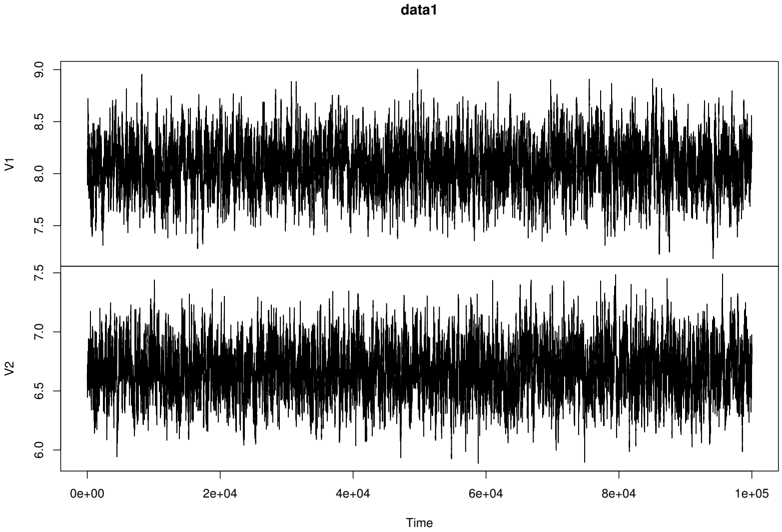
Thousands of plots

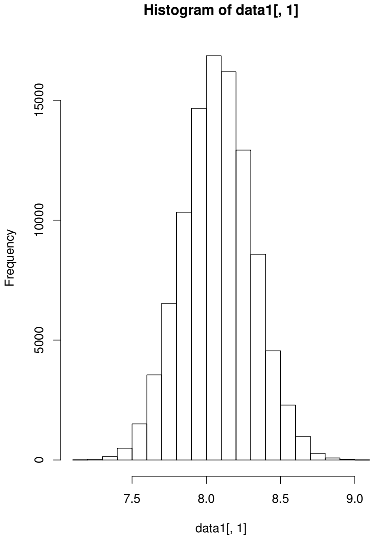
J. Ślęzak, M. Eskelinen, P. Andersson, M. Cornely, D. Salgado, C. Amorino, D. Fidanov

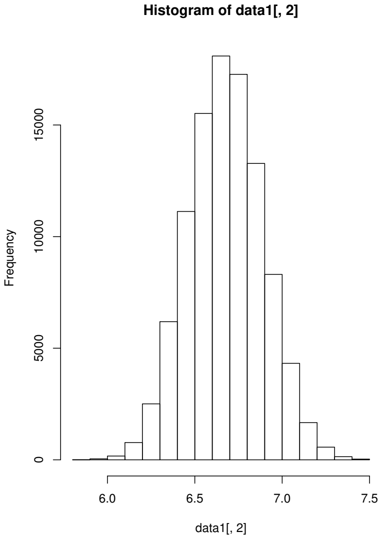
July 22, 2016

We found many folders containing cryptic .txt files. 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:





It looks a bit Gaussian. What say Kolmogorov and Smirnov?
For variable 1, normality is not very probable, but not excluded:

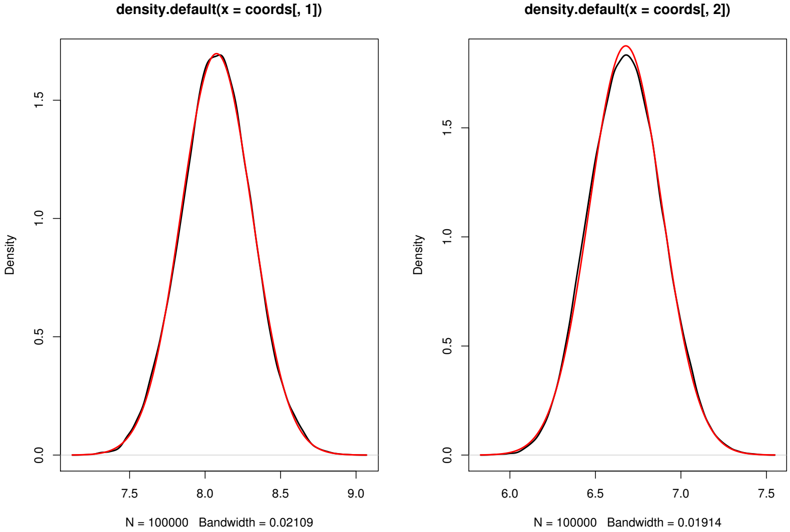
$$\text{p-value} = 0.08741$$

For variable 2, it's very abnormal:

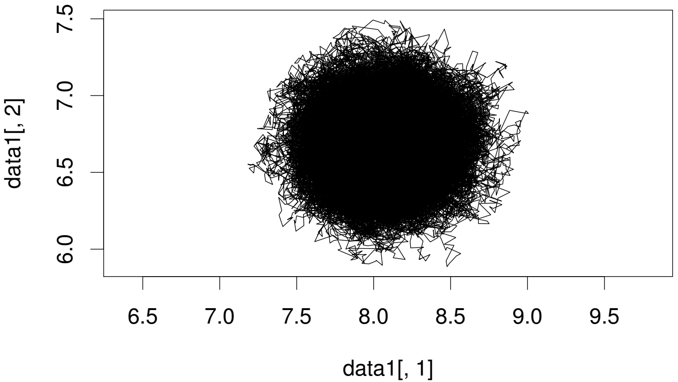
$$\text{p-value} = 1.972e - 05$$

Let's look at the density anyway...

Pretty close, we'll roll with it.



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



Not much relation, but it's not completely random either. Let's compare. . .

Figure: Left: First 2000 pairs. Right: 2000 random pairs

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

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

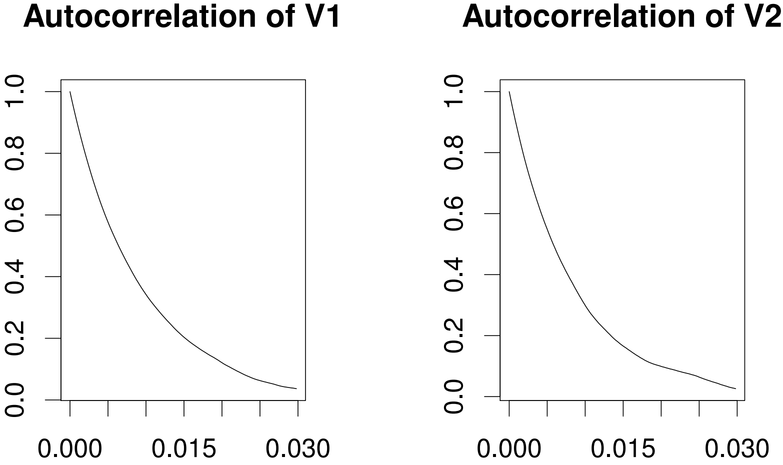
If there are dynamics, we should see the later values depending on the earlier ones. Can we measure it somehow?

Yes!

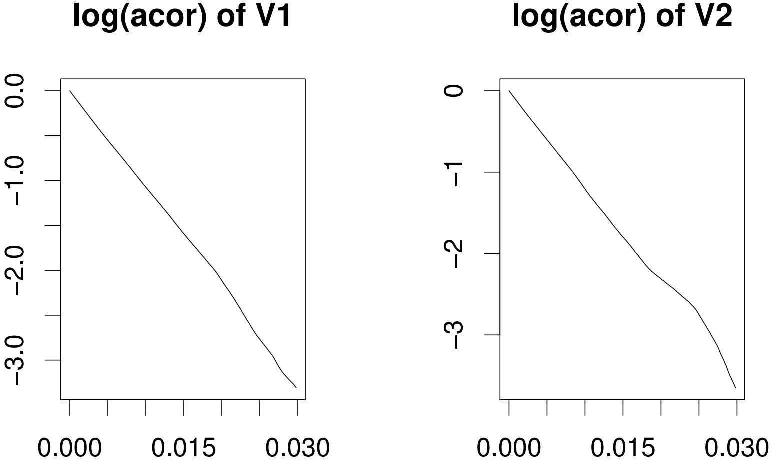
For a variable X with mean μ and variance σ^2, we can define the autocorrelation as a function of time as

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

This tells us the dependence between values of X that are time τ apart. The filename mentions a frequency of 5kHz, so the time difference between the points is probably 0.2 ms.



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...)

Since the data is Gaussian, the only stochastic process (dynamic) which produces this decay is the *Ornstein-Uhlenbeck process*,

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

Here

- k is the stiffness of 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

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

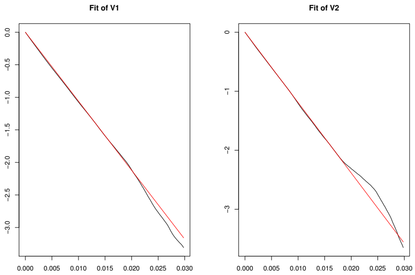
The solution of the equation is

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

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 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:

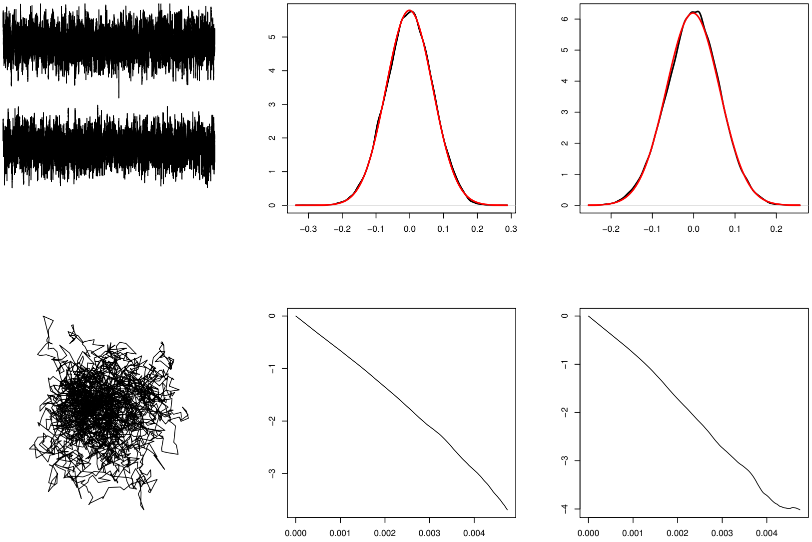
$$x_{n+1} = \alpha x_n + z_n,$$

where

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

and z_n is a white Gaussian noise with variance $(1-\alpha^2)\sigma^2$.

Let's calculate everything from the model using the k values we found:



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.

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.

Conclusion:

The data can be modeled quite well as an Ornstein-Uhlenbeck process!