Journal of
open research software

Rinn, P et al 2016 The Langevin Approach: An R Package for Modeling Markov Processes. *Journal of Open Research Software,* 4: e34, DOI: http://dx.doi.org/10.5334/jors.123

SOFTWARE METAPAPER

The Langevin Approach: An R Package for Modeling Markov Processes

Philip Rinn[1], Pedro G. Lind[1,2], Matthias Wächter[1] and Joachim Peinke[1]

[1] Institute of Physics and ForWind – Center for Wind Energy research, C.v.O. University Oldenburg, 26111 Oldenburg, Germany, DE

[2] Institute of Physics, University Osnabrück, Barbarastraße 7, 49076 Osnabrück, Germany, DE

Corresponding author: Philip Rinn (philip.rinn@uni-oldenburg.de)

We describe an *R* package developed by the research group *Turbulence, Wind energy and Stochastics* (TWiSt) at the Carl von Ossietzky University of Oldenburg, which extracts the (stochastic) evolution equation underlying a set of data or measurements. The method can be directly applied to data sets with one or two stochastic variables. Examples for the one-dimensional and two-dimensional cases are provided. This framework is valid under a small set of conditions which are explicitly presented and which imply simple preliminary test procedures to the data. For Markovian processes involving Gaussian white noise, a stochastic differential equation is derived straightforwardly from the time series and captures the full dynamical properties of the underlying process. Still, even in the case such conditions are not fulfilled, there are alternative versions of this method which we discuss briefly and provide the user with the necessary bibliography.

Keywords: *R*; stochastic processes; Markov processes; data analysis; time series

Funding statement: PGL thanks the German Environment Ministry as part of the research project "Probabilistic loads description, monitoring, and reduction for the next generation offshore wind turbines (OWEA Loads)" under grant number 0325577B.

(1) Overview

Introduction

When dealing with stochastic series of data measurements, standard statistical tools, such as mean and centered moments, are able to catch the essential features of thedistribution of observed values. Last end, sufficient high-order moments will retrieve a good approximation of the probability density function (PDF) associated with the stochastic process. However, PDFs are not able to fully characterize the dynamics underlying the process. A typical example is the Gaussian distribution: if the stochastic variable assumes values according to a Gaussian distribution, the dynamics producing such distribution of values can be as simple as an Ornstein-Uhlenbeck process [25] but it may also be the result of a much more complicated dynamics, as we exemplify below. Thus, while knowing the distribution of observed values is important as a first approach to the data, uncovering the complete dynamics of the process provides a much deeper insight into the system, which cannot be accessed through standard statistical tools.

Starting from a stochastic differential equation, a process can be statistically reconstructed through simple stochastic integration. The inverse problem however is much more complicated: would a set of measurements be enough for a bottom-up approach to infer the underlying dynamics of the process? The short answer is yes, there are cases where this is possible. In this paper we present the long answer implemented as a package for *R* (see [21]), which can be easily used, composing a method which we call the Langevin Approach. This approach was introduced by Peinke and Friedrich in the late 1990s [5, 28] and further developed in the last decades. For a review see Friedrich et al. [6].

Stochastic equations: The Langevin model

A wide range of dynamical systems can be described by a stochastic differential equation, the (non-linear) Langevin equation (cf. [9, 25, 30]).

Consider a general stochastic trajectory $X(t)$ in time t. The time derivative of the system's trajectory $\frac{dX}{dt}$ can be expressed as the sum of two complementary contributions: one being purely deterministic and another one being stochastic, governed by a stochastic "force" $\Gamma(t)$, defined as a δ-correlated Gaussian white noise, i.e., $\langle\Gamma(t)\rangle = 0$ and $\langle\Gamma(t)\Gamma(t')\rangle = 2\delta\,(t - t')$. While the deterministic term is defined by a function, $D^{(1)}(X)$ the stochastic contribution is weighted by another function, $D^{(2)}(X)$, yielding the evolution equation of X

$$\frac{dX}{dt} = D^{(1)}(X) + \sqrt{D^{(2)}(X)}\,\Gamma(t), \qquad (1)$$

where the square root is taken for consistency, as will be clear below. We assume stationary time series here, so $D^{(1)}$ and $D^{(2)}$ are not time dependent but we show briefly how non-stationary time series can be treated in Section "A glimpse beyond the *Langevin* package".

The Langevin equation should be interpreted as follows: for every time t where the system meets an arbitrary but fixed point X in phase space, $X(t+\tau)$ with small τ is defined by the deterministic function $D^{(1)}(X)$ and the stochastic function $\sqrt{D^{(2)}(X)}\Gamma(t)$, through trivial (Euler) stochastic integration [6]:

$$X(t+\tau) = X(t) + D^{(1)}(X)\tau + \sqrt{D^{(2)}(X)\tau}\,\eta(t), \qquad (2)$$

where $\eta(t)$ is a normally distributed random variable. Here we use the Itô picture of stochastic integrals, for further details see Gardiner [7].

Functions $D^{(1)}(X)$ and $D^{(2)}(X)$ are usually called drift and diffusion coefficients respectively and they can be as simple as constants or linear functions of X, as e.g., the Ornstein-Uhlenbeck process, as well as more complicated nonlinear functions, typically polynomials up to a given order. In particular if $D^{(2)}$ is explicitly depending on X, the case is called multiplicative noise.

In all cases, through substitution of the selected functions into Equation 2 one is able to generate samples of series having the same statistical features and obeying the same dynamics.

Figure 1a shows an illustration of a time series obtained through integration of Equation 2 for a cubic drift $D^{(1)}(X) = -X^3 + X$, and a quadratic diffusion, $D^{(2)}(X) = X^2 + 1$. Notice that these drift and diffusion coefficients describe a non-trivial dynamics, namely the underlying deterministic process, i.e., $D^{(2)} \equiv 0$, has two attractive fixed points at $X = \pm 1$. The processes tends to converge to one of two stable states being at the same time perturbed by a stochastic fluctuation ($D^{(2)} \neq 0$) which is able to push the system from one stable fixed point to the other. As shown in **Figure 1b**, despite this non-trivial dynamics, the PDF is a Gaussian distribution with zero mean and unit standard deviation, the same PDF as for a simple Ornstein-Uhlenbeck process with $D^{(1)} = -X$ and $D^{(2)} = 1$.

This is one of many possible examples that illustrates the deep insight, which an evolution equation like in Equation 1 can provide and which is not obtained by looking at a density distribution, see Appendix for further details.

From stochastic data to the Langevin model

As explained previously, it is easy to generate data through the integration of a stochastic equation, such as Equation 2. More difficult is the inverse problem, to derive functions $D^{(1)}$ and $D^{(2)}$ from given data.

A condition to derive the drift and diffusion numerically is that the time-steps τ of the set of X-values are small enough (see Honisch and Friedrich [10] for details). If the system is at time t in the state $x = X(t)$ the drift can be calculated for small τ by averaging over the difference, $X(t+\tau) - X(t)$, of the system state at $t+\tau$ and the state at t. Check Equation 2 above. This average is the first conditional moment of the series and it can be mathematically proven that its time derivative yields the drift coefficient. Similarly, computing the second conditional moment, i.e., the average squared differences between $X(t+\tau)$ and x, yields the diffusion coefficient [25].

Therefore, having a series of data, one estimates the drift and diffusion by computing the averages of the first and second power of the differences between $X(t+\tau)$ and x:

$$M^{(n)}(x,\tau) = \left\langle (X(t+\tau) - X(t))^n \right\rangle \Big|_{X(t)=x}, \qquad (3)$$

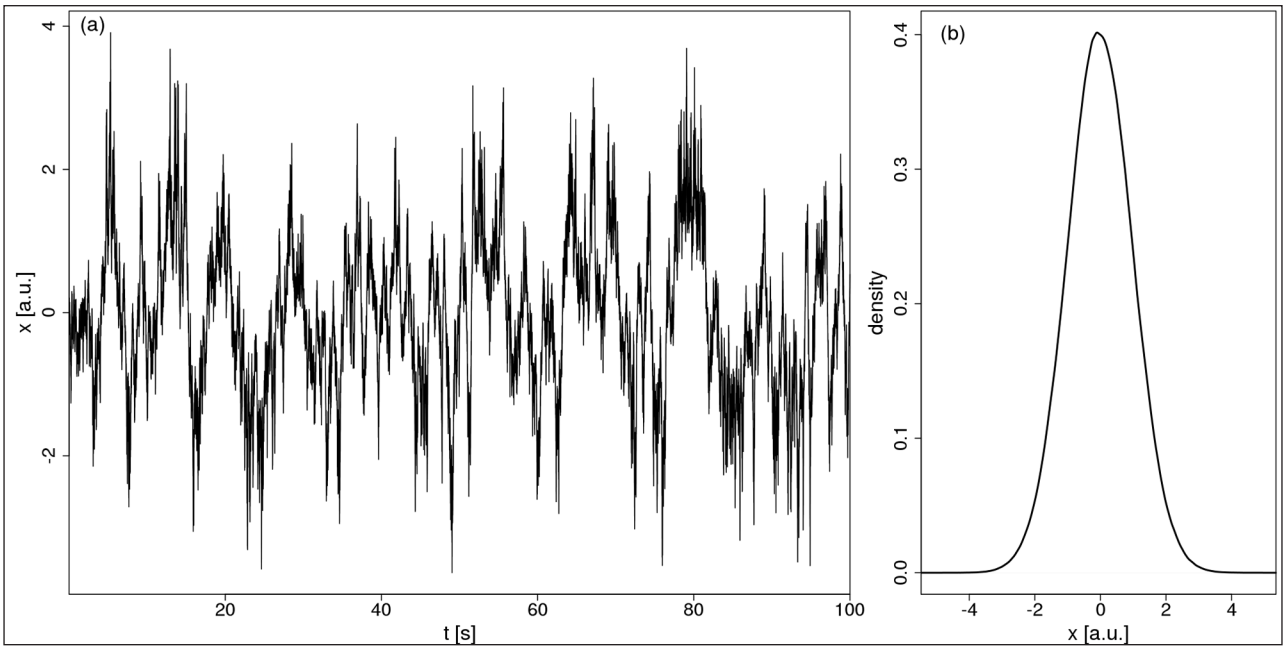

Figure 1: (a) Sketch of a stochastic process in time governed by a cubic drift and quadratic diffusion contributions and (b) its corresponding probability density function (PDF). Though the series shows a bistable dynamics (cubic drift) the PDF follows a Gaussian function, equivalent to an Ornstein-Uhlenbeck process (see text).

where $\langle\cdot\rangle$ represents the average over time t. Mathematically the drift and diffusion coefficients are defined as [25]

$$D^{(n)}(x) = \lim_{\tau \to 0} \frac{1}{n!\tau} M^{(n)}(x, \tau), \quad (4)$$

which means that they are given by derivatives of the corresponding conditional moments $M^{(n)}(x, \tau)$ with respect to τ. In many cases, for a fixed x, the conditional moments depend linearly on τ for the smallest range of τ values and consequently the drift and the diffusion coefficients at this state x are estimated solely by the quotient between the corresponding conditional moment and τ in this range.

Figures 2a and **2b** show respectively the drift and diffusion coefficients of the series integrated in the previous section and sketched in **Figure 1**. The theoretical expressions of the coefficients used when generating the synthetic data through integration of Equation 1 are indicated as red dashed lines, while the estimated values of the coefficients at each selected bin are plotted with bullets.

Back to the data

The Langevin Approach summarized previously is applied under a few conditions, though, as we discuss afterwards, when such conditions are not fulfilled in many cases it is still possible to overcome that and apply an alternative approach which also retrieves the dynamics underlying the stochastic process. For the completeness of this paper and for the consistency of the application of our *R* functions, we advise the user to briefly test the data. Three conditions should in general be tested as a preliminary checking procedure and two further conditions can afterwards be tested as cross-checking.

The first condition is that the data series is stationary. Indeed, the averages for computing the conditional moments have to be taken over all $t = t_i$ where $X(t_i) = x$ (see Equation 3). If the series is non-stationary these averages are in principle not meaningful.

The second condition is that the process should be Markovian, i.e., the present state should depend on the previous state solely. Mathematically it means an equivalence between two-point statistics, $p(X(t + \tau), X(t))$, and any higher-order statistics, $p(X(t + \tau), X(t), ..., X(t - n\tau))$. This equivalence leads to the following equality between conditional probabilities of finding a value of $X(t + \tau)$ under the condition that $X(t), X(t - \tau), ..., X(t - n\tau)$ have selected values:

$$p(X(t+\tau)|X(t)) = p(X(t+\tau)|X(t), ..., X(t-n\tau)). \quad (5)$$

This should hold for any positive integer n. In practice, one tests the equality for three-point statistics ($n = 1$) only and assumes that if the equality holds it will also hold for higher-order statistics, since all correlations shall decrease monotonically with time.

To test if the process is Markovian one can also use alternatively the Wilcoxon test [32], in case one is dealing with single variable stochastic processes. For details see Renner et al. [22, Appendix A].

The third condition to be tested comes from a mathematical result called Pawula Theorem [25], from which it follows a necessary condition for Equation 1 to be valid: the fourth conditional moment must be constant, i.e., $D^{(4)} = 0$. To test that one computes its derivative with respect to the time-lag, the fourth coefficient

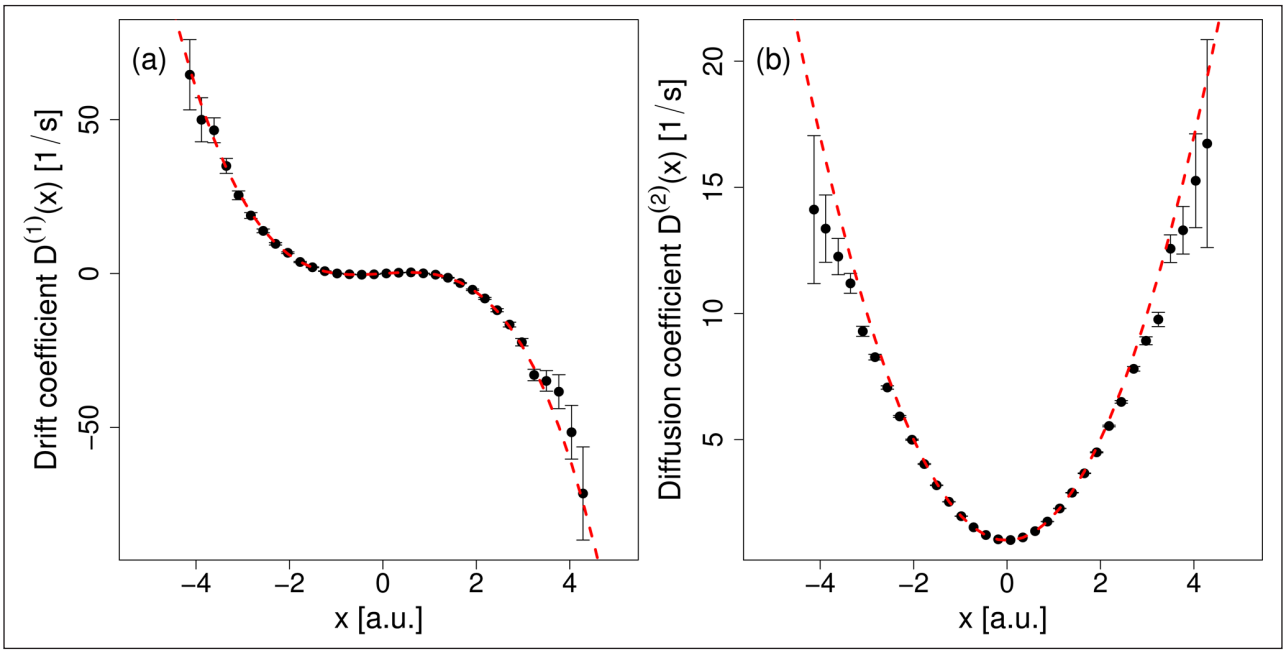

Figure 2: One-dimensional Langevin Approach: (a) drift coefficient, $D^{(1)}(x) = -x^3 + x$, and (b) diffusion coefficient, $D^{(2)}(x) = x^2 + 1$. Circles indicate the numerical results while the red dashed line indicates the theoretical coefficient, used when generating the synthetic data. Here 10^7 data points from the series illustrated in **Figure 1a**, were used for computing the averaged conditional moments.

$$D^{(4)}(x) = \lim_{\tau \to 0} \frac{1}{4!\tau} \left\langle (X(t+\tau) - X(t))^4 \right\rangle \Big|_{X(t)=x} \tag{6}$$

and checks if it vanishes, i.e., if it is small compared to the diffusion coefficient: $D^{(4)}(x) \ll (D^{(2)}(x))^2 \; \forall x$. This coefficient is also useful for computing the numerical error of the diffusion coefficient [19].

The tests whether conditions two and three hold ensure that $\Gamma(t)$ (see Equation 1) is δ-correlated and Gaussian distributed.

If all these conditions are fulfilled the Langevin Approach can be carried out and the two functions, drift coefficient $D^{(1)}$ and diffusion coefficient $D^{(2)}$, can be derived from the given data. With the derived coefficients two additional cross-checking tests can be done.

The first one is to check if the stochastic force in Equation 1 fulfills the two conditions of a δ-correlated Gaussian white noise. To that end, one substitutes in Equation 2 the derived $D^{(1)}(X)$ and $D^{(2)}(X)$ and solves it with respect to $\eta(t)$:

$$\eta(t) = \frac{X(t+\tau) - X(t) - D^{(1)}(X)\tau}{\sqrt{D^{(2)}(X)\tau}}. \tag{7}$$

Taking τ as the time-step of the observed time-series and substituting in $X(t + \tau)$ and $X(t)$ successive values of that series one re-obtains a series for $\eta(t)$ which should be normally distributed.

The second cross-checking test is to substitute in Equation 2 the derived $D^{(1)}(X)$ and $D^{(2)}(X)$ coefficients, generate synthetic series and compare if its increments

$$\Delta_\tau(t) = X(t+\tau) - X(t) \tag{8}$$

have the same distribution as the original series for a fixed τ spanning from the time-step of the original series up to two or more orders of magnitude larger.

Some extra care should be taken if the derived $D^{(1)}(X)$ and $D^{(2)}(X)$ coefficients show linear drift and quadratic diffusion forms as this is also the case for every Langevin process if the sampling interval is large compared to the relaxation time of the process. Riera and Anteneodo [23] presented a method for cross-checking in this case.

Notice that, though the fulfillment of all such conditions through the proposed preliminary tests and cross-checking tests guarantees that the Langevin Approach can be applied, the rejection of one or more of these tests is still no reason for avoiding this approach. In Section "A glimpse beyond the *Langevin* package" we will come back to this issue.

Implementation and architecture

In this section we present the implementation of the Langevin Approach describing the two available *R* functions, `Langevin1D` and `Langevin2D`. The function `Langevin1D` deals with single time-series while `Langevin2D` should be used for two-dimensional cases, when one has two stochastic variables to be analyzed simultaneously.

The one-dimensional case deals with an evolution equation similar to Equation 1 and the two-dimensional case comprehends two stochastic variables, $X_1(t)$ and $X_2(t)$, governed by:

$$\frac{d}{dt}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} D_1^{(1)}(X_1, X_2) \\ D_2^{(1)}(X_1, X_2) \end{bmatrix} + \begin{bmatrix} g_{11}(X_1, X_2) & g_{12}(X_1, X_2) \\ g_{21}(X_1, X_2) & g_{22}(X_1, X_2) \end{bmatrix} \begin{bmatrix} \Gamma_1(t) \\ \Gamma_2(t) \end{bmatrix} \tag{9}$$

where clearly now the drift function $\mathbf{D}^{(1)} = (D_1^{(1)}, D_2^{(1)})$ is a two-dimensional vector and the diffusion coefficient is a 2×2-matrix given by $\mathbf{D}^{(2)} = \mathbf{g}\mathbf{g}^T$, i.e., $= D_{ij}^{(2)} = \sum_k g_{ik} g_{jk}$. Similar to the one-dimensional case the integration of Equation 9 follows from a simple Euler scheme leading to:

$$\begin{bmatrix} X_1(t+\tau) \\ X_2(t+\tau) \end{bmatrix} = \begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} + \tau \begin{bmatrix} D_1^{(1)}(X_1, X_2) \\ D_2^{(1)}(X_1, X_2) \end{bmatrix} + \sqrt{\tau} \begin{bmatrix} g_{11}(X_1, X_2) & g_{12}(X_1, X_2) \\ g_{21}(X_1, X_2) & g_{22}(X_1, X_2) \end{bmatrix} \begin{bmatrix} \eta_1(t) \\ \eta_2(t) \end{bmatrix} \tag{10}$$

where $\eta_1(t)$ and $\eta_2(t)$ are two independent normally distributed random variables. In our implementation the conditional moments $M^{(n)}(x, \tau)$, Equation 3, are estimated by dividing the state space of x in N intervals, or bins, $(I_1, ..., I_N)$ and calculating the mean values for each interval I_i:

$$M^{(n)}(x, \tau) = \left\langle (X(t+\tau) - X(t))^n \right\rangle \Big|_{X(t) \in I_i}. \tag{11}$$

For estimating the drift and diffusion coefficients from the conditional moments we insert Equation 2 into Equation 11 and apply the conditional averages for $n = 1, 2$ leading to:

$$M^{(1)}(x, \tau) \approx D^{(1)}(x)\tau, \tag{12}$$

$$M^{(2)}(x, \tau) \approx 2D^{(2)}(x)\tau + (D^{(1)}(x)\tau)^2. \tag{13}$$

Important to notice is that for $M^{(2)}(x, \tau)$, Equation 13, a term quadratic in $D^{(1)}(x)$ and τ has to be considered. We estimate drift and diffusion coefficients from the slope of a weighted linear regression of Equations 12 and 13.

The implementation of the functions heavily relies on the *C++* linear algebra library `Armadillo` [26] for which *RcppArmadillo* and *Rcpp* provide the integration with *R* [3, 4]. We choose `Armadillo` as it results in fast code especially for large data sets and has an easy readable syntax. The functions `Langevin1D` and `Langevin2D` use `OpenMP` [2] if available to take advantage of shared memory multiprocessing. Here we parallelize the evaluation of the drift and diffusion coefficients for the bins as their evaluation is independent for each bin.

In the following subsections we present one- and two-dimensional examples of Langevin processes and walk through the analysis based on the framework described in the previous section.

Example for analyzing one-dimensional data sets[1]

As an example we integrate the Langevin equation illustrated in **Figure 1a** with cubic drift and quadratic diffusion, namely

$$\frac{dx}{dt} = x(t) - x^3(t) + \sqrt{x^2(t) + 1}\,\Gamma(t). \tag{14}$$

The presented package provides the function `time-series1D` to do the integration using an Euler integration scheme:

```
R> library("Langevin")
R> sf <- 1000
R> set.seed(4711)
R> x <- timeseries1D(N = 1e7, d11 = 1, d13 = -1,
+    d22 = 1, d20 = 1, sf = sf)
```

Extracting drift and diffusion coefficients from the generated time series is done by the function `Langevin1D`. Here two parameters that are important for the estimation have to be given as arguments.

The first one is the number of *bins* dividing the variable space x in discrete bins at which drift and diffusion are estimated. This integer should not be so large that each bin does no longer include a significant number of points (typically ~ 100) and also not so small that no dependence of the drift and diffusion on the state variable can be observed.

The second parameter is the vector *steps* to calculate the conditional moments for different τ values (Equation 3). The conditional moments will be computed for each bin and for each step. For each bin, a linear fit is computed for all steps in *steps.* Typically a vector of up to ten steps is given in samples (= $\tau \cdot$ sf).

```
R> bins <- 40
R> steps <- c(1:3)

R> ests <- Langevin1D(x, bins, steps)
```

From the resulting list *ests,* plots of the estimated drift and diffusion coefficients can be generated (see **Figure 2**). Here we use *plotrix* [18] to add errorbars.

```
R> library("plotrix")
R> attach(ests)
R> par(mfrow = c(1, 2))
R> plotCI(mean_bin, D1, uiw = eD1, xlab ="x [a.u.]",
+    ylab = expression(paste("Drift coefficient",
+    D^(1), "(x) [a.u.]")),
+    cex = 2, pch = 20)
R> lines(mean_bin, mean_bin - mean_bin^3,
+    col = "red", lwd = 3, lty = 2)
R> plotCI(mean_bin, D2, uiw = eD2, xlab =
+    "x [a.u.]", ylab = expression(paste
+    ("Diffusion coefficient", D^(2),
+    "(x) [a.u.]")),
+    cex = 2, pch = 20)
R> lines(mean_bin, mean_bin^2 + 1, col = "red",
+    lwd = 3, lty = 2)
```

We now want to walk through some of the remarks given in Section "Back to the data" to check if the conditions under which we applied the Langevin Approach are fulfilled. We do not check if the time series is stationary and fulfills the Markovian properties, since here we already know this (as we are using synthetic data).

Therefore we concentrate on cross-checking the estimated drift and diffusion coefficients. For checking if $D^{(4)}(X)$ is small compared to $D^{(2)}(X)$ (Pawula Theorem) we use the function `summary` which also computes the ratio between $D^{(4)}$ and $(D^{(2)})^2$:

```
R> summary(ests)
Number of bins: 40
Population of the bins:
       Min. : 3
       Median: 32034
       Mean : 250000
       Max. : 1053446
Number of NA's for D1: 7
Number of NA's for D2: 7
Ratio between D4 and D2^2:
       Min. : 0.002004
       Median: 0.002102
       Mean : 0.002385
       Max. : 0.004487
```

The result shows that $D^{(4)}(X)$ is smaller than 0.5% of the squared diffusion coefficient, indicating the necessary condition of the Pawula Theorem holds.

As a second cross-check we compare the increments, as defined in Equation 8, of the original time series with the ones computed from the reconstructed time series based on the estimated drift and diffusion functions.

To this end we fit a cubic function to the estimated drift coefficient and a quadratic function to the diffusion coefficient:

```
R> estD1 <- coef(lm(D1 ~ mean_bin
+    I(mean_bin^2) + I(mean_bin^3),
+    weights = 1/eD1))
R> estD2 <- coef(lm(D2 ~ mean_bin
+    I(mean_bin^2), weights = 1/eD2))
```

The resulting coefficients are used to generate a new time series with `timeseries1D`:

```
R> rec_x <- timeseries1D(N = 1e7, d10 =
+    estD1[1], d11 = estD1[2],
+    d12 = estD1 [3], d13 = estD1[4], d20 =
+    estD2[1], d21 = estD2[2],
+    d22 = estD2[3], sf = sf)
```

We want to emphasize here that the Langevin Approach does not require the drift and the diffusion coefficients to be of any particular functional form, from the estimated coefficients one could directly integrate a stochastic time series which can be used to calculate the increments. We fit the estimated coefficients to polynomials only to be able to use the function `timeseries1D` for the integration.

From the original and the reconstructed time series we now calculate PDFs of the increments for different τ and plot them to inspect their agreement visually:

```
R> plot(1,1, log = "y", type = "n", xlim =
c(-11, 12), ylim = c(1e-17, 5),
+    xlab = expression(Delta[tau]/
+    sigma[Delta[tau]]), ylab = "density")

R> tau <- c(1,10,100,1000)
R> for(i in 1:4) {
+      delta <- diff(Ux, lag = tau[i])
+      rec_delta <- diff(rec_x, lag = tau[i])
+      den <- density(delta)
+      den$x <- den$x/sd(delta, na.rm = TRUE)
+      rec_den <- density(rec_delta)
+      rec_den$x <- rec_den$x/sd(rec_delta,
+      na.rm = TRUE)
+      lines(den, lwd = 2, col = i)
+      lines(rec_den, lwd=2, lty = 2, col = i)
+ }
```

Figure 3 shows the output: there is indeed good agreement of both increment PDFs for a wide range of τ values.

Therefore we can assume that our estimated drift and diffusion coefficients describe the process sufficiently.

Notice once again that while the PDF of the series generated by Equation 14 is the same as the one of the simple Ornstein-Uhlenbeck process, $\frac{dx}{dt} = -x(t) + \Gamma(t)$, our Langevin Approach is able to uncover the correct dynamics with a bistable drift and a non-constant diffusion (see Appendix).

Example for analyzing two-dimensional data sets[1]

As a two-dimensional example we integrate the coupled Langevin equations in Equations 9 for a particular choice of the drift and diffusion coefficients, namely [28]

$$\frac{X_1}{dt} = X_2 + a\Gamma_1(t) \tag{15a}$$

$$\frac{X_2}{dt} = 0.02X_1 + 0.03X_2 - X_1^3 - X_1^2X_2 + a\Gamma_2(t), \tag{15b}$$

where a is a constant. **Figure 4a** shows the integrated trajectory (X_1, X_2) for $a = 0$, a case where no stochastic contribution is present, whereas in **Figure 4b** the same trajectory is plotted now with stochastic forces having a constant amplitude of $a = 0.05$.

The integration is performed by `timeseries2D`. Drift and diffusion functions are full cubic and quadratic polynomials respectively and the elements a_{ij} of the matrices are defined by the corresponding equations for the drift and diffusion terms (see Equations 9 and 10):

$$D^{(1)}_{1,2} = \sum_{i=1}^{4}\sum_{j=1}^{5-i} a_{ij}\, x_1^{(i-1)} x_2^{(j-1)} \quad \text{and}$$

$$g_{11,12,21,22} = \sum_{i=1}^{3}\sum_{j=1}^{4-i} a_{ij}\, x_1^{(i-1)} x_2^{(j-1)}\,.$$

Estimating the drift and diffusion coefficients is done by `Langevin2D`, here the same rules for *bins* and *steps* apply as for the one-dimensional case.

The results shown in **Figure 5** are generated by the following command lines (the source code for plotting the figure can be found in the aforementioned *examples.r*):

```
R> D1_1 <- matrix(0, nrow = 4, ncol = 4)
R> D1_1 [1, 2] <- 1
R> D1_2 <- matrix(0, nrow = 4, ncol = 4)
R> D1_2 [2, 1] <- 0.02
R> D1_2[1, 2] <- 0.03
R> D1_2[4, 1] <- -1
R> D1_2[3, 2] <- -1

R> g_11 <- matrix(0, nrow = 3, ncol = 3)
R> g_12 <- matrix(0, nrow = 3, ncol = 3)
R> g_21 <- matrix(0, nrow = 3, ncol = 3)
R> g_22 <- matrix(0, nrow = 3, ncol = 3)
R> g_11[1, 1] <- 0.0025
R> g_22[1, 1] <- 0.0025

R> set.seed(4711)
R> x <- timeseries2D(N = 1e8, 0.145, 0.0002,
+    D1_1, D1_2, + g_11, g_12, g_21, g_22,
+    sf = sf)
R> ests <- Langevin2D(x, bins, steps)
```

The numerical results can be properly fitted through the functions used for the integration in Equations 15, namely: $D_1^{(1)} = X_2$, $D_2^{(1)} = 0.02X_1 + 0.03X_2 - X_1^3 - X_1^2X_2$, $D_{11}^{(2)} = D_{22}^{(2)} = 0.05^2$ and $D_{12}^{(2)} = D_{21}^{(2)} = 0$. Notice that the large deviations in the boundaries are due to the finite length of the time series and thus the lower population in the boundary bins resulting in a poorer estimation of the drift and the diffusion.

A glimpse beyond the *Langevin* package

The two examples exposed above show cases where all conditions are fulfilled. When analyzing real empirical data sets this is often not the case: one or more of the

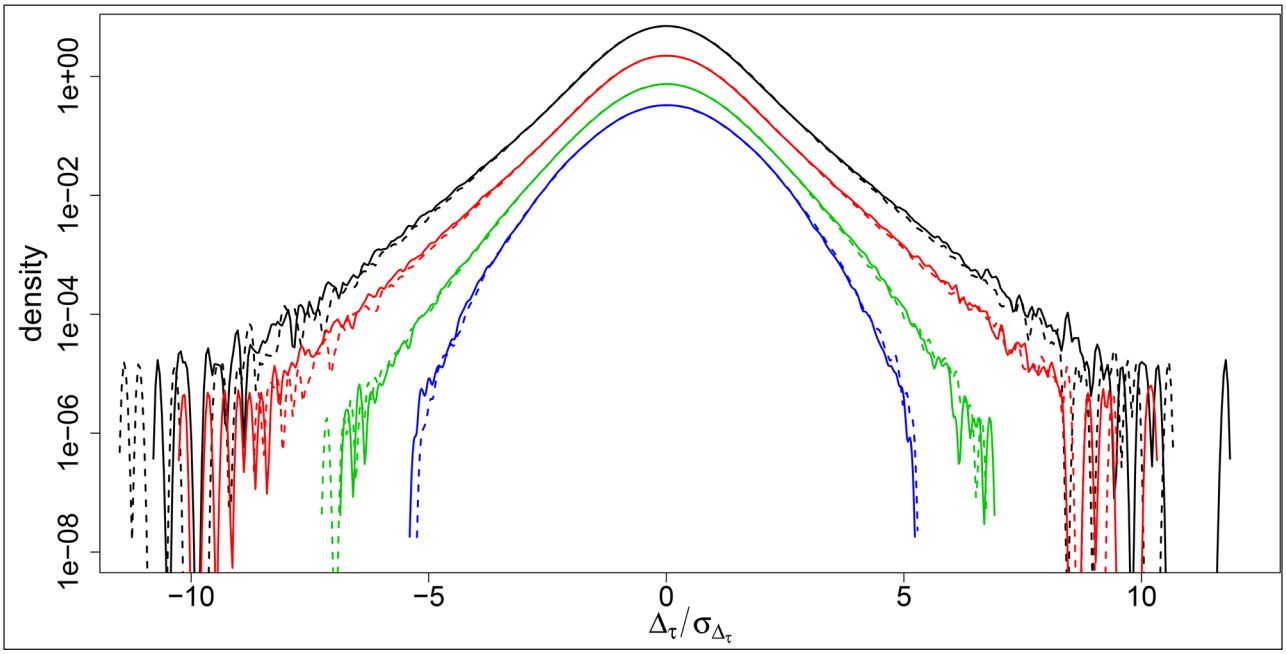

Figure 3: PDFs of the increments for $\tau = 1$, $\tau = 10$, $\tau = 100$ and $\tau = 1000$ time lags (from top to bottom). Solid lines show the results for the original time series, broken lines the result for the reconstructed time series.

conditions under which the Langevin Approach is applied are not met. Still, in the last years we developed different alternatives and extensions to this approach to overcome specific situations in stochastic data analysis. In this section we briefly describe these alternatives and extensions.

One first problem that researchers face is the non-stationary character often appearing in real data. Here, one of two approaches may be possible. One is to ascertain if for "shorter" time-windows of the data series stationarity may be assumed. In case the data set can be decomposed

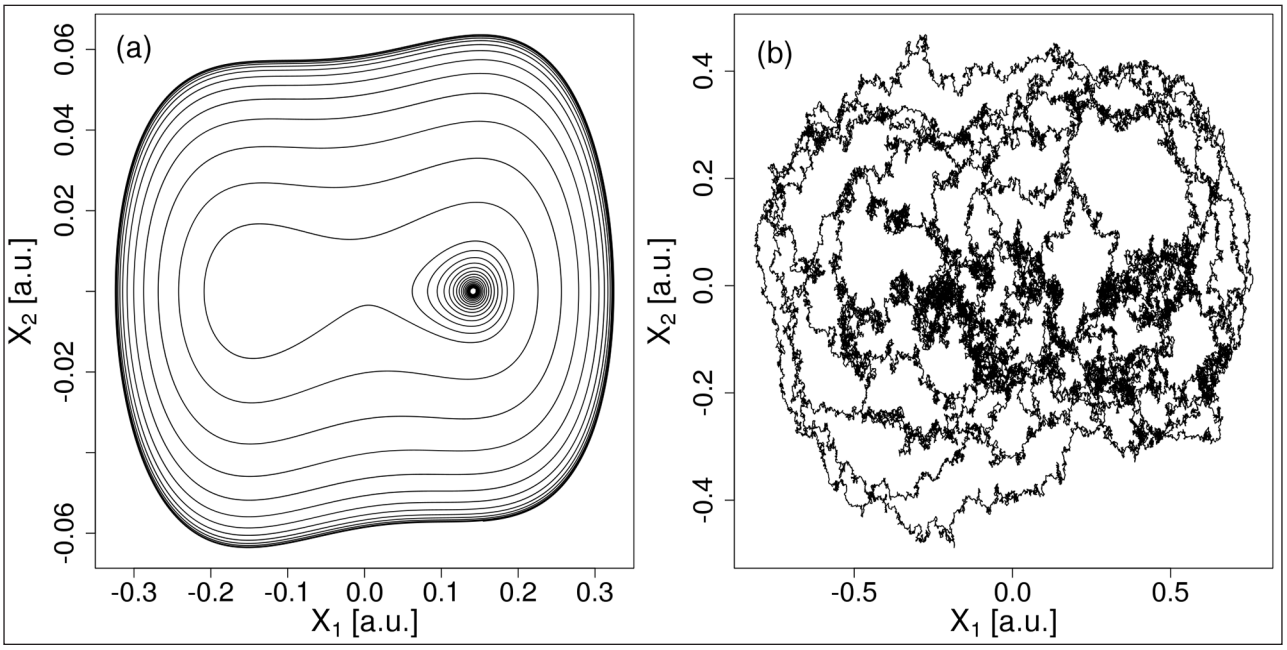

Figure 4: (a) Trajectory $(X_1(t), X_2(t))$ from Equations 15 with $a = 0$ and (b) the same trajectory integrating the same equations with non-zero stochastic terms ($a = 0.05$). For plotting 10^6 resp. 10^5 data points where used.

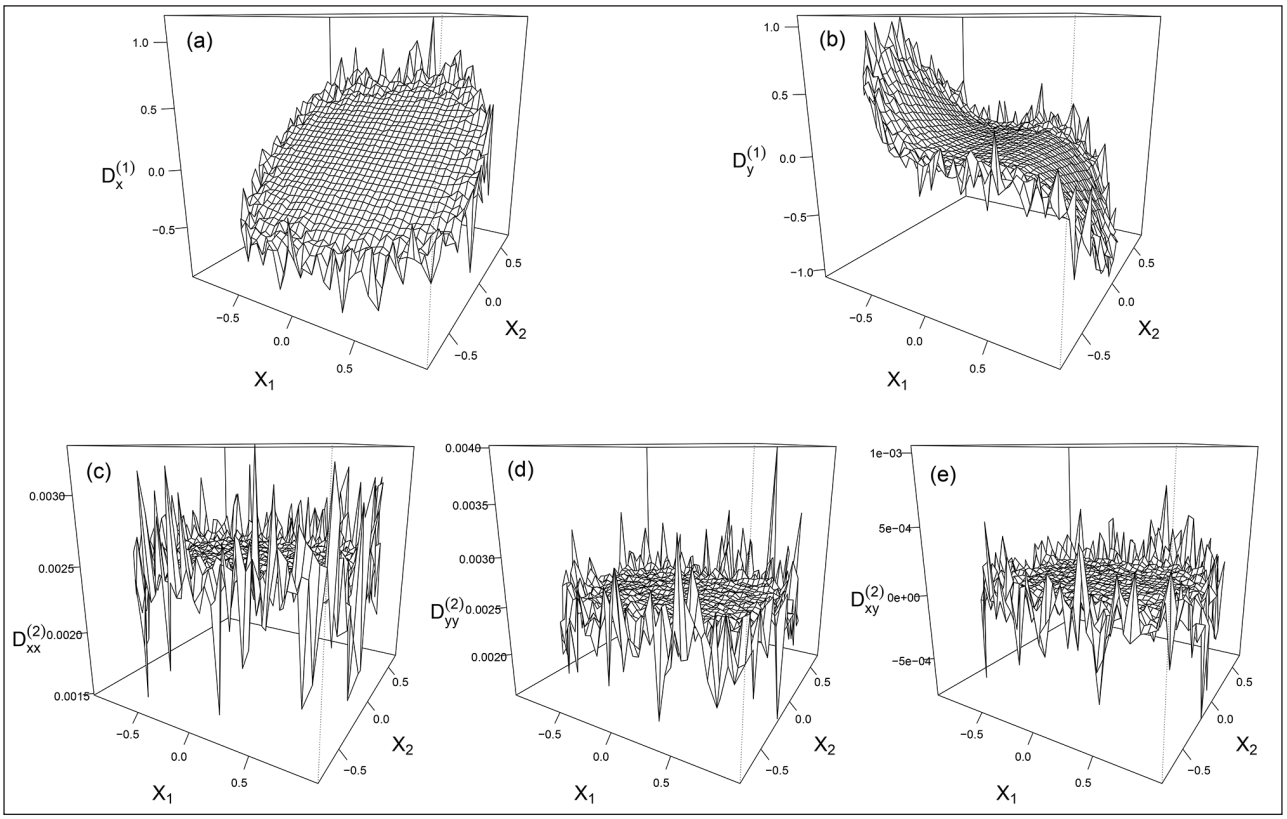

Figure 5: Drift coefficient of (a) the X_1 component, $D_1^{(1)}$, and (b) the X_2 component, $D_2^{(1)}$, together with all diffusion coefficients, namely (c) $D_{11}^{(2)}$, (d) $D_{22}^{(2)}$, (e) $D_{12}^{(2)} = D_{21}^{(2)}$. See Equation 9. Estimated with added noise, i.e., $a = 0.05$ in Equation 15.

in a series of time-windows which may overlap, each one having more or less constant statistical moments of the observable, the Langevin approach can be applied separately to each one of them, yielding a set of drift and diffusion coefficients, one for each time-window. In the end one extracts one drift and one diffusion coefficient, both functions of the observable and also of time.

Another possibility to handle non-stationary data sets is to check if they can be conditioned to other observables. In that case, considering the periods of the data sets associated to a particular value of the conditioning observable may be itself stationary. This is the case of the stochastic series measured of wind turbines [20, 31]. The power output of one wind turbine or the loads applied to it by the wind field are two observables whose measurement series are by themselves non-stationary. The wind velocity is the observable driving those properties and it is also non-stationary. However, we have shown that both wind power production [31] and instantaneous loads [20] can be analyzed through the Langevin Approach if we conditioned both the drift and the diffusion coefficients to each particular admissible value of the wind speed.

Another problem researchers often face are situations where the correlation function of the process is not fully resolved, i.e. the sampling rate of the data is too low. When this is the case, the correlation length is overestimated leading to wrong estimation of the time scale associated with the drift [12]. Kleinhans and Friedrich [12], Lade [14] and Honisch and Friedrich [10] developed optimization methods to still resolve drift and diffusion coefficients properly for cases where data is poorly sampled. These optimizations are computationally demanding, particularly when no functional form of the drift and diffusion coefficients is known a priori.

The second condition listed above is the Markov property. When the series of measurements fails to fulfill the Markov tests described above, it cannot be reconstructed through stochastic Euler integration since the next state cannot be estimated from the present state alone (see Equation 2). This happens, for instance, when a Markov process is spoiled by additional additive noise when a measurement is taken (see [13]). While the process alone, $X(t)$, is Markovian, the actual measurement, which retrieves $X(t) + Y(t)$, does not fulfill the Markov property. In such cases the limits computed for the coefficients $D^{(1)}$ and $D^{(2)}$ diverge (see Equation 4): when $\tau \to 0$ the conditional moment for the measured values (numerator) does not vanish. Still, it is frequently possible to obtain the correct drift and diffusion coefficients for the Markov process $X(t)$ through simple changes of their estimates [1, 16, 17, 27]. In cases of correlated noise where $\left\langle \tilde{\Gamma}(t) \right\rangle = 0$ holds the drift coefficient $D^{(1)}(X)$ can still be reconstructed correctly.

A third problem that may appear during preliminary tests of empirical data is the non-vanishing fourth coefficient $D^{(4)}$. As stated above in Section "Back to the data", according to Pawula Theorem [25] the fourth conditional moment must be independent of the time-lag τ. If not, one cannot assume that the stochastic process is governed by a Langevin equation, Equation 1. However, in such cases, although no evolution equation can be extracted and therefore the estimated functions $D^{(1)}$ and $D^{(2)}$ have not the meaning of drift and diffusion contributions, one can still use both to provide valuable insight about the system being analyzed. One example is the work of Rinn et al. [24] on in-situ analysis of the elastic features of a mechanical beam structure for realistic excitations with correlated noise as it appears in real-world situations. They could show that the slope of the drift coefficient $D^{(1)}$ is a sensitive indicator of the damage and compared to frequency based approaches, like power spectra, which estimate changes of the eigenfrequency of the structure, it is even more sensitive to small damages.

Finally, it is also important to stress that, while the functions of the presented package were prepared for analyzing data series as processes in time, the Langevin Approach can be adapted for analyzing processes in scale. In fact, when the process is not Markovian in time, violating Equation 5, there is the possibility that it is Markovian in "scale". What does this mean? It means that the increments Δ_τ introduced above follow a Markovian process in τ i.e., in time-lags but are instationary. Such analysis in scale is able to reproduce e.g., turbulence energy cascades [5, 29] or ocean rogue waves [8].

More details on all these extensions and alternatives to the Langevin Approach can be found in Friedrich et al. [6].

Discussion and conclusions

In this paper we present an *R* package for stochastic data analysis that is able to extract the stochastic evolution equations of physical properties from sets of their measurements.

The introduced functions serve as a framework to analyze one- and two-dimensional time series. They provide estimation of drift and diffusion coefficients describing the deterministic and the stochastic part of the analyzed process respectively. Integrating Langevin processes numerically enables one for cross-checking the obtained result and for generation of synthetic data sets.

Through illustrative examples we have shown that the Langevin evolution equation is able to uncover complex dynamics, even in cases when the associated statistics is identical to many other stochastic processes.

The presented package can be straightforwardly applied by *R*-users and it implies only a few preliminary tests to ascertain if all conditions on which the Langevin Approach is built are fulfilled. In case they are not, we briefly explain how to overcome them with simple extensions to the method that were already successfully applied in several applications [6].

Still, additional improvements of the presented functions are possible. For instance, instead of using the common average bin value when performing the binning of the data, one can incorporate a kernel-based regression of such values [15] or a maximum likelihood framework [11] for estimating the drift and diffusion functions.

Appendix: Different stochastic dynamics, same stationary distribution

In this appendix we show that a large family of two-point statistical distributions, each one univocaly defining one Langevin equation, Equation 1, corresponds to a one-point statistics given by the standard normal distribution

$$P_0(X) \propto \exp\left(-\frac{X^2}{2}\right), \tag{16}$$

i.e., a Gaussian distribution with zero mean and unit variance.

To that end, we start with one important remark concerning the evolution equation of one stochastic variable X, Equation 1: this equation is related to an another evolution equation, namely the one of the probability density function (PDF) of X, so-called Fokker-Planck Equation [25]:

$$\frac{\partial P(X)}{\partial t} = \left(-\frac{\partial}{\partial X} D^{(1)}(X) + \frac{\partial^2}{\partial X^2} D^{(2)}(X)\right) P(X). \tag{17}$$

The stationary solution $\left(\frac{\partial P}{\partial t} = 0\right)$ of the one-dimension Fokker-Planck is given by Risken [25]:

$$P(X) \propto \frac{1}{D^{(2)}(X)} \exp\left(\int_X \frac{D^{(1)}(x)}{D^{(2)}(x)} dx\right). \tag{18}$$

For the simple Ornstein-Uhlenbeck process, governed by Equation 1 with $D^{(1)} = -x$ and $D^{(2)} = 1$, the stationary PDF reduces to P_0 in Equation 16.

One could, however, consider a much more complex dynamics such as the one exemplified in this paper, with a bistable (cubic) drift coefficient and a non-constant diffusion, depending quadratically on the stochastic variable X:

$$D^{(1)}(X) = aX(b - X)(b + X), \tag{19a}$$

$$D^{(2)}(X) = c + dX^2. \tag{19b}$$

Here, $D^{(1)}$ has two stable fixed points at $\pm b$, with a maximum amplitude between them proportional to a, while $D^{(2)}$ has a minimum value c and a broadness proportional to $1/d$.

Substituting the cubic drift and the quadratic diffusion, given in Equations 19, into the stationary solution, Equation 18, and integrating, yields the stationary solution:

$$\begin{aligned} P(X) &\propto \frac{1}{c + dX^2} \exp\left(\int_X \frac{ab^2 x - ax^3}{c + dx^2} dx\right) \\ &= \frac{1}{d}\left(\frac{c}{d} + X^2\right)^{\frac{a}{2d}\left(b^2 + \frac{c}{d}\right) - 1} \exp\left(-\frac{X^2}{2\frac{d}{a}}\right). \end{aligned} \tag{20}$$

As one sees, the solution in Equation 20 has, in general, not only a Gaussian part, like Ornstein-Uhlenbeck processes, but also a polynomial part with an exponent depending on all parameters of $D^{(1)}$ and $D^{(2)}$. However, if the exponent is exactly zero,

$$\frac{a}{2d}\left(b^2 + \frac{c}{d}\right) - 1 = 0 \tag{21}$$

the polynomial part vanishes and the stationary solution reduces to the Gaussian distribution. In the example used in Section "Example for analyzing one dimensional data sets" with $a = b = c = d = 1$, this is the case (see **Figures 1** and **2** and Equation 14).

In general, the one-point statistic in the stationary regime given by Equation 18 yields Gaussian distributions even in more complex dynamics than the one here chosen. One only needs to have a drift coefficient given by one polynomial of odd degree $n > 0$ and simultaneously have a diffusion coefficient given by a polynomial of degree $n - 1$. In that case, whatever general expression both coefficients have, it is always possible to find a combination of their parameter values for which the quotient $D^{(1)}/D^{(2)}$ in the stationary solution reduces to a linear function in x yielding the PDF of a Gaussian distribution. The two-point statistic, $P(X(t)|X(t-\tau))$, however is able to distinguish between sets of $(D^{(1)}, D^{(2)})$ yielding the same one-point statistic.

The ambiguity of one-point statistics in characterizing the dynamics of stochastic processes in general, motivates the Langevin Approach implemented in our *R* package. Our approach has the advantage of being parameter free: since it computes numerically $D^{(1)}$ and $D^{(2)}$ without any given Ansatz, it can easily distinguish between higher-order drift and diffusion coefficients.

Quality control

All functions of the presented package have been tested against analytically solvable problems. Both example sections of this paper show how those tests where carried out in principle.

(2)Availability

Operating system

Any system capable of running R ≥ 3.0.2.

Programming language

R ≥ 3.0.2.

Dependencies

R ≥ 3.0.2, Rcpp ≥ 0.11.0 and RcppArmadillo ≥ 0.4.600.0.

List of contributors

1. Philip Rinn (Developer)
2. Pedro G. Lind (Contributed to `timeseries2D`)
3. David Bastine (Contributed to `timeseries1D`)

Software location

Archive

Name: CRAN

Persistent identifier: https://cran.r-project.org/web/packages/Langevin/

Licence: GPL-2+

Publisher: Philip Rinn
Version published: 1.1.1
Date published: 03/11/2015

Code repository
Name: GitLab, C.v.O. University of Oldenburg
Persistent identifier: https://gitlab.uni-oldenburg.de/TWiSt/Langevin
Licence: GPL-2+
Date published: 11/03/2016

Language
English.

(3) Reuse potential

The *Langevin* package serves as a basis for the analysis of a wide range of dynamic systems, therefor it is applicable for a wide range of scientific fields. We already outlined where the Langevin Approach was already used in the section "A glimpse beyond the *Langevin* package" and believe that this provides a good basis for potential users to judge if the presented package might be suitable for there research. The authors welcome interested developers to contribute code by mail or as a pull request in the code repository.

Notes

[1] The source code of these examples is availiable at https://gitlab.uni-oldenburg.de/TWiSt/Langevin/blob/master/examples.r

Acknowledgements
The authors thank Constantino Garcia Martínez (University of Santiago de Compostela, Spain) for useful discussions about our methods and motivating in sharing their implementation with a broader audience during his visit to our group.

Competing Interests
The authors declare that they have no competing interests.

References

1. **Böttcher, F, Peinke, J, Kleinhans, D, Friedrich, R, Lind, P G** and **Haase, M** 2006 Reconstruction of complex dynamical systems affected by strong measrement noise. *Physical Review Letters,* 97: 090603, DOI: http://dx.doi.org/10.1103/PhysRevLett.97.090603
2. **Dagum, L** and **Menon, R** 1998 OpenMP: An industry standard API for shared-memory programming. *Computational Science & Engineering,* IEEE, 5(1): 46–55, DOI: http://dx.doi.org/10.1109/99.660313
3. **Eddelbuettel, D** and **François, R** 2011 Rcpp: Seamless R and C++ integration. *Journal of Statistical Software,* 40(8): 1–18, URL http://www.jstatsoft.org/v40/i08/. DOI: http://dx.doi.org/10.18637/jss.v040.i08
4. **Eddelbuettel, D** and **Sanderson, C** RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis,* 71: 1054–1063, March 2014. DOI: http://dx.doi.org/10.1016/j.csda.2013.02.005
5. **Friedrich, R** and **Peinke, J** 1997 Description of a turbulent cascade by a Fokker-Planck equation. *Physical Review Letters,* 78: 863–866, DOI: http://dx.doi.org/10.1103/PhysRevLett.78.863
6. **Friedrich, R, Peinke, J, Sahimi, M** and **Tabar, M R R** Approaching complexity by stochastic methods: From biological systems to turbulence. *Physics Reports,* 506(5): 87–162, September 2011. ISSN 0370–1573. DOI: http://dx.doi.org/10.1016/j.physrep.2011.05.003
7. **Gardiner, C W** 2004 *Handbook of Stochastic Methods.* Springer Verlag, Berlin, third edition, DOI: http://dx.doi.org/10.1007/978-3-662-05389-8
8. **Hadjihosseini, A, Peinke, J** and **Hoffmann, N P** 2014 Stochastic analysis of ocean wave states with and without rogue waves. *New Journal of Physics,* 16(5): 053037, DOI: http://dx.doi.org/10.1088/1367-2630/16/5/053037
9. **Hänggi, P** and **Thomas, H** 1982 Stochastic processes: Time-evolution, symmetries and linear response. *Physics Reports,* 88: 207–319, DOI: http://dx.doi.org/10.1016/0370-1573(82)90045-X
10. **Honisch, C** and **Friedrich, R** Estimation of kramers-moyal coefficients at low sampling rates. *Physical Review E,* 83(6): 066701, June 2011. DOI: http://dx.doi.org/10.1103/PhysRevE.83.066701
11. **Kleinhans, D** Estimation of drift and diffusion functions from time series data: A maximum likelihood framework. *Physical Review E,* 85(2): 026705, February 2012. DOI: http://dx.doi.org/10.1103/PhysRevE.85.026705
12. **Kleinhans, D** and **Friedrich, R** 2007 Maximum likelihood estimation of drift and diffusion functions. *Physics Letters A,* 368(3–4): 194–198, DOI: http://dx.doi.org/10.1016/j.physleta.2007.03.082
13. **Kleinhans, D, Friedrich, R, Wächter, M** and **Peinke, J** Markov properties in presence of measurement noise. *Physical Review E,* 76: 041109, Oct 2007. DOI: http://dx.doi.org/10.1103/PhysRevE.76.041109
14. **Lade, S J** 2009 Finite sampling interval effects in kramers-moyal analysis. *Physics Letters A,* 373: 3705–3709, DOI: http://dx.doi.org/10.1016/j.physleta.2009.08.029
15. **Lamouroux, D** and **Lehnertz, K** 2009 Kernel-based regression of drift and diffusion coefficients of stochastic processes. *Physics Letters A,* 373: 3507–3512, DOI: http://dx.doi.org/10.1016/j.physleta.2009.07.073
16. **Lehle, B** Analysis of stochastic time series in the presence of strong measurement noise. *Physical Review E,* 83(2): 021113, February 2011. ISSN 1550–2376. DOI: http://dx.doi.org/10.1103/PhysRevE.83.021113
17. **Lehle, B** Stochastic time series with strong, correlated measurement noise: Markov analysis in n dimensions. *Journal of Statistical Physics,* 152(6): 1145–1169, September 2013. ISSN 1572–9613. DOI: http://dx.doi.org/10.1007/s10955-013-0803-z
18. **Lemon, J** 2006 Plotrix: A package in the red light district of R. *R-News,* 6(4): 8–12, URL https://cran.r-project.org/doc/Rnews/Rnews_2006-4.pdf.

19. **Lind, P G, Haase, M, Böttcher, F, Peinke, J, Kleinhans, D** and **Friedrich, R** 2010 Extracting strong measurement noise from stochastic time series: Applications to empirical data. *Physical Review E,* 81: 041125, DOI: http://dx.doi.org/10.1103/PhysRevE.81.041125
20. **Lind, P G, Herráez, I, Wächter, M** and **Peinke, J** 2014 Fatigue loads estimation through a simple stochastic model. *Energies,* 7: 8279–8293, DOI: http://dx.doi.org/10.3390/en7128279
21. **R Core Team** 2015 *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, URL http: //www.R-project.org/.
22. **Renner, Ch, Peinke, J** and **Friedrich, R** 2001 Experimental indications for Markov properties of small-scale turbulence. *Journal of Fluid Mechanics,* 433: 383–409, DOI: http://dx.doi.org/10.1017/S0022112001003597
23. **Riera, R** and **Anteneodo, C** 2010 Validation of drift and diffusion coefficients from experimental data. *Journal of Statistical Mechanics: Theory and Experiment,* 2010(04): P04020, DOI: http://dx.doi.org/10.1088/1742-5468/2010/04/P04020
24. **Rinn, P, Heißelmann, H, Wächter, M** and **Peinke, J** 2012 Stochastic method for in-situ damage analysis. *The European Physical Journal B,* 86: 1–5, ISSN 1434–6028. DOI: http://dx.doi.org/10.1140/epjb/e2012-30472-8
25. **Risken, H** 1996 *The Fokker-Planck Equation.* Springer-Verlag, DOI: http://dx.doi.org/10.1007/978-3-642-61544-3_4
26. **Sanderson, C** 2010 Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments. Technical report, NICTA, URL http://arma.sourceforge.net/armadillo_nicta_2010.pdf.
27. **Scholz, T, Raischel, F, Lopes, V V, Lehle, B, Wächter, M, Peinke, J** and **Lind, P G** 2015 Parameter-free resolution of the superposition of stochastic signals. *submitted,* URL http://arxiv.org/abs/1510.07285.
28. **Siegert, S, Friedrich, R** and **Peinke, J** 1998 Analysis of data sets of stochastic systems. *Physics Letters A,* 243: 275–280, DOI: http://dx.doi.org/10.1016/S0375-9601(98)00283-7
29. **Stresing, R** and **Peinke, J** 2010 Towards a stochastic multi-point description of turbulence. *New Journal of Physics,* 12(10): 103046, DOI: http://dx.doi.org/10.1088/1367-2630/12/10/103046
30. **Van Kampen, N G** 2007 *Stochastic Processes in Physics and Chemistry.* North-Holland Personal Library. Elsevier, Amsterdam, third edition.
31. **Wächter, M, Milan, P, Mücke, T** and **Peinke, J** 2011 Power performance of wind energy converters characterized as stochstic process: Applications of the langevin power curve. *Wind Energy,* 14: 711–717, DOI: http://dx.doi.org/10.1002/we.453
32. **Wilcoxon, F** 1945 Individual comparisons by ranking methods. *Biometrics,* 1: 80–83, DOI: http://dx.doi.org/10.2307/3001968

How to cite this article: Rinn, P, Lind, P G, Wächter, M and Peinke, J 2016 The Langevin Approach: An R Package for Modeling Markov Processes. *Journal of Open Research Software,* 4: e34, DOI: http://dx.doi.org/10.5334/jors.123

Submitted: 14 March 2016 **Accepted:** 01 July 2016 **Published:** 23 August 2016

]u[*Journal of Open Research Software* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS