The Journal of Open Source Software

# bayes_traj: A Python package for Bayesian trajectory analysis

**James C. Ross** [1,3] **and Tingting Zhao**[2]

**1** Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States of America **2** College of Business, University of Rhode Island, Kingston, RI, United States of America **3** For correspondence, contact jcross186@gmail.com

**DOI:** 10.21105/joss.07323

**Software**

- Review
- Repository
- Archive

**Editor:** Mark A. Jensen

**Reviewers:**

- @gvieralopez
- @gchure

**Submitted:** 18 September 2024
**Published:** 15 April 2025



## Statement of need

Trajectory analysis broadly refers to techniques and modeling paradigms that explain heterogeneity in longitudinal data. These methods identify the most suitable number of subgroups (trajectories) in the data, the distinct patterns of change characterizing each trajectory, and the most likely assignment of study participants to trajectories. Methods of trajectory analysis have been applied to a wide range of fields including psychology, criminology, behavioral research, and epidemiology. (These methods are distinct from those that track people, animals, vehicles, and natural phenomena – also referred to as trajectory analysis – and which have their own dedicated set of techniques and frameworks. See, e.g., (Shenk et al., 2021) and (Viera-López et al., 2023)).

Although trajectory analysis has been applied in multiple domains, the motivation for developing **bayes_traj** has been to improve our understanding of heterogeneity in the context of chronic obstructive pulmonary disease (COPD), a leading cause of death worldwide. Research has shown that there are multiple patterns of lung function development and decline, with some patterns associated with greater risk of developing COPD (Lange et al., 2015). Furthermore, there is a growing recognition that COPD is better conceived of as a multi-faceted syndrome, requiring consideration of other disease facets (such as clinical presentation and structural assessment from medical images) (Lowe et al., 2019). Researchers have applied techniques of trajectory analysis to longitudinal measures of lung function to delineate distinct patterns of progression for further analysis (Agusti & Faner, 2019). Existing trajectory approaches are predominantly frequentist in nature and use maximum likelihood to identify point estimates of unknown parameters. These approaches do not permit incorporation of prior information. Challenges arise when study cohorts lack sufficient longitudinal data characteristics to adequately power frequentist-based trajectory algorithms. Bayesian approaches are well-suited for data-limited scenarios given their ability to incorporate prior knowledge in the model fitting process, though existing Bayesian trajectory approaches use sampling-based inference (i.e., Markov chain Monte Carlo) which can be slow to converge and can suffer from the so-called "label switching" problem (the unidentifiability of the permutation of latent variables). There is thus a need for scalable approaches that can simultaneously model distinct progression patterns across multiple health measures, especially in data-limited scenarios.

## Summary

**bayes_traj** is a Python package for Bayesian trajectory analysis, offering a suite of command-line tools for prior specification, model fitting, and posterior evaluation. **Figure 1** illustrates the key tools and their role within the workflow. The package is domain-agnostic and applicable across various disciplines. It is particularly suited for researchers who require scalable trajectory analysis methods, especially in scenarios where traditional frequentist approaches struggle due

to limited data or the need to incorporate prior knowledge. By providing a scalable Bayesian alternative, **bayes_traj** complements existing tools and broadens the range of methodologies available for trajectory analysis.

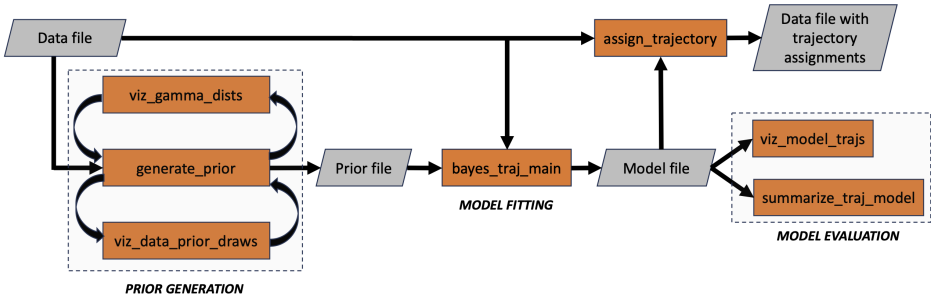


**Figure 1:** Workflow of **bayes_traj** command-line tools (orange). The process begins with an input **data file**, which informs prior specification using the `generate_prior` routine. (`viz_data_prior_draws` and `viz_gamma_dists` provide feedback for prior evaluation.) Model fitting (`bayes_traj_main`) take a prior and input data to perform Bayesian inference. The fitted model is evaluated through visualization (`viz_model_trajs`) and quantitative summary (`summarize_traj_model`). Finally, `assign_trajectory` applies the fitted model to assign individuals to trajectory groups. Each command-line tool supports the `-h` flag for detailed usage instructions.

**bayes_traj** has several distinguishing features:

- Simultaneously models multiple continuous and binary target variables as functions of predictor variables.
- Uses Bayesian nonparametrics (Dirichlet Process mixture modeling) to automatically identify the number of groups in a data set given an estimate of the number of trajectories.
- Makes the assumption that target variables are conditionally independent given trajectory assignments, enabling the algorithm to scale well to multiple targets.
- Performs Bayesian approximate inference using coordinate ascent variational inference, which is fast and scales well to large data sets.
- Independently estimates residual variance posteriors for each trajectory and each target variable
- Allows specification of random effects for continuous target variables using unstructured covariance matrices
- Provides a suite of tools to facilitate prior specification, model visualization, and summary statistic computation.

These features make **bayes_traj** a great fit for investigating COPD heterogeneity, and we have used it in several publications. In an early implementation, we used it to identify disease subtypes using five measures of emphysema computed from medical images (Ross et al., 2016). Later we used it to identify distinct lung function trajectories in one cohort and to then probabilistically assign individuals in another cohort to their most likely trajectory for further analysis (Ross et al., 2018). Recently, we applied **bayes_traj** to multiple measures of lung function in a cohort of middle-aged and older adults, using an informative prior to capture known information about lung function in early adulthood (Ross et al., 2024).

While **bayes_traj** offers several advantages over existing trajectory analysis tools, it also has some limitations. The underlying model assumes conditional independence of target variables given trajectory assignments and predictors, which, although common, may not always hold in real-world data. Additionally, while variational inference is scalable, provides computational efficiency, and is less susceptible to label-switching, it may not capture posterior characteristics as accurately as sampling-based methods. The model also assumes that errors are uncorrelated, a simplification that may not be appropriate for all use cases. Finally, although **bayes_traj** supports both continuous and binary target variables, it does not currently handle count data.

## State of the field

There are numerous approaches to trajectory analysis that make different modeling assumptions and use different inference strategies, and implementations are available in R, SAS, Stata, and MpLus. Van der Nest et al. (Nest et al., 2020), Lu (Lu, 2024), and Lu et al. (Lu et al., 2023) provide excellent reviews of the state of the art. Two broad and commonly used model-based approaches are group-based trajectory modeling (GBTM) and latent class mixed effect modeling (LCMM). GBTM assumes identical trajectories within clusters, while LCMM generalizes this by allowing individual trajectories to deviate from the cluster mean. Zang and Max describe a Bayesian group-based trajectory modeling approach that relies on MCMC for inference with an implementation available in R (Zang & Max, 2022). Other Bayesian approaches to trajectory analysis such as Bayesian mixture modeling (Komárek & Komárková, 2013) and Bayesian consensus clustering (Lock & Dunson, 2013) have implementations in R (Komárek & Komárková, 2014; Tan et al., 2022) and fall within the LCMM category. These methods also rely on MCMC for inference. The model implemented in **bayes_traj** can be considered a Bayesian nonparametric version of LCMM that is capable of modeling multiple longitudinal markers. To our knowledge, **bayes_traj** is the only full-featured Python package for Bayesian nonparametric trajectory analysis that uses variational inference for model fitting across multiple target variables, making it a scalable and versatile tool for researchers across disciplines that complements the collection of existing trajectory analysis approaches.

## Acknowledgements

Special thanks to Fritz Obermeyer for contributing code toward incorporating Pyro probabilistic programming language capabilities into the **bayes_traj** environment as part of a subcontract funded through NIH R01-HL164380. Continued development of **bayes_traj** is supported by the US National Heart, Lung, and Blood Institute (R01-HL164380). **bayes_traj** would not be possible without numerous other open-source Python packages, especially numpy (Harris et al., 2020), scipy (Virtanen et al., 2020), matplotlib (Hunter, 2007), PyTorch (Paszke et al., 2019), and pandas (Reback et al., 2020).

## References

Agusti, A., & Faner, R. (2019). Lung function trajectories in health and disease. *The Lancet Respiratory Medicine*, *7*(4), 358–364. https://doi.org/10.1016/S2213-2600(18)30529-0

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., & others. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Komárek, A., & Komárková, L. (2013). *Clustering for multivariate continuous and discrete longitudinal data*. https://doi.org/10.1214/12-AOAS580

Komárek, A., & Komárková, L. (2014). Capabilities of r package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software*, *59*, 1–38. https://doi.org/10.18637/jss.v059.i12

Lange, P., Celli, B., Agustí, A., Boje Jensen, G., Divo, M., Faner, R., Guerra, S., Marott, J. L., Martinez, F. D., Martinez-Camblor, P., & others. (2015). Lung-function trajectories leading to chronic obstructive pulmonary disease. *New England Journal of Medicine*, *373*(2), 111–122. https://doi.org/10.1056/NEJMoa1411532

Lock, E. F., & Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, *29*(20),

2610–2616. https://doi.org/10.1093/bioinformatics/btt425

Lowe, K. E., Regan, E. A., Anzueto, A., Austin, E., Austin, J. H., Beaty, T. H., Benos, P. V., Benway, C. J., Bhatt, S. P., Bleecker, E. R., & others. (2019). COPDGene® 2019: Redefining the diagnosis of chronic obstructive pulmonary disease. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, *6*(5), 384. https://doi.org/10.15326/jcopdf.6.5.2019.0149

Lu, Z. (2024). Clustering longitudinal data: A review of methods and software packages. *International Statistical Review*. https://doi.org/10.1111/insr.12588

Lu, Z., Ahmadiankalati, M., & Tan, Z. (2023). Joint clustering multiple longitudinal features: A comparison of methods and software packages with practical guidance. *Statistics in Medicine*, *42*(29), 5513–5540. https://doi.org/10.1002/sim.9917

Nest, G. van der, Passos, V. L., Candel, M. J., & Breukelen, G. J. van. (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research*, *43*, 100323. https://doi.org/10.1016/j.alcr.2019.100323

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1912.01703

Reback, J., McKinney, W., Van Den Bossche, J., Augspurger, T., Cloud, P., Klein, A., Hawkins, S., Roeschke, M., Tratner, J., She, C., & others. (2020). Pandas-dev/pandas: Pandas 1.0.5. *Zenodo*. https://doi.org/10.5281/zenodo.3898987

Ross, J. C., Castaldi, P. J., Cho, M. H., Chen, J., Chang, Y., Dy, J. G., Silverman, E. K., Washko, G. R., & Estépar, R. S. J. (2016). A bayesian nonparametric model for disease subtyping: Application to emphysema phenotypes. *IEEE Transactions on Medical Imaging*, *36*(1), 343–354. https://doi.org/10.1109/TMI.2016.2608782

Ross, J. C., Castaldi, P. J., Cho, M. H., Hersh, C. P., Rahaghi, F. N., Sánchez-Ferrero, G. V., Parker, M. M., Litonjua, A. A., Sparrow, D., Dy, J. G., & others. (2018). Longitudinal modeling of lung function trajectories in smokers with and without chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, *198*(8), 1033–1042. https://doi.org/10.1164/rccm.201707-1405OC

Ross, J. C., Estépar, R. S. J., Ash, S., Pistenmaa, C., Han, M., Bhatt, S. P., Bodduluri, S., Sparrow, D., Charbonnier, J.-P., Washko, G. R., & others. (2024). Dysanapsis is differentially related to lung function trajectories with distinct structural and functional patterns in COPD and variable risk for adverse outcomes. *EClinicalMedicine*, *68*. https://doi.org/10.1016/j.eclinm.2023.102408

Shenk, J., Byttner, W., Nambusubramaniyan, S., & Zoeller, A. (2021). Traja: A python toolbox for animal trajectory analysis. *Journal of Open Source Software*, *6*(63), 3202. https://doi.org/10.21105/joss.03202

Tan, Z., Lu, Z., & Shen, C. (2022). *A joint modeling approach for clustering mixed-type multivariate longitudinal data: Application to the CHILD cohort study*. https://doi.org/10.32614/CRAN.package.BCClong

Viera-López, G., Morgado-Vega, J., Reyes, A., Altshuler, E., Almeida-Cruz, Y., & Manganini, G. (2023). Pactus: A python framework for trajectory classification. *Journal of Open Source Software*, *8*(89), 5738. https://doi.org/10.21105/joss.05738

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D.,

Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Zang, E., & Max, J. T. (2022). Bayesian estimation and model selection in group-based trajectory models. *Psychological Methods*, *27*(3), 347. https://doi.org/10.1037/met0000359