The Journal of Open Source Software

Python KM: Kramers–Moyal coefficients for stochastic processes

Leonardo Rydin Gorjão[1, 2, 3, 4] **and Francisco Meirinhos**[5]

1 Department of Epileptology, University of Bonn, Venusberg Campus 1, 53127 Bonn, Germany **2** Helmholtz Institute for Radiation and Nuclear Physics, University of Bonn, Nußallee 14–16, 53115~Bonn, Germany **3** Institute for Theoretical Physics, University of Cologne, 50937~Köln, Germany **4** Forschungszentrum Jülich, Institute for Energy and Climate Research - Systems Analysis and Technology Evaluation (IEK-STE), 52428~Jülich, Germany **5** Physikalisches Institut and Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany

DOI: 10.21105/joss.01693

Software

- Review
- Repository
- Archive

Submitted: 23 August 2019
Published: 30 August 2019

Summary

A general problem for evaluating Markovian stochastic processes is the retrieval of the moments or the Kramers–Moyal coefficients $\mathcal{M}$ from data or time-series. The Kramers–Moyal coefficients are derived from an Taylor expansion of the master equation that describes the probability evolution of a Markovian stochastic process.

Given a set of stochastic data, ergodic or quasi-stationary, the extensive literature of stochastic processes awards a set of measures, such as the Kramers–Moyal coefficients or its moments, which link stochastic processes to a probabilistic description of the process or of the family of processes (Risken, H., 1996). Most commonly known is the Fokker–Planck equation or truncated forward Kolmogorov equation, partial differential equations, obtained from the Taylor expansion of the master equation.

Of particular relevance is the growing evidence that real-world data displays higher-order ($n > 2$) Kramers–Moyal coefficients, which has a two-fold consequence: The common truncation at third order of the forward Kolmogorov equation, giving rise to the Fokker–Planck equation, is no longer valid. The evidence of higher-order ($n > 2$) Kramers–Moyal coefficients in recorded data thus invalidates the aforementioned common argument for truncation, based on Pawula's theorem and thus rendering the Fokker–Planck description unjustified (Tabar, M. R. R., 2019). A clear and common example is the presence of discontinuous jumps in data(Anvari, M., Tabar, M. R. R., Peinke, J., and Lehnertz, K., 2016,Sahalia), which can give rise to higher-order Kramers–Moyal coefficients, and are evidenced in (Rydin Gorjão, L., Heysel, J., Lehnertz, K. and Tabar, M. R. R., 2019) and references within.

Calculating the moments or Kramers–Moyal coefficients strickly from data can be computationally heavy for long data series and is prone to be innacurate especially where the density of data points is scarce, e.g. usually at the boundaries on the domain of the process. The most straightforward approach is to perform a histogram-based estimation to evaluate the moments of the system at hand. This has two main drawbacks: it requires a discrete space of examination of the process, and is shown to be less accurate than using kernel-based estimators (Lamouroux, D. and Lehnertz, K., 2009).

This library is based on a kernel-based estimation, which allows for more robust results given both a wider range of possible kernel shapes to perform the calculation, as well as retrieving the results in a non-binned coordinate space, unlike histogram regressions (Silverman, B. W., 2018). It further employs a convolution of the series in studied with the selected kernel, circumventing the computational issue of sequential array summation, the most common bottleneck in integration time and computer memory.

The Journal of Open Source Software

contains rather than comprises

I wouldn't use manifold here given the mathematical audience → many works fine.

The package presented here comprises a manifold of options: A general open-source toolbox for the calculation of Kramers–Moyal coefficients for any given data series of any dimension and to any order, with a selection of commonly used kernel estimators.

Mathematics

sp.

For a general N-dimensional Markovian process $\boldsymbol{x}'(t) \in \mathbb{R}^N$ tke Kramers–Moyal yields all orders of the cumulants of the conditional probability distribution $P(\boldsymbol{x}', t+\Delta T|\boldsymbol{x}, t)$ as (Risken, H., 1996)

$$\mathcal{M}^\sigma(\boldsymbol{x}, t) = \lim_{\Delta t\to 0} \frac{1}{\Delta t}\int dx'[\boldsymbol{x}(t)' - \boldsymbol{x}(t)]^\sigma P(\boldsymbol{x}', t+\Delta T|\boldsymbol{x}, t),$$

with $[\dots]^\sigma$ a dyadic multiplication, and the power σ allowing for a set of powers depending on the dimensionality of the process.

The exact evaluation of the Kramers–Moyal coefficients for discrete or discretised datasets $\boldsymbol{y}(t)$—any human measure of a process is discrete, as well as any computer generated data—is bounded by the timewise limit imposed. Taking as an example a two-dimensional case with $\boldsymbol{y}(t) = (y_1(t), y_2(t)) \in \mathbb{R}^2$, the Kramers–Moyal coefficients $\mathcal{M}^{[\ell,m]} \in \mathbb{R}^2$ take the form

$$\mathcal{M}^{[\ell,m]}(x_1, x_2, t) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int \mathrm{d}y_1\mathrm{d}y_2 (y_1(t+\Delta t) - x_1(t))^\ell (y_2(t+\Delta t) - x_2(t))^m \cdot P(y_1, y_2; t+\Delta t|x_1, x_2; t$$

at a certain measure point (x_1, x_2). The order of the Kramers–Moyal coefficients is given here by the superscripts ℓ and m.

multi-line ?

an explicit

too many commas

sp.

Theoretically, there are still two details to attend to: Firstly, for non-stationary data, there is and explicity dependence on t, but, as is the case discussed here, for stationary (or quasi-stationary) data, $P(\boldsymbol{x}', t+\Delta T|\boldsymbol{x}, t) = P(\boldsymbol{x}', \Delta T|\boldsymbol{x})$. This entails time-independent Kramers–Moyal coefficients $\mathcal{M}^\sigma(\boldsymbol{x})$. Secondly, Δt should take the limiting case of $\Delta t \to 0$, but the restriction of any measuring or storing device—or the nature of the observables themselves—permits only time-sampled or discrete recordings. In the limiting case where Δt is equivalent to the minimal sampling rate of the data, the Kramers–Moyal coefficients take the form, in our two-dimensional example, as

$$\mathcal{M}^{[\ell,m]}(x_1, x_2) = \frac{1}{\Delta t}\langle \Delta y_1^\ell \Delta y_2^m|_{y_1(t)=x_1, y_2(t)=x_2}\rangle, \text{ with } \Delta y_i = y_i(t+\Delta t) - y_i(t).$$

It is straightforward to generalise this to any dimensions. The relevance and importance of adequate time-sampling was extensively studied and discussed in (Lehnertz, K., Zabawa, L. and Tabar, M. R. R., 2018). Notice here that if the sampling resolution of the process is known, than that is Δt.

then?

The Kramers–Moyal coefficients exist on an underlying probabilistic space, i.e., there exists a probabilistic measure assigned to the process, stemming from the master equation describing the family of such processes. The conventional procedure, as mentioned previously, is to utilise a histogram regression of the observed process and retrieve, via approximation or fitting, the Kramers–Moyal coefficient. The choice of a histogram measure for the Kramers–Moyal coefficient results in an acceptable measure of the probability density functions of the process but requires a new mathematical space (a distribution space). The employment of a kernel estimation approach, implemented in this library, permits an identical overview without the necessity of a new (discretised) distribution space, given that the equivalent space of the observable can be taken.

Like the histogram approach for the measure of the Kramers–Moyal coefficients, each single measure of the observable $\boldsymbol{y}(t)$ is averaged, with a designed weight, into the distribution space. The standing difference, in comparison to the histogram approach, is the riddance of a (discrete) binning system. All points are averaged, in a weighted fashion, into the distribution space—aiding especially in cases where the number of point in a dataset is small—and awarding a continuous measurable space (easier for fitting, for example) (Lamouroux, D. and Lehnertz, K., 2009).

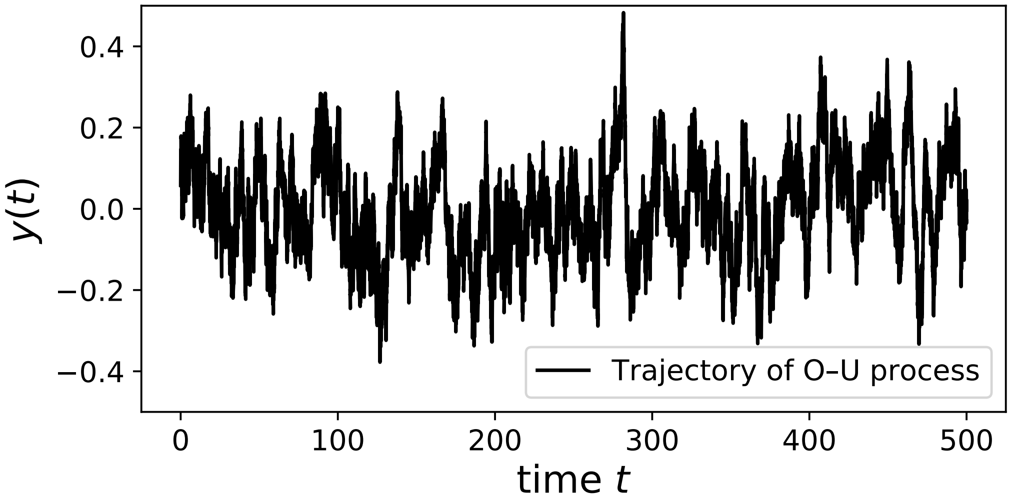
Exemplary one-dimensional Ornstein–Uhlenbeck process

A one-dimensional Ornstein–Uhlenbeck process $y(t)$ takes the form

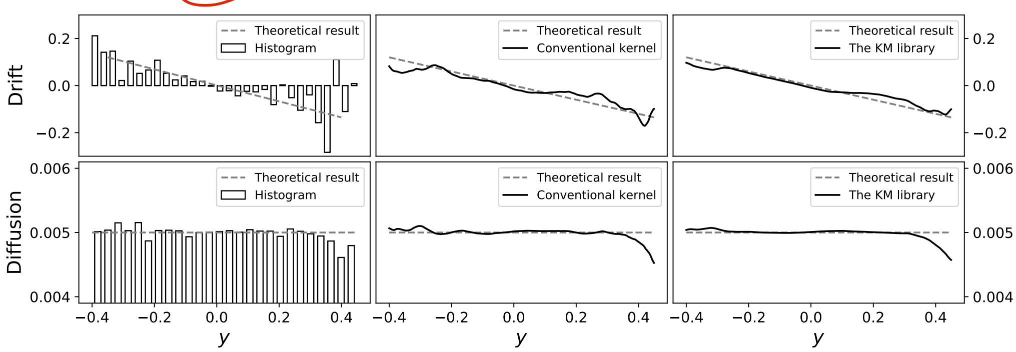
$$dy = -\theta y dt + \sigma dW(t),$$

with θ denoted as the *drift* or mean-reverting term, σ the *diffusion* term or stochastic amplitude, and $W(t)$ is a Brownian motion, i.e., a Wiener process. For this particular example set $\theta = 0.3$ and $\sigma = 0.1$.

To be able to test the library and the retrieval on the Kramers–Moyal coefficients, and subsequently recover the drift and diffusion term, one can numerically integrate the process. For the present case we employ a Euler–Maruyama integrator, for simplicity. There are more reliable and faster integrators, ~~as~~ for example `JiTCSDE` (Ansmann, G., 2018).



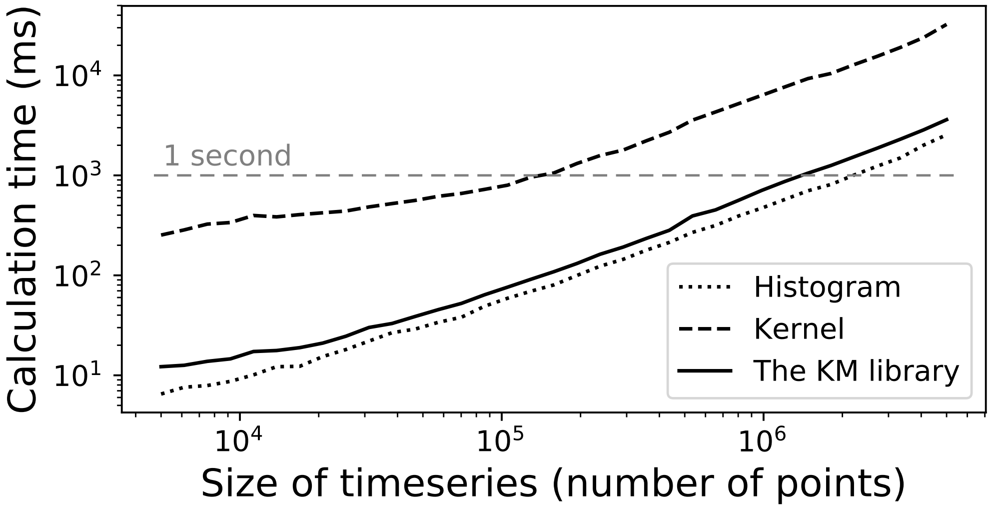
For the present case, with an integration over 500 time units and with a timestep of 0.001, which can be seen in Fig.~. The first and second Kramers–Moyal coefficients are presented in Fig.~, where as well the conventional histogram-based estimation, a non-convolution based kernel estimation, and this library implementing a convolution of the kernel with the terms the right-hand side in (). An Epanechnikov kernel was chosen for both kernel-based estimations.



Library

The presented library is comprised of two separate blocks, `kernels` and `km`, and is a standalone package for a non-parametric retrieval of Kramers–Moyal coefficients, solely dependent on `numpy`, `scipy`, and `functools`. The sub-module `kernels` comprises the kernels for the kernel-based estimation, similarly available in `sklearn`, and `km` performs the desired Kramers–Moyal calculations to any desired power (Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E., 2011).

In order compare the computational speed up of the library the aforementioned Ornstein-Uhlenbeck (**??**) was used (with $\theta = 0.3$ and $\sigma = 0.1$), and the total time of integration of the process was increased iteratively. In Fig.~ the comparative results of employing a histogram estimation with 200 bins, a conventional kernel-based regression in a space with 5500 numerical points, and this library's kernel-convolution method, over similarly 5500 numerical points.



Acknowledgements

L. R. G. and F. M. contributed equally to this project with their respective expertise. L. R. G. thanks Klaus Lehnertz and M. Reza Rahimi Tabar for all the help in understanding stochastic processes and developing this package, Dirk Witthaut for the support during the process of writing and reviewing, Gerrit Ansmann for the help in understanding python's intricacies, and Marieke Helmich for the text reviews. L. R. G. gratefully acknowledges support by the Helmholtz Association, via the joint initiative *Energy System 2050 - A Contribution of the Research Field Energy*, the grant No. VH-NG-1025, the scholarship funding from *E.ON Stipendienfonds*, and the STORM project of the Department of Mathematics, University of Oslo, under the supervision of Giulia di Nunno. F. M. gratefully acknowledges the fund, in part, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 277625399 - CRC 185.

References

Ansmann, G. (2018). Efficiently and easily integrating differential equations with JiTCODE, JiTCDDE, and JiTCSDE. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(4), 043116. doi:10.1063/1.5019320

Anvari, M., Tabar, M. R. R., Peinke, J., and Lehnertz, K. (2016). Disentangling the stochastic behavior of complex time series. *Scientific reports*, *6*, 35435. doi:10.1038/srep35435

Lamouroux, D. and Lehnertz, K. (2009). Kernel-based regression of drift and diffusion coefficients of stochastic processes. *Physics Letters A*, *373*(39), 3507–3512. doi:10.1016/j.physleta.2009.07.073

Lehnertz, K., Zabawa, L. and Tabar, M. R. R. (2018). Characterizing abrupt transitions in stochastic dynamics. *New Journal of Physics*, *20*(11), 113043. doi:10.1088/1367-2630/aaf0d7

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from http://www.jmlr.org/papers/v12/pedregosa11a.html

Risken, H. (1996). *The fokker–Planck Equation*. Springer, Berlin. doi:10.1007/978-3-642-61544-3

Rydin Gorjão, L., Heysel, J., Lehnertz, K. and Tabar, M. R. R. (2019). Analysis of bivariate jump-diffusion processes. *arXiv preprint arXiv:1907.05371*. doi:https://arxiv.org/abs/1907.05371

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge, New York. doi:10.1201/9781315140919

Tabar, M. R. R. (2019). *Analysis and data-based reconstruction of complex nonlinear dynamical systems*. Springer International Publishing. doi:10.1007/978-3-030-18472-8