. https://doi.org/10.1016/j.envsoft.2003.08.003
13. Karlis D, Xekalaki E. Choosing Initial Values for the EM Algorithm for Finite Mixtures [Internet]. Rochester, NY: Social Science Research Network; 2007 [cited 2026 Jan 1]. https://papers.ssrn.com/abstract=1029320. Accessed 1 Jan 2026.
14. Ma J, Fu S. On the correct convergence of the EM algorithm for Gaussian mixtures. Pattern Recogn. 2005;38:2602–11. https://doi.org/10.1016/j.patcog.2005.03.010
15. McParland D, Gormley IC. Model based clustering for mixed data: clustMD. Adv Data Anal Classif. 2016;10:155–69. https://doi.org/10.1007/s11634-016-0238-x
16. Drachal K. dynmix: an R package for the estimation of dynamic finite mixtures. SoftwareX. 2023;22:101388. https://doi.org/10.1016/j.softx.2023.101388
17. McNicholas PD, ElSherbiny A, McDaid AF, Murphy TB. pgmm: Parsimonious Gaussian mixture models [Internet]. 2011 [cited 2026 Jan 1]. p. 1.2.8. https://doi.org/10.32614/CRAN.package.pgmm
18. Nagode M, Panić B, Klemenc J, Oman S. Fault detection and classification with the rebmix R package. Computers Industrial Eng [Internet]. 2023;185:109628. https://doi.org/10.1016/j.cie.2023.109628. [cited 2026 Jan 1].
19. Grün B. Jan. CRAN Task View: Cluster Analysis & Finite Mixture Models [Internet]. Comprehensive R Archive Network (CRAN); 2025 [cited 2026 Jan 1]. https://CRAN.R-project.org/view=Cluster. Accessed 1 2026.
20. Tsay JS, Asmerian H, Germine LT, Wilmer J, Ivry RB, Nakayama K. Large-scale citizen science reveals predictors of sensorimotor adaptation. Nat Hum Behav [Internet] Nat Publishing Group.

2024;8:510–25. https://doi.org/10.1038/s41562-023-01798-0. [cited 2026 Jan 2].
21. Chetverikov A, Jehee JFM. Motion direction is represented as a bimodal probability distribution in the human visual cortex. Nat Commun [Internet] Nat Publishing Group. 2023;14:7634. https://doi.org/10.1038/s41467-023-43251-w. [cited 2026 Jan 2].
22. Tuite AR, Fisman D, Abe KT, Rathod B, Pasculescu A, Colwill K, et al. Estimating SARS-CoV-2 Seroprevalence in Canadian blood donors. April 2020 to March 2021: improving accuracy with multiple assays. Microbiology Spectrum [Internet]. Volume 10. American Society for Microbiology; 2022. pp. e02563–21. [cited 2026 Jan 2];. https://doi.org/10.1128/spectrum.02563-21.
23. Barnett A, Martino E, Knibbs LD, Shaw JE, Dunstan DW, Magliano DJ, et al. The neighbourhood environment and profiles of the metabolic syndrome. Environ Health [Internet]. 2022;21:80. https://doi.org/10.1186/s12940-022-00894-4. [cited 2026 Jan 2].
24. Roberts G, Fontanella S, Selby A, Howard R, Filippi S, Hedlin G, et al. Connectivity patterns between multiple allergen specific IgE antibodies and their association with severe asthma. J Allergy Clin Immunol. 2020;146:821–30. https://doi.org/10.1016/j.jaci.2020.02.031.
25. Chen Y, Culpepper S, Liang FA. Sparse latent class model for cognitive diagnosis. Psychometrika. 2020;85:121–53. https://doi.org/10.1007/s11336-019-09693-2
26. Elliott MR, Zhao Z, Mukherjee B, Kanaya A, Needham BL. Methods to account for uncertainty in latent class assignments when using latent classes as predictors in regression models, with application to acculturation strategy measures. Epidemiology. 2020;31:194–204. https://doi.org/10.1097/EDE.0000000000001139
27. May PB, Schlund M, Armston J, Kotowska MM, Brambach F, Wenzel A, et al. Mapping aboveground biomass in Indonesian lowland forests using GEDI and hierarchical models. Remote Sens Environ. 2024;313:114384. https://doi.org/10.1016/j.rse.2024.114384
28. Wolfgang Seis P, Pouault D, Steffelbauer P. 2024 [cited 2026 Feb 23];252:121186. https://doi.org/10.1016/j.watres.2024.121186
29. Kain MP, Epstein JH, Ross N. Rethinking statistical approaches for serological data analysis for viral surveillance. J Virological Methods [Internet]. 2025;335:115149. https://doi.org/10.1016/j.jviromet.2025.115149. [cited 2026 Jan 2].
30. Lin L, Hejblum BP. Bayesian mixture models for cytometry data analysis. WIREs Computational Statistics [Internet]. 2021 [cited 2026 Jan 2];13:e1535. https://doi.org/10.1002/wics.1535
31. Li S, Batterman S, Su F-C, Mukherjee B. Addressing extrema and censoring in pollutant and exposure data using mixture of normal distributions. Atmospheric Environ [Internet]. 2013;77:464–73. https://doi.org/10.1016/j.atmosenv.2013.05.004. [cited 2026 Feb 22].
32. Batterman S, Su F-C, Li S, Mukherjee B, Jia C, HEI Health Review Committee. Personal exposure to mixtures of volatile organic compounds: modeling and further analysis of the RIOPA data. Res Rep Health Eff Inst Health Eff Inst; 2014;3–63.
33. Cai Y, Hansell AL, Granell R, Blangiardo M, Zottoli M, Fecht D, et al. Prenatal, Early-Life, and Childhood Exposure to Air Pollution and Lung Function: The ALSPAC Cohort. Am J Respir Crit Care Med. 2020;202:112–23. https://doi.org/10.1164/rccm.201902-0286OC.
34. Briazu RA, Bell L, Dodd GF, Blackburn S, Massri C, Chang B, et al. The effectiveness of personalised food choice advice tailored to an individual's socio-demographic, cognitive characteristics, and sensory preferences. Appetite [Internet]. 2024;201:107600. https://doi.org/10.1016/j.appet.2024.107600. [cited 2026 Jan 2].
35. Keenan K, Papathomas M, Mshana SE, Asiimwe B, Kiiru J, Lynch AG, et al. Intersecting social and environmental determinants of multidrug-resistant urinary tract infections in East Africa beyond antibiotic use. Nat Commun [Internet] Nat Publishing Group. 2024;15:9418. https://doi.org/10.1038/s41467-024-53253-x. [cited 2026 Jan 1].
36. Griffin EK, Aristizabal-Henao J, Timshina A, Ditz HL, Camacho CG, Da Silva BF et al. Assessment of per- and polyfluoroalkyl substances (PFAS) in the Indian River Lagoon and Atlantic coast of Brevard County, FL, reveals distinct spatial clusters. Chemosphere [Internet]. 2022 [cited 2026 Jan 1];301:134478. https://doi.org/10.1016/j.chemosphere.2022.134478
37. D'Angelo L, Canale A, Yu Z, Guindani M. Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. Biometrics [Internet]. 2023;79:1370–82. https://doi.org/10.1111/biom.13626. [cited 2026 Jan 2].
38. Malsiner-Walli G, Grün B, Frühwirth-Schnatter S. Without pain – clustering categorical data using a Bayesian mixture of finite mixtures of latent class analysis models. Adv Data Anal Classif [Internet]. 2025 [cited 2026 Jan 2]; https://doi.org/10.1007/s11634-025-00640-x
39. Hughes DM, Komárek A, Czanner G, Garcia-Fiñana M. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. Volume 27. SAGE Publications Ltd STM; 2018. pp. 2060–80. [cited 2026 Jan 2];. https://doi.org/10.1177/0962280216674496. Stat Methods Med Res [Internet].
40. Jaspers S, Komárek A, Aerts M. Bayesian estimation of multivariate normal mixtures with covariate-dependent mixing weights, with an application in antimicrobial resistance monitoring. Biometrical J [Internet]. 2018;60:7–19. https://doi.org/10.1002/bimj.201600253. [cited 2026 Jan 2].
41. Lu Z, Lou W. Bayesian consensus clustering for multivariate longitudinal data. Stat Med. 2022;41:108–27. https://doi.org/10.1002/sim.9225.
42. Katz-Summercorn AC, Jammula S, Frangou A, Peneva I, O'Donovan M, Tripathi M, et al. Multi-omic cross-sectional cohort study of pre-malignant Barrett's esophagus reveals early structural variation and retrotransposon activity. Nat Commun [Internet] Nat Publishing Group. 2022;13:1407. https://doi.org/10.1038/s41467-022-28237-4. [cited 2026 Jan 2].
43. Ross GJ, Markwick D. dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models [Internet]. https://cloud.r-project.org/web/packages/dirichletprocess/vignettes/dirichletprocess.pdf
44. Rousseau J, Mengersen K. Asymptotic behaviour of the posterior distribution in overfitted mixture models. J R Stat Soc Ser B Stat Methodol. 2011;73:689–710. . https://doi.org/10.1111/j.1467-9868.2011.00781.x
45. Corradin R, Canale A, Nipoti B, BNPmix. An *R* Package for Bayesian Nonparametric Modeling via Pitman-Yor Mixtures. J Stat Soft [Internet]. 2021 [cited 2026 Jan 1];100. https://doi.org/10.18637/jss.v100.i15
46. Murphy K, Viroli C, Gormley IC. Infinite mixtures of infinite factor analysers. Bayesian Anal [Internet]. 2020 [cited 2026 Jan 1];15. https://doi.org/10.1214/19-BA1179
47. Miller JW, Harrison MT. Inconsistency of Pitman-Yor process mixtures for the number of components. J Mach Learn Res 2014;15:3333–70. http://jmlr.org/papers/v15/miller14a.html Accessed 1 Jan 2026
48. Miller JW, Harrison MT. Mixture models with a prior on the number of components. J Am Stat Assoc 113: 340–56. https://doi.org/10.1080/01621459.2016.1255636
49. Ascolani F, Lijoi A, Rebaudo G, Zanella G. Clustering consistency with Dirichlet process mixtures. Biometrika. 2023;110:551–8. https://doi.org/10.1093/biomet/asac051
50. Lu X, Li Y, Love T. On Bayesian analysis of Parsimonious Gaussian mixture models. J Classif. 2021;38:576–93. https://doi.org/10.1007/s00357-021-09391-8
51. Xu Y, Müller P, Telesca D. Bayesian Inference for Latent Biologic Structure With Determinantal Point Processes (DPP). Biometrics

[Internet]. 2016 [cited 2026 Jan 1];72:955–64. https://doi.org/10.1111/biom.12482
52. Müller P. Bayesian Nonparametric mixture models. Handbook of Mixture Analysis. Chapman and Hall/CRC; 2019.
53. White A, Murphy TB. BayesLCA: an R package for Bayesian latent class analysis. J Stat Softw. 2014;61:1–28. https://doi.org/10.18637/jss.v061.i13
54. Nobile A, Fearnside AT. Bayesian finite mixtures with an unknown number of components: the allocation sampler. Stat Comput. 2007;17:147–62. https://doi.org/10.1007/s11222-006-9014-7
55. Papastamoulis P, Rattray M. Jan. BayesBinMix: an R package for model based clustering of multivariate binary data. The R Journal [Internet]. Technische Universität Wien, Institut für Statistik und Wahrscheinlichkeitstheorie; 2017 [cited 2026 Jan 7]; https://research.manchester.ac.uk/en/publications/bayesbinmix-an-r-package-for-model-based-clustering-of-multivaria/. Accessed 7 2026.
56. Castillo I. Bayesian Nonparametric Statistics: École d'Été de Probabilités de Saint-Flour LI – 2023 [Internet]. Cham: Springer Nature Switzerland; 2024. [cited 2026 Jan 1]. https://doi.org/10.1007/978-3-031-74035-0.
57. Frühwirth-Schnatter S, Malsiner-Walli G, Grün B. Generalized mixtures of finite mixtures and telescoping sampling. Bayesian Anal [Internet]. 2021 [cited 2026 Jan 1];16. https://doi.org/10.1214/21-BA1294
58. Benaglia T, Chauveau D, Hunter DR, Young DS. mixtools: an R package for analyzing mixture models. J Stat Softw. 2010;32:1–29. https://doi.org/10.18637/jss.v032.i06
59. Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S. PReMiuM : An *R* Package for Profile Regression Mixture Models Using Dirichlet Processes. J Stat Soft [Internet]. 2015. https://doi.org/10.18637/jss.v064.i07. [cited 2026 Jan 1];64.
60. Kalli M, Griffin JE, Walker SG. Slice sampling mixture models. Stat Comput [Internet]. 2011;21:93–105. https://doi.org/10.1007/s11222-009-9150-y. [cited 2026 Jan 1].
61. Sun J, Herazo-Maya JD, Kaminski N, Zhao H, Warren JL. A Dirichlet process mixture model for clustering longitudinal gene expression data. Stat Med [Internet]. 2017;36:3495–506. https://doi.org/10.1002/sim.7374. [cited 2026 Jan 7].
62. Komárek A, Komárková L. Capabilities of R Package mixAK for Clustering Based on Multivariate Continuous and Discrete Longitudinal Data. Journal of Statistical Software [Internet]. 2014 [cited 2026 Jan 7];59:1–38. https://doi.org/10.18637/jss.v059.i12
63. Chakraborty S, Wong SWK. BAMBI: an R package for fitting bivariate angular mixture models. J Stat Soft [Internet]. 2021 [cited 2026 Jan 1]; 99. https://doi.org/10.18637/jss.v099.i11
64. Pirani M, Best N, Blangiardo M, Liverani S, Atkinson RW, Fuller GW. Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. Environ Int. 2015;79:56–64. https://doi.org/10.1016/j.envint.2015.02.010.
65. LaLonde A, Love T. Using the Seychelles child development study to cluster multiple outcomes into domains to improve estimation of the overall effect of mercury on neurodevelopment. Math Appl [Internet]. 2018;7:53–62. https://doi.org/10.13164/ma.2018.05. [cited 2026 Jan 1].
66. XieF, Xu Y. Bayesian repulsive gaussian mixture model. J Am Stat Assoc 115:187–203. https://doi.org/10.1080/01621459.2018.1537918
67. Frühwirth-Schnatter S, Hosszejni D, Lopes HF. Sparse Bayesian factor analysis when the number of factors is unknown (with discussion). Bayesian Analysis [Internet]. International Society for Bayesian Analysis; 2025 [cited 2026 Jan 1];20:213–344. https://doi.org/10.1214/24-BA1423
68. Maruotti A, Bulla J, Lagona F, Picone M, Martella F. Dynamic mixtures of factor analyzers to characterize multivariate air pollutant exposures. Ann Appl Stat [Internet]. 2017. https://doi.org/10.1214/17-AOAS1049. [cited 2026 Jan 1];11.