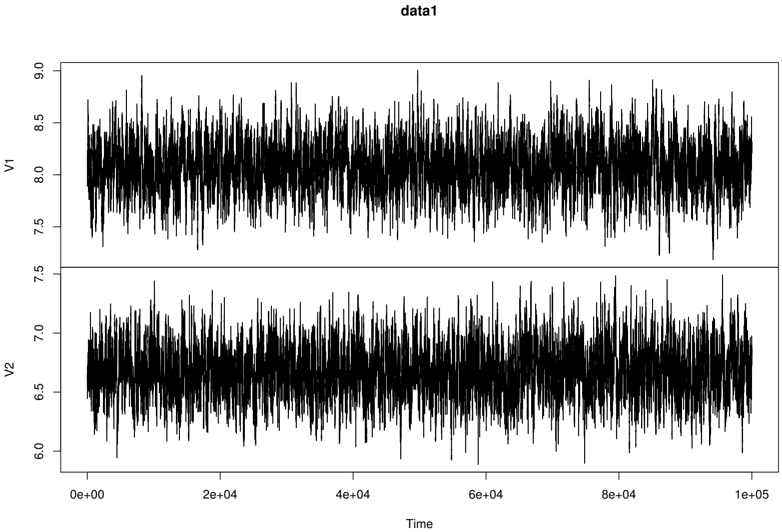
Project 6: Modeling unknown data

J. Ślęzak, C. Amorino, P. Andersson, M. Cornely, M. Eskelinen,
D. Fidanov, D. Salgado

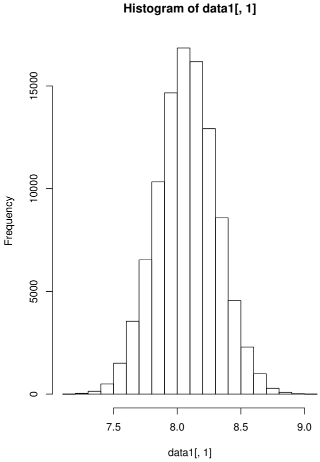
ECMI MW 2016

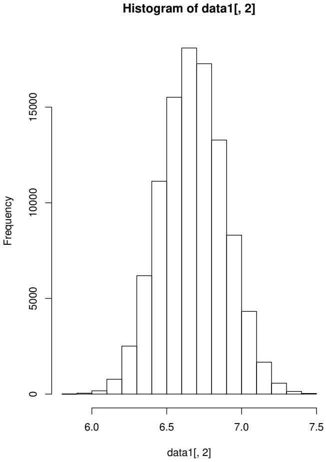
We found many directories containing files with two columns of some experimental data.

We tried plotting the numbers in one of them as a time series:

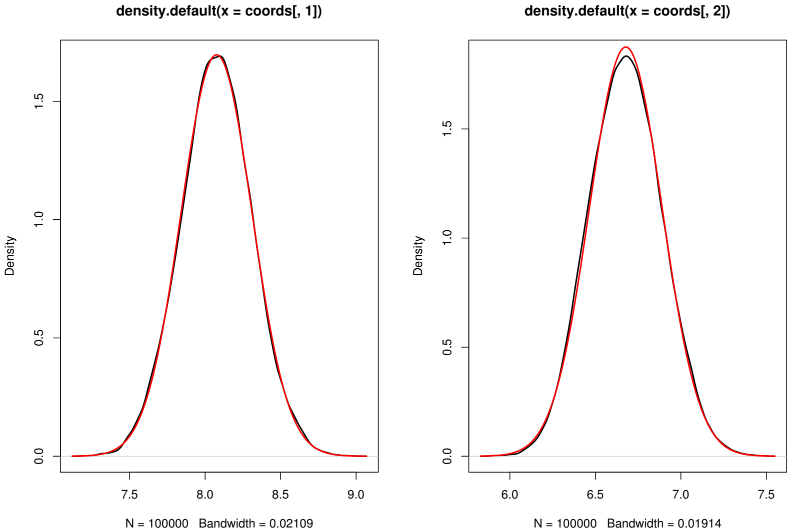


Let's look at the distributions of values for the two variables. Seems pretty Gaussian. . .

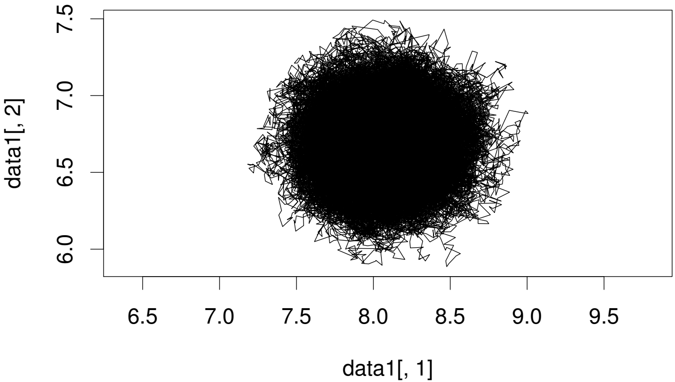




Plotting the kernel density and the theoretical Gaussian, it seems pretty close, we'll roll with it.



Wait, what about correlation between the two variables? Let's do a line plot:



Not much relation between the variables, but the overall behaviour is not completely random either. Let's compare...

Figure: Left: First 2000 pairs from the data. Right: 2000 random pairs of independent white Gaussian noise

The jumps between the consecutive points are small compared to actual random variables. Hints of dynamics...

If there are dynamics behind the data, we should see the later values depending on the earlier ones. Can we measure it somehow?

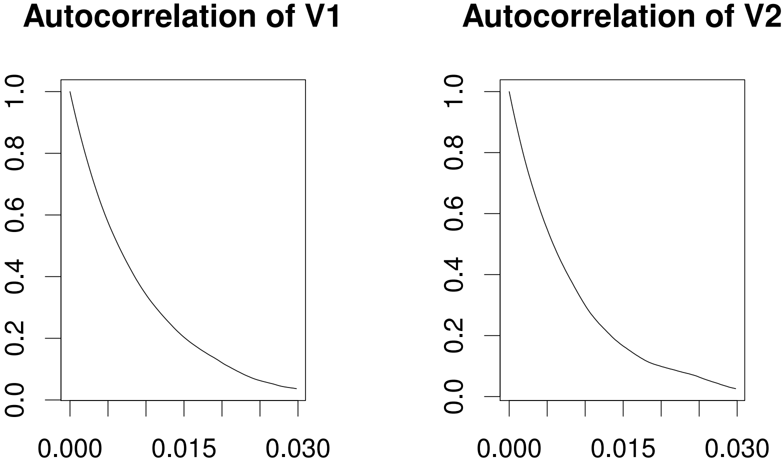
If there are dynamics behind the data, we should see the later values depending on the earlier ones. Can we measure it somehow?

Yes!

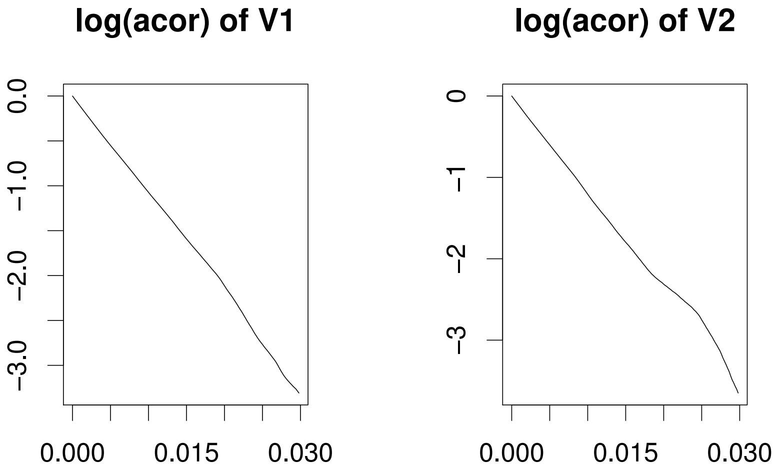
For a variable X with mean μ and variance σ^2, we can define the autocorrelation as a function of time as

$$\mathrm{acor}(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2}.$$

This tells us the dependence between values of X that are time τ apart.



Noise would only have a peak at 0, so we have something! Let's check these in log scale...



The dependence between values seems to decay exponentially (for a while...)

The only stochastic process which produces this decay for data that seems Gaussian is the *Ornstein-Uhlenbeck process*,

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -kx(t) + \eta(t).$$

Here

- k is the stiffness parameter for a harmonic force
- η is white Gaussian noise

The solution of the equation is

$$x(t) = \int_{-\infty}^{t} e^{-k(t-s)} \mathrm{d}B(s),$$

with B denoting Brownian motion ($\mathrm{d}B(s) = \eta(s)\mathrm{d}s$).
We find the autocorrelation for this process to be

$$\begin{aligned}
\mathrm{acor}(x(\tau)) &= \frac{E[x(0)x(\tau)]}{\sigma^2} \\
&= \dots \\
&= e^{-kt}.
\end{aligned}$$

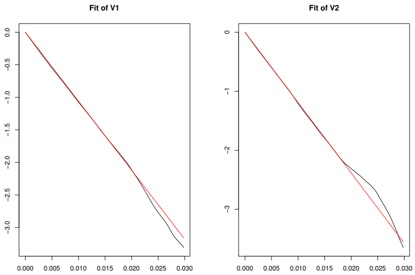
The solution of the equation is

$$x(t) = \int_{-\infty}^{t} e^{-k(t-s)} \mathrm{d}B(s),$$

with B denoting Brownian motion ($\mathrm{d}B(s) = \eta(s)\mathrm{d}s$).
We find the autocorrelation for this process to be

$$\begin{aligned} \mathrm{acor}(x(\tau)) &= \frac{E[x(0)x(\tau)]}{\sigma^2} \\ &= \dots \\ &= e^{-kt}. \end{aligned}$$

We can fit this to the autocorrelation of the data to determine k!



The fit gives us the values of k for variables V1 and V2

$$k_{V1} = 106.00,$$
$$k_{V2} = 119.25.$$

How do we know if the k we find is right? *Modelling!*
We can integrate the Ornstein-Uhlenbeck process between two consecutive points Δt apart to derive a formula for simulation:

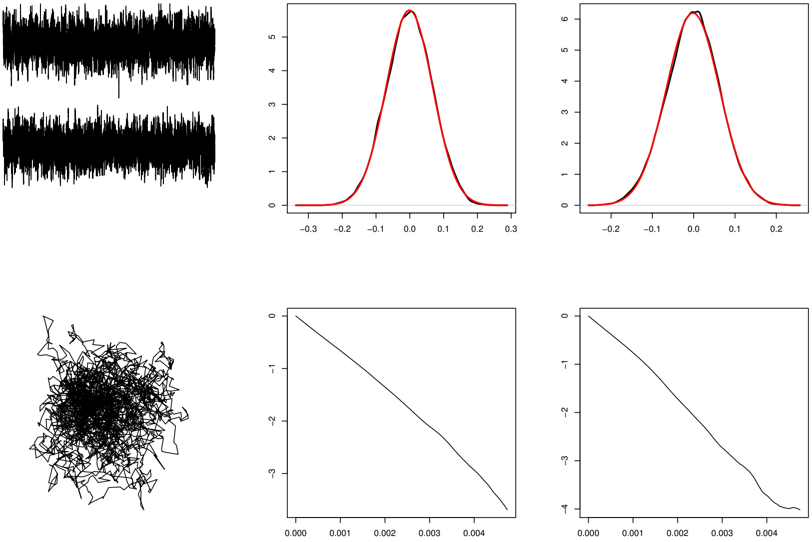
$$x_{n+1} = \alpha x_n + z_n,$$

where

$$\alpha = e^{-k\Delta t}$$

and z_n is a white Gaussian noise with variance $(1-\alpha^2)\sigma^2$.
The filename mentions a frequency of 5kHz, so the time difference Δt between the points is probably 0.2 ms.

Let's calculate everything from the model using Δt and fitted k values:



For the simulated data, the linear fit results in the k values

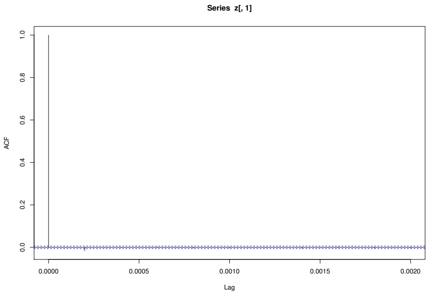
$$k_{V1} = 109.37,$$
$$k_{V2} = 138.61,$$

which mean that for the first and second variables we are within 4% and 17% of the actual values in the simulation, respectively.

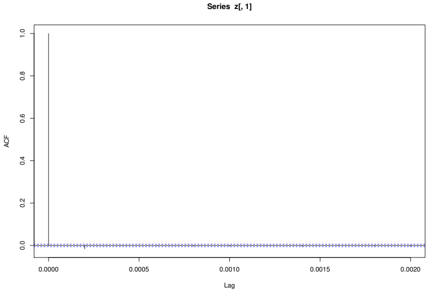
Reversing the exact discrete model, we can also solve for the noise z_n:

$$\begin{aligned} x_{n+1} &= \alpha x_n + z_n \\ z_n &= x_{n+1} - \alpha x_n. \end{aligned}$$

If we know α, we can now calculate z_n from the data!



Calculating the autocovariance shows only a value of 1 at 0, indicating that the values z_n are in fact white noise.



Calculating the autocovariance shows only a value of 1 at 0, indicating that the values z_n are in fact white noise.

Conclusion:

The data can be modeled quite well as an Ornstein-Uhlenbeck process (and we didn't even need to know it was describing optical tweezers)!