



OST

Ostschweizer
Fachhochschule

Introduction to recursive probabilistic learning

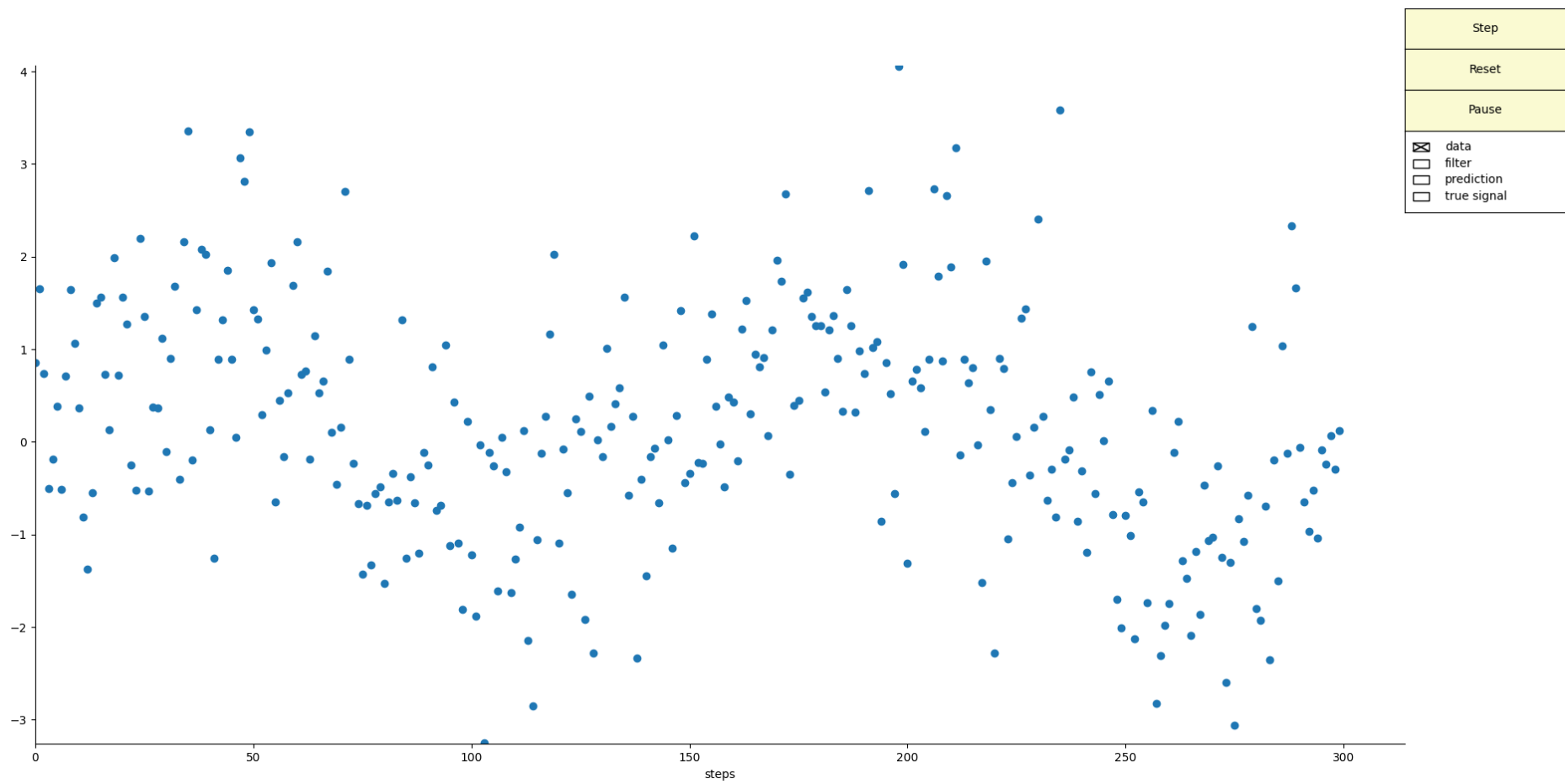
Modelling

JuanPi Carbajal
01.10.2021

Institute for Energy Technology

Overview

Our goal



source: `s_filtering_interactive.py`

The goal of the workshop is that you are able to understand and configure Kalman filters. You should be able to understand when to use it, how to use it, and what to do in case it is not applicable.

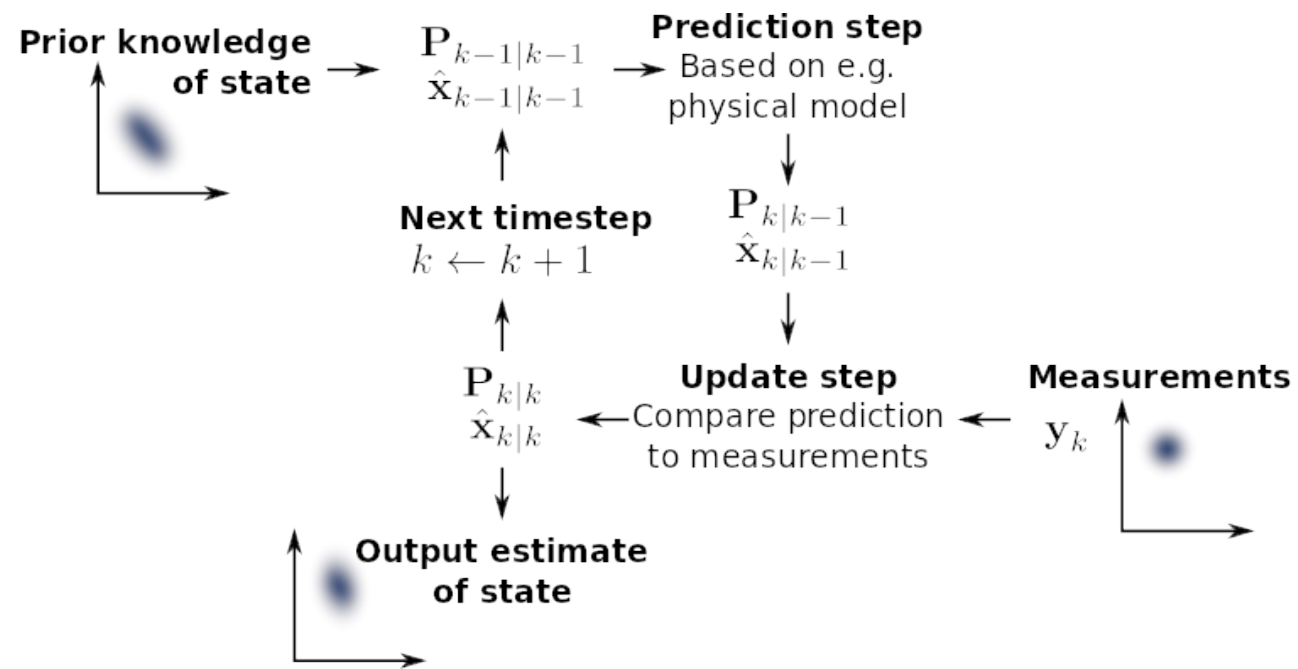
The needed ingredients

- An initial state → **Prior** distribution
- Model for the dynamics of the inferred states: how the states change over time → Linear Iterated maps
- Model for how the measurements are obtained from the states → Linear Map, **likelihood**
- A method to update the inferred states → Bayes rule, **posterior** distribution

Expert knowledge is realized in the models of the dynamics and the measurements.

Model dynamics (structure and parameter values) can also be learned from data (*out fo scope*: data driven dynamical systems, system identification). Given the model structure parameters can be tuned to the data (*within scope*: parameter estimation)

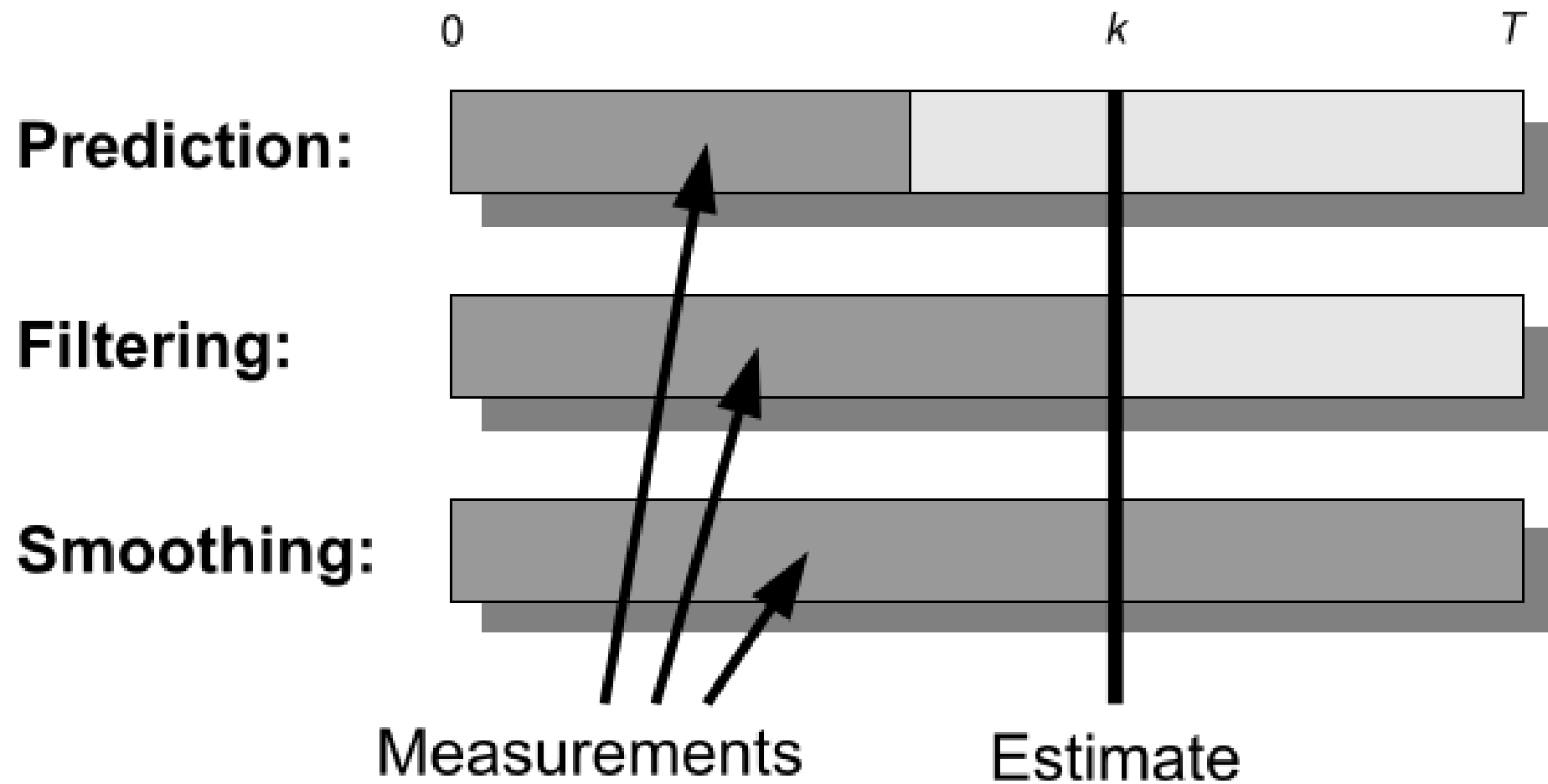
The needed ingredients



source: Wikipedia, Kalman filter

Our goal of today is to understand the **prediction step**. In the following session we will cover the **update step**. Finally we will work several examples.

Filtering and smoothing



source: Särkkä, S. doi:10.1017/CBO9781139344203

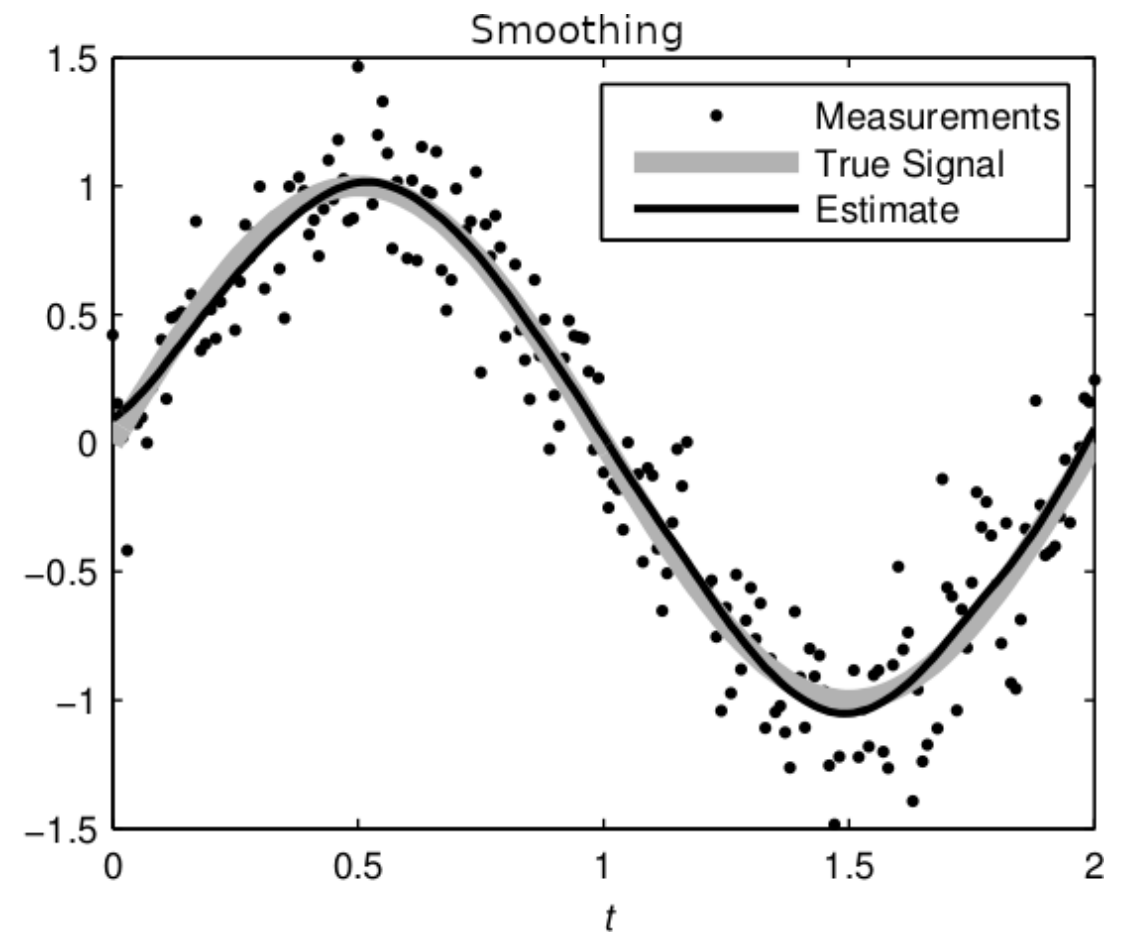
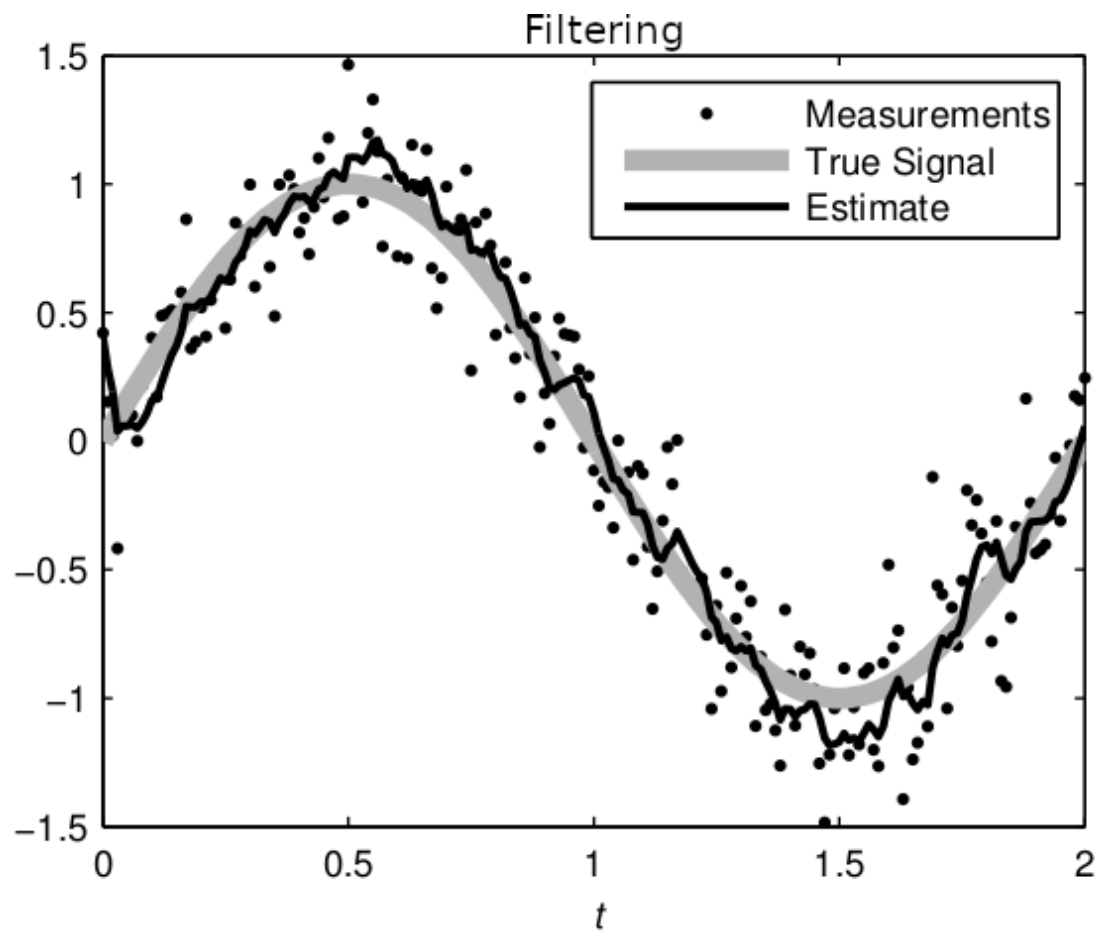
State estimation problems can be divided into optimal prediction, filtering, and smoothing depending on the time span of the measurements available with respect to the time of the estimated state.

Filtering: current state given the previous and current measurements.

Prediction: future state beyond the current state.

Smoothing: current state given previous, current, and future measurements.

Filtering and smoothing



source: Särkkä, S. doi:10.1017/CBO9781139344203

Smoothing uses future information to update the past, hence the estimation is smoother.

Let's start with prediction.

Iterated maps

Iterated map: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y$$

What do we get if $\alpha = 1$?

Iterated map: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} \quad |\alpha| \leq 1$$

Do you expect a straight trajectory?

What do we get if $\alpha = 1$ and $\alpha = -1$?

Iterated map: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} \quad |\alpha| \leq 1$$

0. x_0

Iterated map: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} \quad |\alpha| \leq 1$$

0. x_0
1. $x_1 = \alpha x_0$

Iterated map: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} \quad |\alpha| \leq 1$$

0. x_0
1. $x_1 = \alpha x_0$
2. $x_2 = \alpha x_1 = \alpha \alpha x_0 = \alpha^2 x_0$

Iterated map: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} \quad |\alpha| \leq 1$$

0. x_0
1. $x_1 = \alpha x_0$
2. $x_2 = \alpha x_1 = \alpha \alpha x_0 = \alpha^2 x_0$
3. $x_3 = \alpha x_2 = \alpha \alpha^2 x_0 = \alpha^3 x_0$

Iterated map: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} \quad |\alpha| \leq 1$$

0. x_0
1. $x_1 = \alpha x_0$
2. $x_2 = \alpha x_1 = \alpha \alpha x_0 = \alpha^2 x_0$
3. $x_3 = \alpha x_2 = \alpha \alpha^2 x_0 = \alpha^3 x_0$
4. ...

Iterated map: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y$$

Iterated map (use the result as the next value of x):

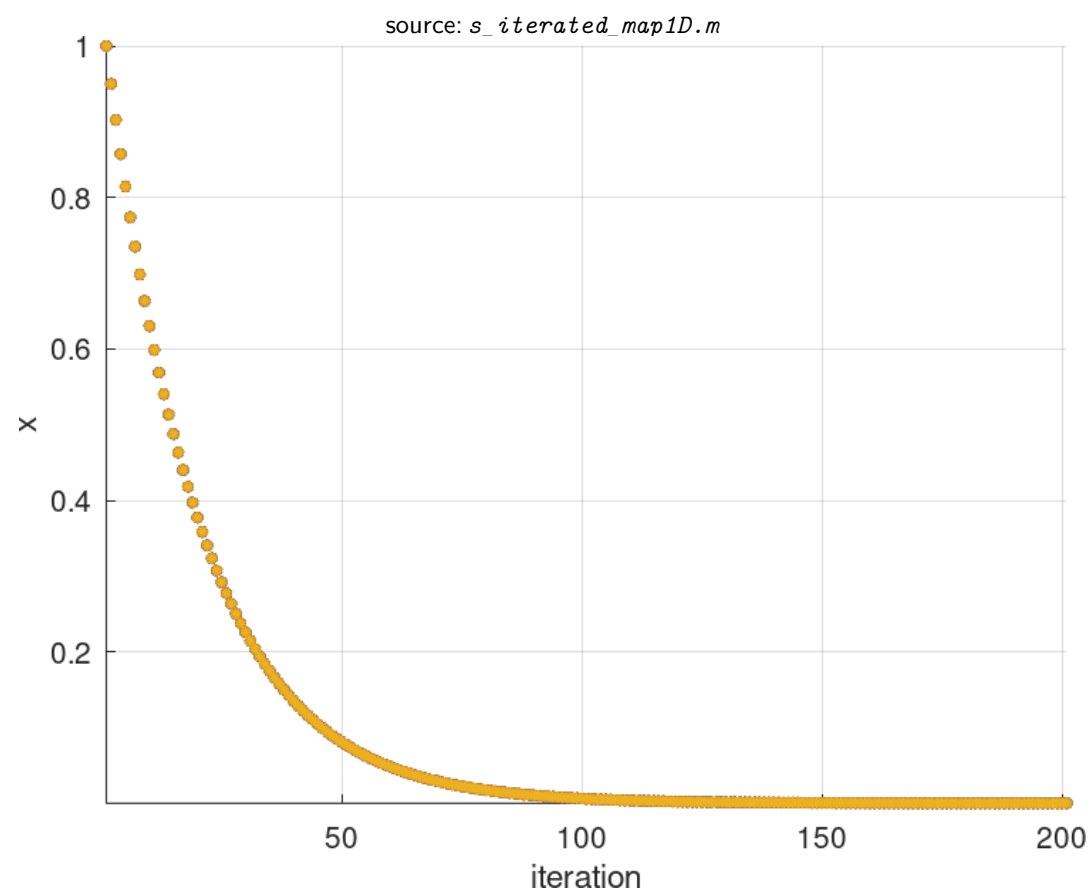
$$x_k = \alpha x_{k-1} \quad |\alpha| \leq 1$$

0. x_0
1. $x_1 = \alpha x_0$
2. $x_2 = \alpha x_1 = \alpha \alpha x_0 = \alpha^2 x_0$
3. $x_3 = \alpha x_2 = \alpha \alpha^2 x_0 = \alpha^3 x_0$
4. ...
5. $x_k = \alpha^k x_0$

The composition of linear maps gives a different linear map at each step: not a straight trajectory

Iterated map: single variable

$$x_k = (1 - 0.05)x_{k-1}$$



Observation: linear map but not straight trajectory

Test other values of $\alpha \in [-1, 1]$. For what values of α do you see qualitative change in behavior?

What happens when $\alpha < -1$ or $\alpha > 1$?

Iterated map: two variables

Transform two variables x_1, x_2 into y_1, y_2

$$y_1 = a_{11}x_1 + a_{12}x_2$$

$$y_2 = a_{21}x_1 + a_{22}x_2$$

The color of the indices suggest a way to order the equations. Do you have any idea?

Iterated map: two variables

Transform two variables x_1, x_2 into y_1, y_2

$$y_1 = a_{11}x_1 + a_{12}x_2$$

$$y_2 = a_{21}x_1 + a_{22}x_2$$

Organize the values in tables (matrices)

$$\text{matrix: } \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\text{single column: } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$\text{single row: } \mathbf{x}^\top = [x_1 \ x_2]$$

Define the product of rows and columns

$$[x \ y] \cdot \begin{bmatrix} u \\ w \end{bmatrix} = xu + yw$$

The shape (sometimes size) of a matrix with n rows and m columns is written $n \times m$. One can also specify the set to which the entries (or cells) of the matrix belong to, e.g. $\mathbb{R}^{2 \times 3}$ are matrices with 2 rows and 3 columns with entries in the real numbers. Other examples would be $\mathbb{Z}^{n \times n}$, $\{0, 1\}^{n \times m}$, etc.

If \mathbf{x} is a column vector, what gives $\mathbf{x}\mathbf{x}^\top$?

Work this out, because it will help with the definition of the variance of a random vector.

Iterated map: two variables

Transform two variables x_1, x_2 into y_1, y_2

$$y_1 = a_{11}x_1 + a_{12}x_2$$

$$y_2 = a_{21}x_1 + a_{22}x_2$$

Organize the values in tables (matrices)

$$\text{matrix: } \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\text{single column: } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$\text{single row: } \mathbf{x}^\top = [x_1 \ x_2]$$

The transformation of the variables now looks like

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

Define the product of rows and columns

$$[x \ y] \cdot \begin{bmatrix} u \\ w \end{bmatrix} = xu + yw$$

Re-write the transformation using row vectors.

Iterated map: two variables

Map (mapping, transformation, etc.)

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

Iterated map:

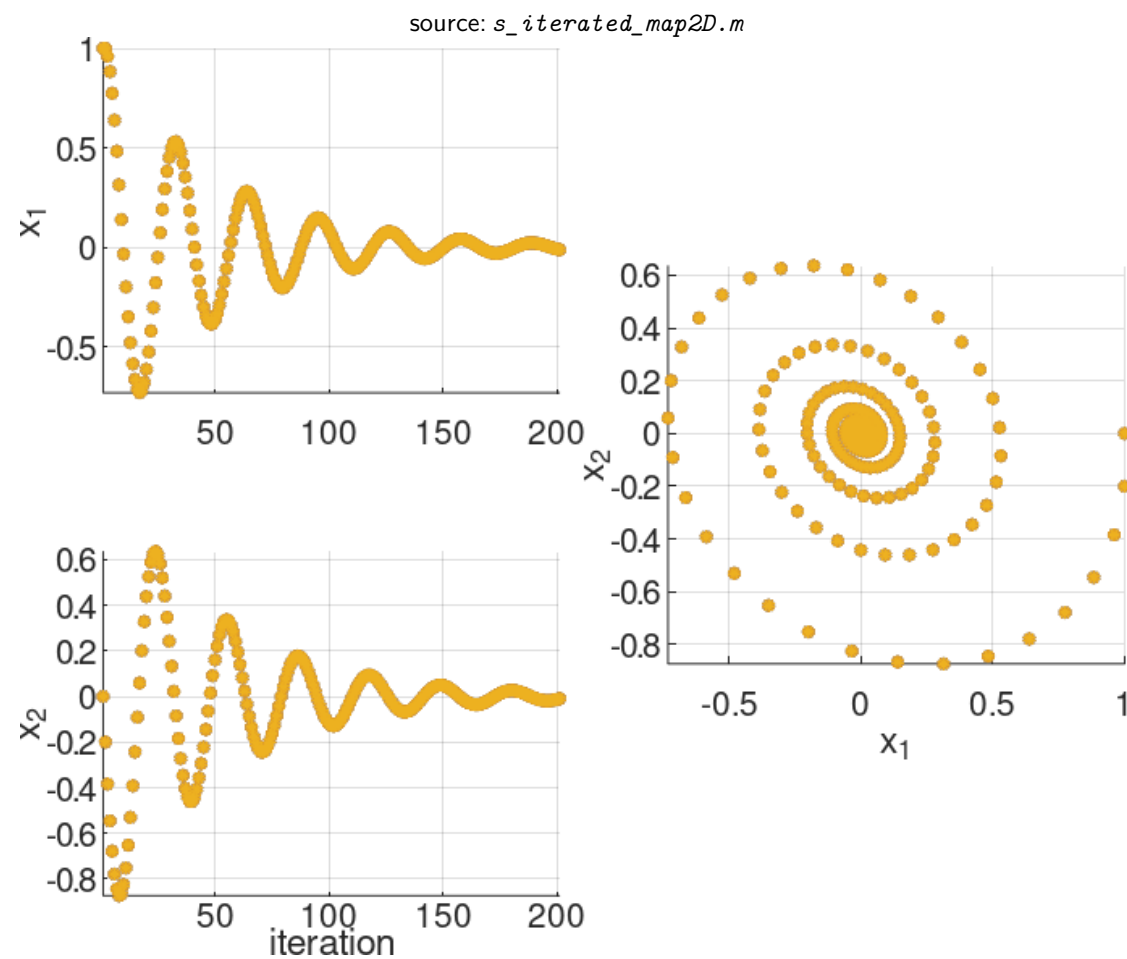
$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1}$$
$$\mathbf{A} = \mathbf{I} + \alpha \begin{bmatrix} 0 & 1 \\ -1 & -2\alpha \end{bmatrix} \quad 0 \leq \alpha \leq 1$$

0. \mathbf{x}_0
1. $\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0$
2. $\mathbf{x}_2 = \mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{A}\mathbf{x}_0 = \mathbf{A}^2\mathbf{x}_0$
3. $\mathbf{x}_3 = \mathbf{A}\mathbf{x}_2 = \mathbf{A}\mathbf{A}^2\mathbf{x}_0 = \mathbf{A}^3\mathbf{x}_0$
4. ...
5. $\mathbf{x}_k = \mathbf{A}^k\mathbf{x}_0$

The composition of linear maps gives a different linear map at each step: not a straight trajectory

Iterated map: two variables

$$\mathbf{x}_k = \left(\mathbf{I} + 0.2 \begin{bmatrix} 0 & 1 \\ -1 & -0.4 \end{bmatrix} \right) \mathbf{x}_{k-1}$$



Test other values of $\alpha \in [0, 1]$. For what values of α do you see qualitative changes in behavior?

What happens when $\alpha < 0$ or $\alpha > 1$?

Measurement model

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1}$$

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k \quad \text{Measurement model}$$

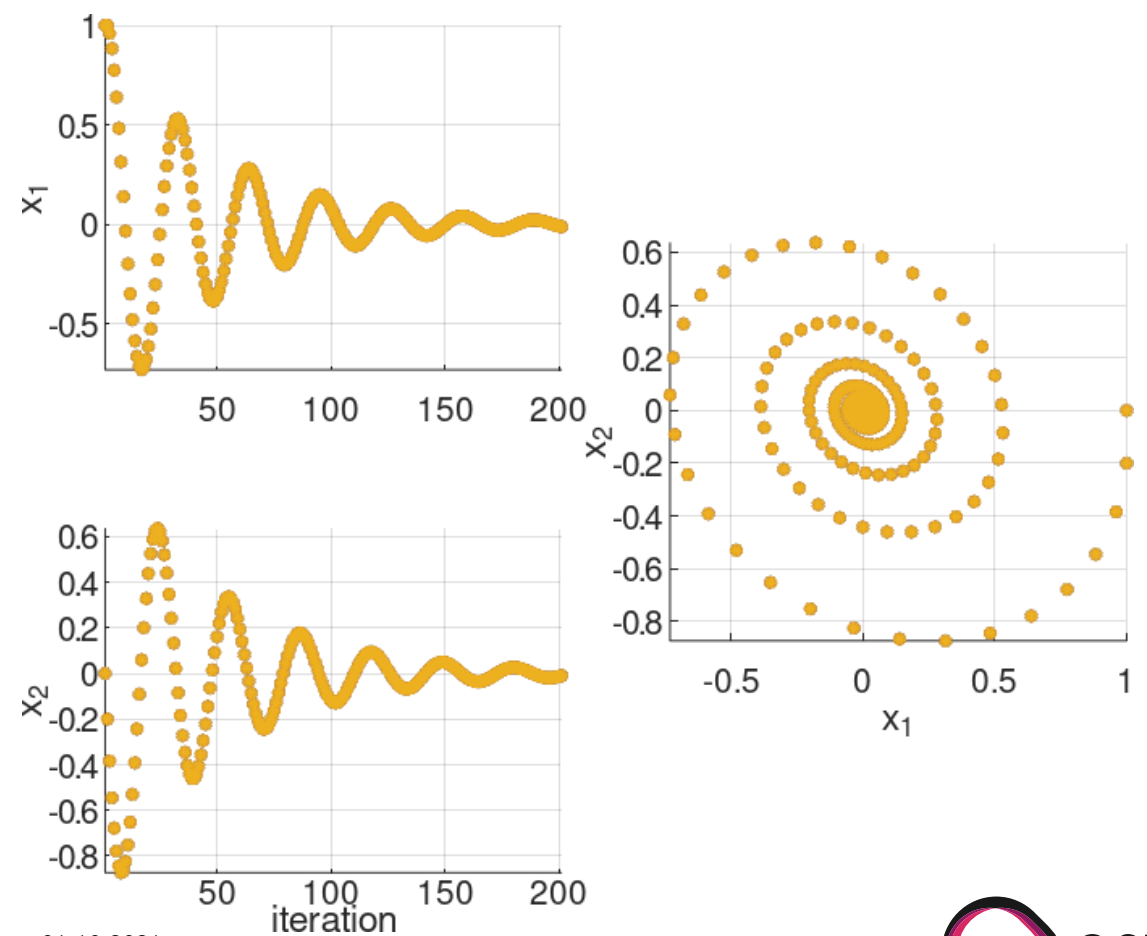
Measure 1st component: $\mathbf{H} = \begin{bmatrix} 1 & 0 \end{bmatrix}$

Measure 2nd component: $\mathbf{H} = \begin{bmatrix} 0 & 1 \end{bmatrix}$

What does this measure?

$$\mathbf{H} = \begin{bmatrix} -1 & 1 \end{bmatrix}$$

What's \mathbf{H} if we measure the mean of the two components?



The dimension of \mathbf{y} could be anything. For m measurements and a n dimensional state we have that: $\mathbf{x} \in \mathbb{R}^{n \times 1}$, $\mathbf{y} \in \mathbb{R}^{m \times 1}$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{H} \in \mathbb{R}^{m \times n}$.

In practise it is common to have m (number of measurements) lower than n . Hence \mathbf{H} is usually a wider than taller (short-fat matrix).

Modify `s_iterated_map2D.m` (or implement your own) and implement different measurement matrices \mathbf{H} . Plot the measurements.

Inputs

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_{k-1} \quad \text{Input matrix, inputs}$$

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k$$

For n states and k inputs, the input matrix has the shape $n \times k$.

Can you provide an example, even if it is 1-dimensional?

Inputs

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_{k-1} \quad \text{Input matrix, inputs}$$

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k$$

Example from kinematics

$$p_t = p_{t-1} + v_{t-1}\Delta t \quad v_t = v_{t-1} + \underbrace{a_{t-1}}_{\frac{F_{t-1}}{m}}\Delta t$$

define

$$\mathbf{x}_t = \begin{bmatrix} p_t \\ v_t \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}$$

$$u_t = a_t \quad \mathbf{B} = \begin{bmatrix} 0 \\ \Delta t \end{bmatrix}$$

then we get (assuming we observe only the position)

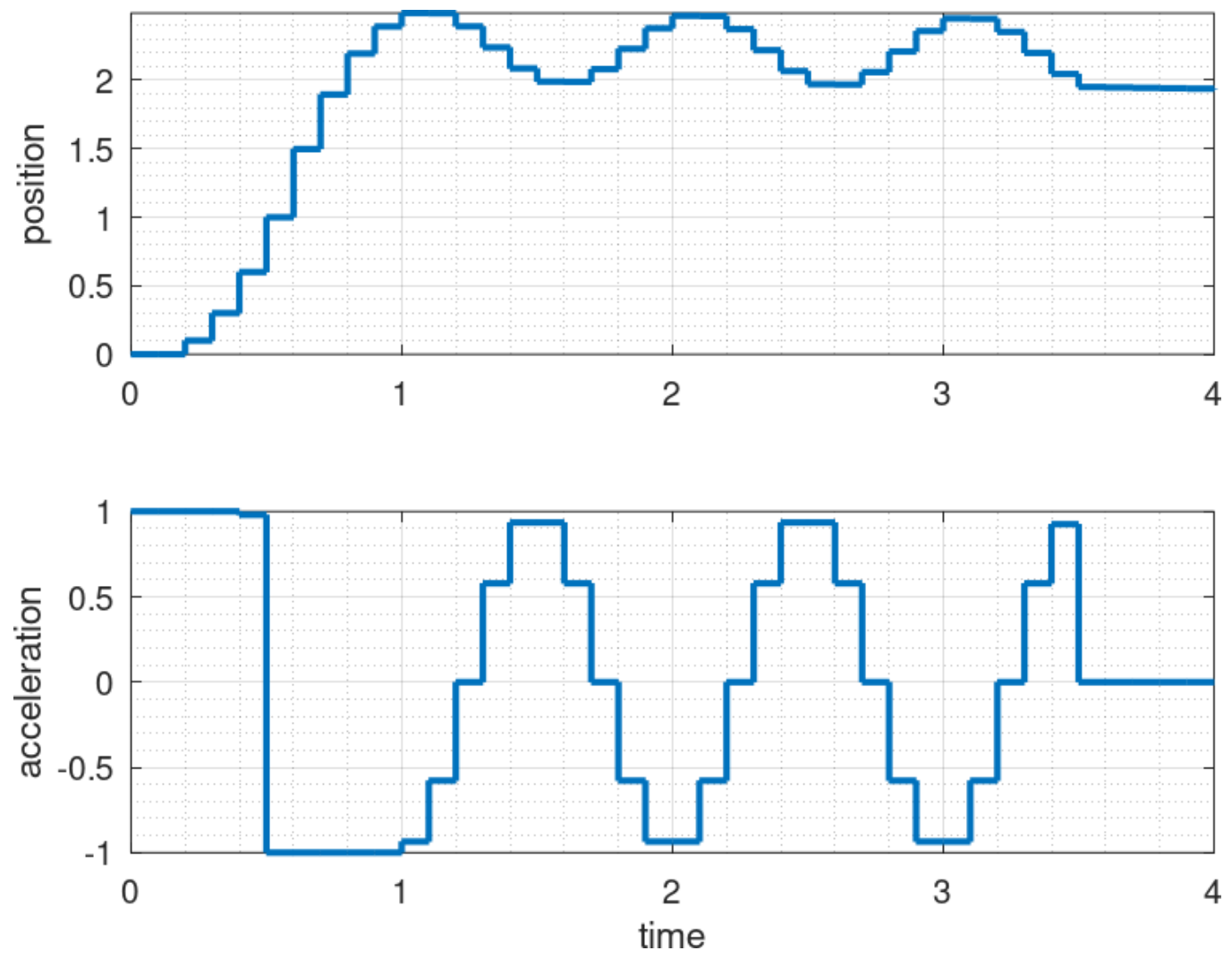
$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}u_{t-1}$$
$$y_t = [1 \ 0] \mathbf{x}_t$$

The example is the Euler method, which is not a good discretization unless Δt is very small. An alternative map would be

$$p_t = p_{t-1} + v_{t-1}\Delta t + \frac{1}{2}a_{t-1}\Delta t^2$$
$$\mathbf{B} = \begin{bmatrix} \frac{\Delta t^2}{2} \\ \Delta t \end{bmatrix}$$

and a_{t-1} is the mean acceleration between $t - 1$ and t .

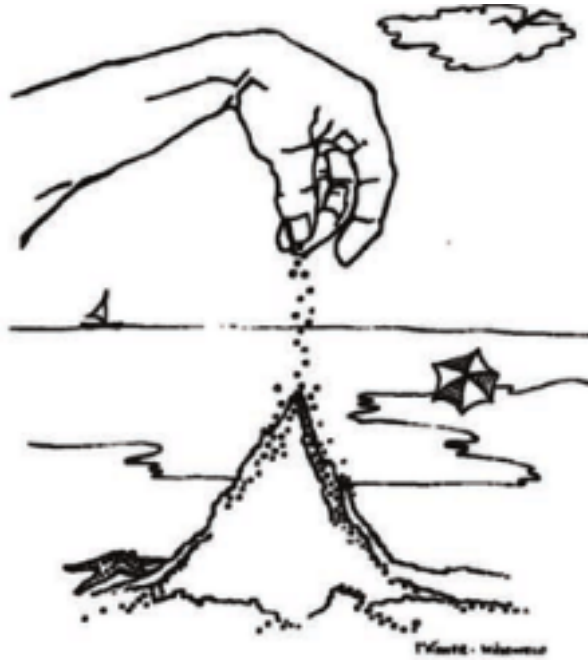
Inputs



Error propagation

Linearly related variables

Take two variables related by: $y = b + ax$



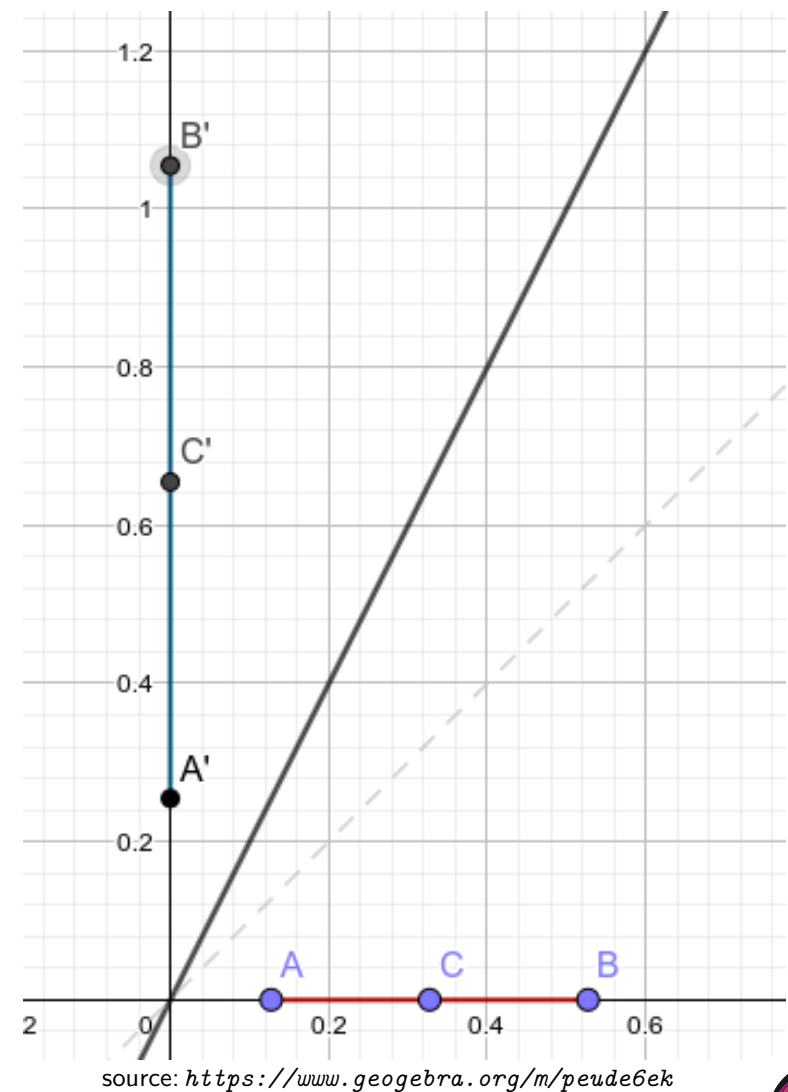
Linearly related variables

Take two variables related by: $y = b + ax$

At a given value x_o we get: $y_o = b + ax_o$

If we modify x_o by a given amount Δx :

$$\hat{y} = b + a(x_o + \Delta x) = \underbrace{b + ax_o}_{=y_o} + a\Delta x$$



source: <https://www.geogebra.org/m/peude6ek>

Linearly related variables

Take two variables related by: $y = b + ax$

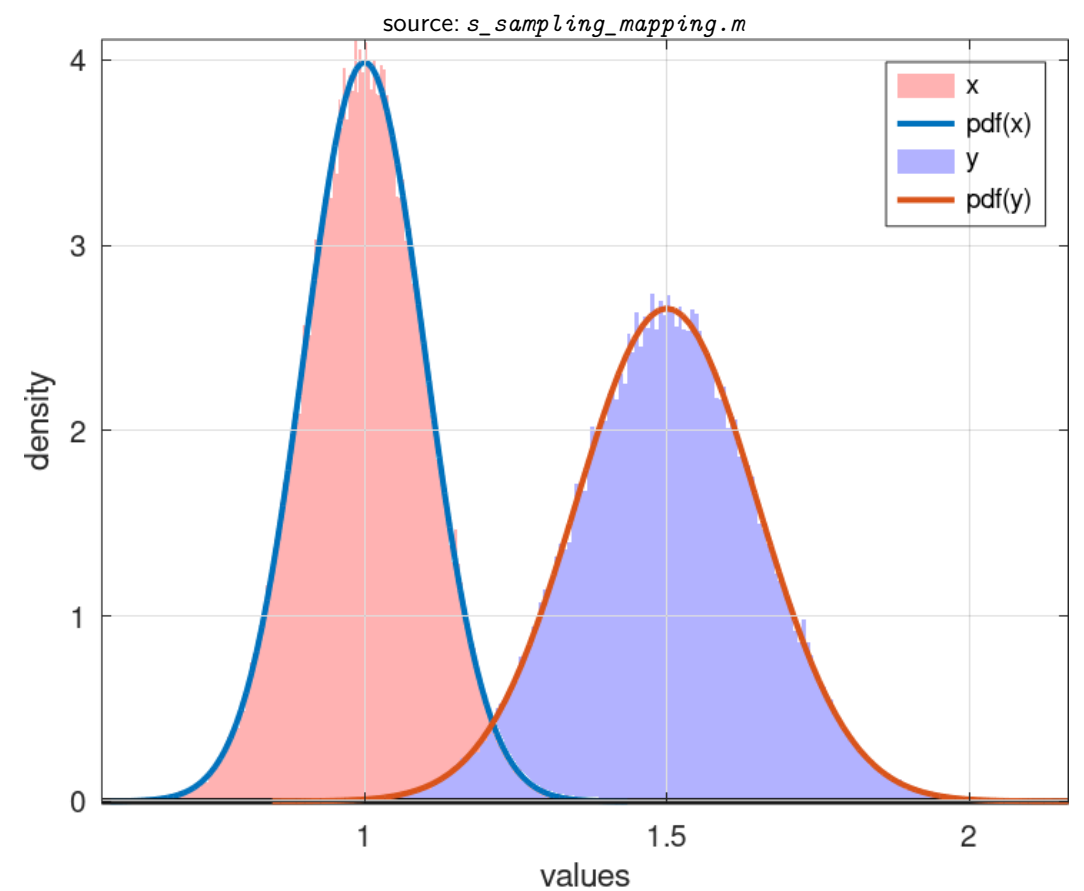
At a given value x_o we get: $y_o = b + ax_o$

If we modify x_o by a given amount Δx :

$$\hat{y} = b + a(x_o + \Delta x) = \underbrace{b + ax_o}_{=y_o} + a\Delta x$$

Then y_o changes an amount Δy :

$$\begin{aligned}\hat{y} &= y_o + \Delta y \\ \Delta y &:= a\Delta x\end{aligned}$$



Linearly related variables

Take two variables related by: $y = b + ax$

$$\Delta y := a\Delta x$$

The "error" Δx propagates to y via the slope a of the relation. The intercept b doesn't play a role in the induced "error" Δy .

Can you propagate and initial value error on an iterated map?

Error propagation in an 1D iterated map

$$x_k = ax_{k-1}$$

Using our previous result

$$\Delta x_k = a\Delta x_{k-1}$$

It follows the same dynamics as the state!

Let's propagate an initial value error

0. $x_0 + \Delta x_0$
1. $x_1 = a(x_0 + \Delta x_0)$
2. $x_2 = a^2(x_0 + \Delta x_0)$
3. ...
4. $x_k = a^k(x_0 + \Delta x_0)$

It is just the iterated map on a different initial condition!

Implement these formulas in the computer and run some simulations for the same x_0 and different values of Δx_0 .

For many simulations with different values of Δx_0 , what's the mean value of x_k ?

Open question: what if we made an error at each iteration step?

What changes if the state is multi-dimensional \mathbf{x}_k ? Consider these matrices:

$$\mathbf{A}_d = \frac{1}{2}\mathbf{I} \quad \mathbf{A}_m = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{3} & \frac{1}{2} \end{bmatrix}$$

and propagate an error in the 1st component of the initial value, i.e. $\mathbf{x}_o^\top = [x_{o1} + \Delta x_{o1} \quad x_{o2}]$

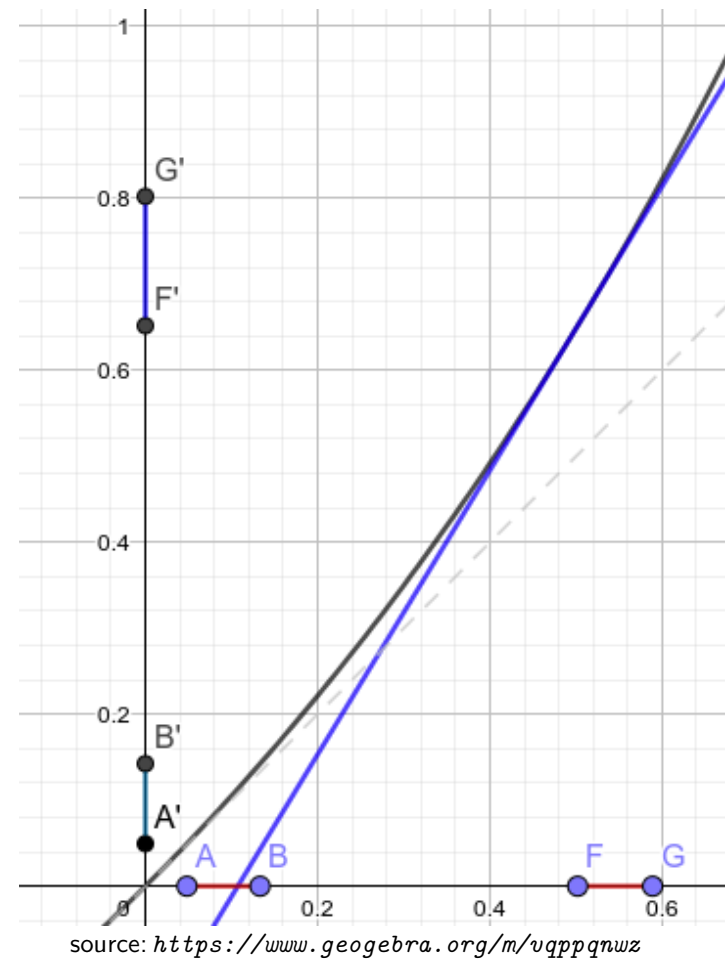
Non-linearly connected variables and small error

Consider two variables connected by a nonlinear relation: $y = f(x)$

At a given value x_o we get $y_o = f(x_o)$

If we modify x_o by a given amount Δx :

$$\hat{y} = f(x_o + \Delta x)$$

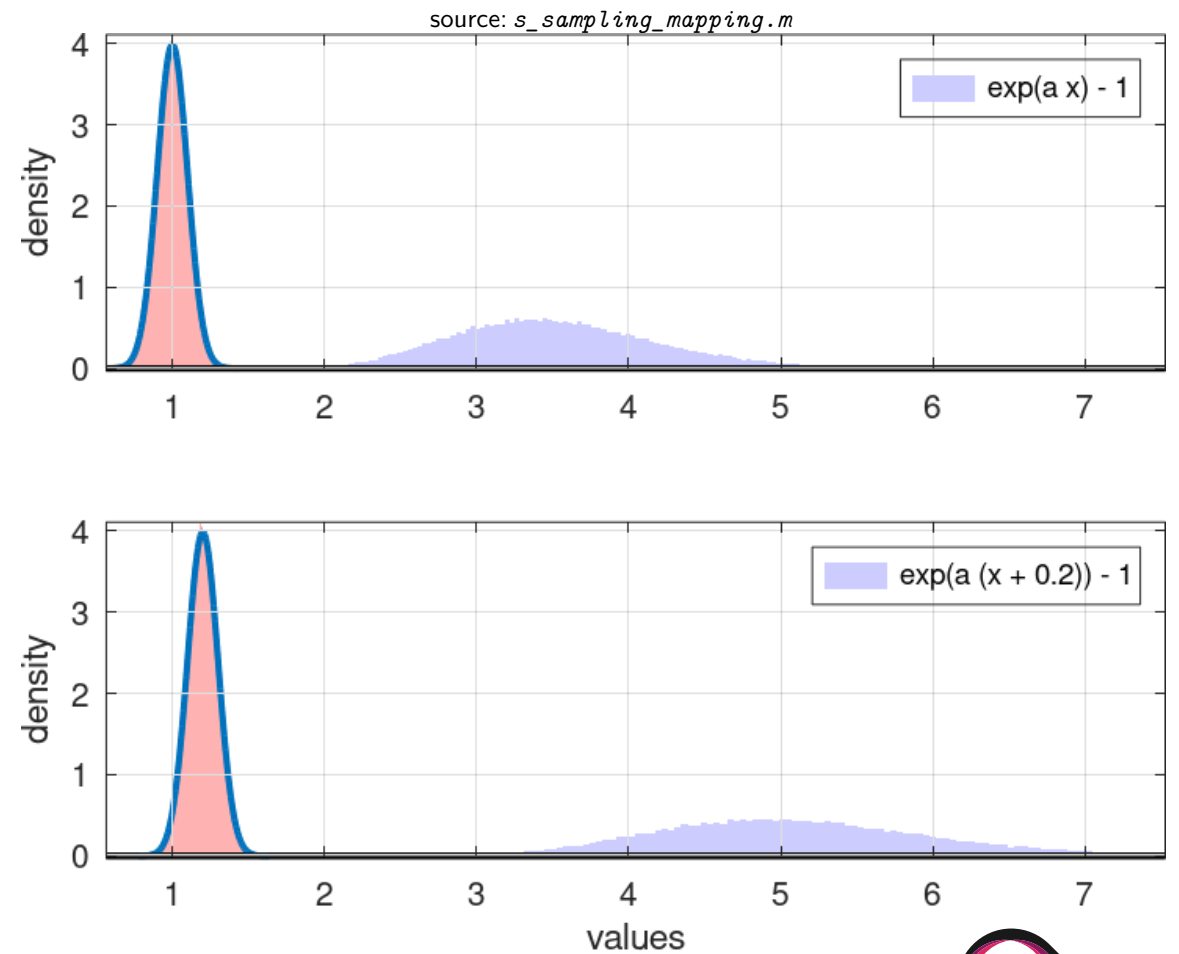


Non-linearly connected variables and small error

Consider two variables connected by a nonlinear relation: $y = f(x)$

If Δx is small (and f analytic), then we can proceed:

$$\hat{y} = f(x_o + \Delta x) \approx f(x_o) + \left. \frac{\partial f}{\partial x} \right|_{x_o} \Delta x$$
$$\hat{y} = y_o + \left. \frac{\partial f}{\partial x} \right|_{x_o} \Delta x = y_o + \Delta y$$
$$\Delta y := \left. \frac{\partial f}{\partial x} \right|_{x_o} \Delta x$$



Non-linearly connected variables and small error

Consider two variables connected by a nonlinear relation: $y = f(x)$

$$\Delta y := \left. \frac{\partial f}{\partial x} \right|_{x_0} \Delta x$$

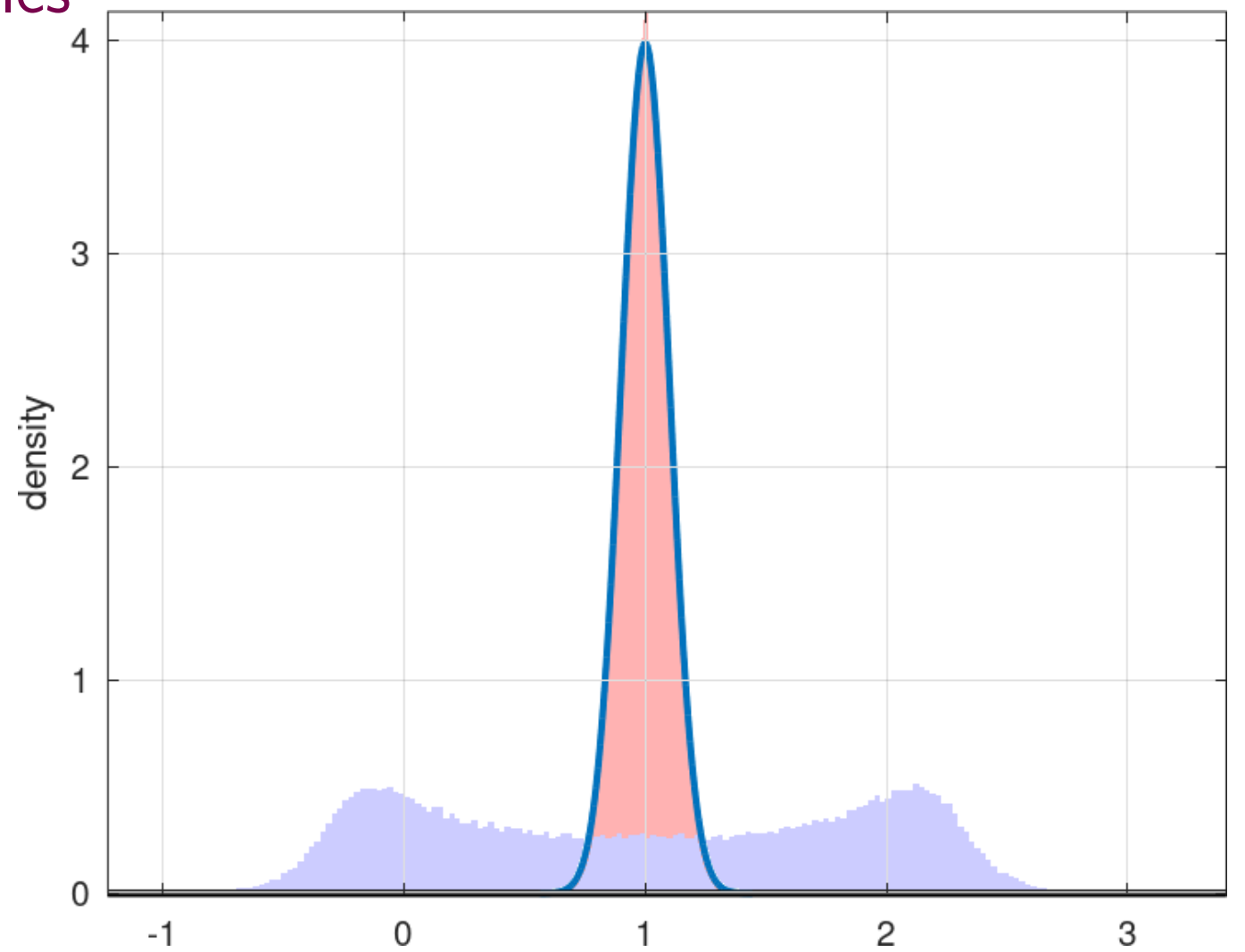
The small "error" Δx propagates to y via the **local** slope of the relation. For different values of x_0 the "error" propagates differently.

There are other approaches for the non-linear case that do not make the same assumptions we did here. This is not on the scope of the course. A good place to start would be Wikipedia's article on propagation of uncertainty https://en.wikipedia.org/wiki/Propagation_of_uncertainty.

Non-linearly connected variables

Nonlinear mappings can radically change the distribution

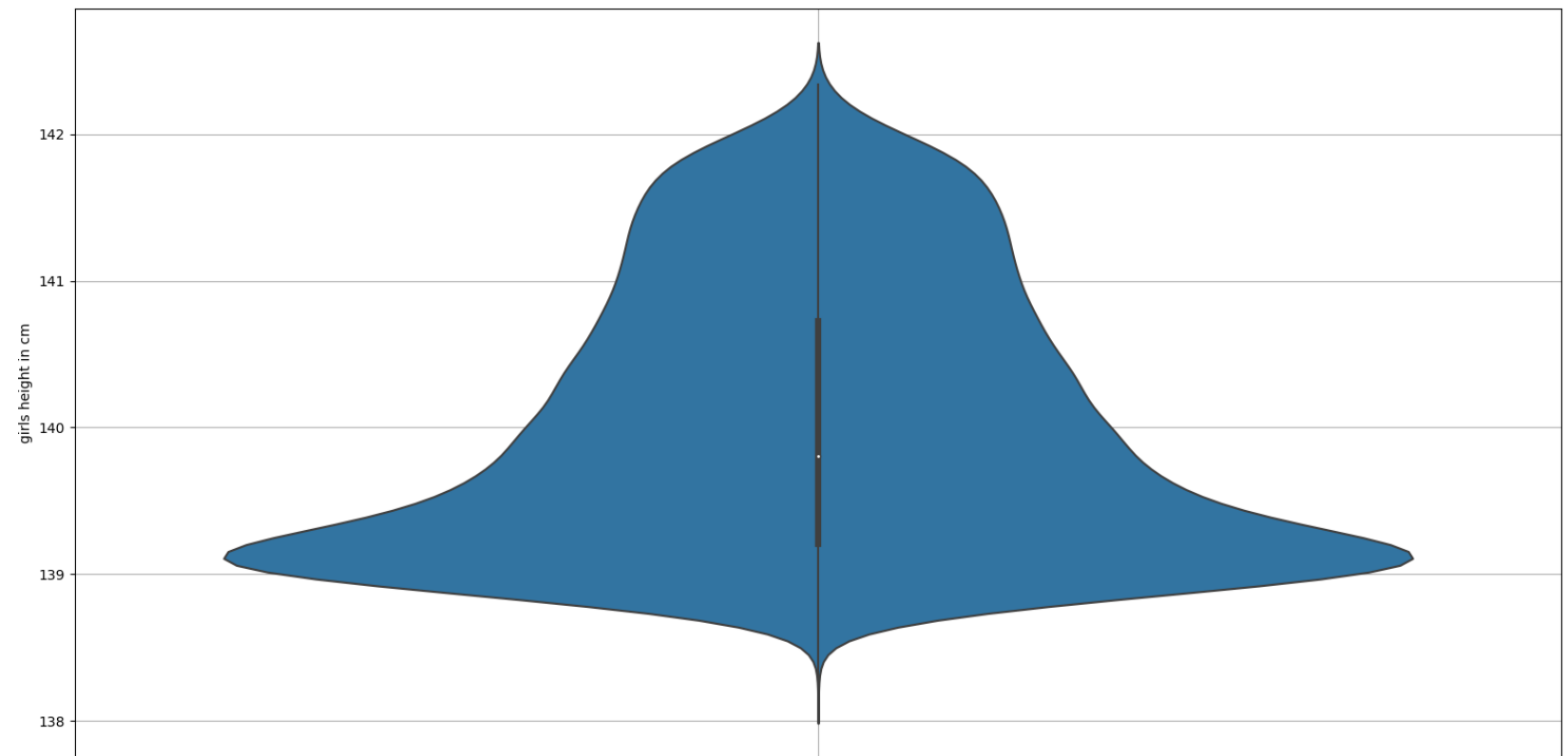
$$y = \frac{1}{2} \tanh(15(x - \mu)) (e^{3|x-\mu|} + 1) + 1$$



Gaussian (Normal) distribution

1-dimensional (single variable) distribution

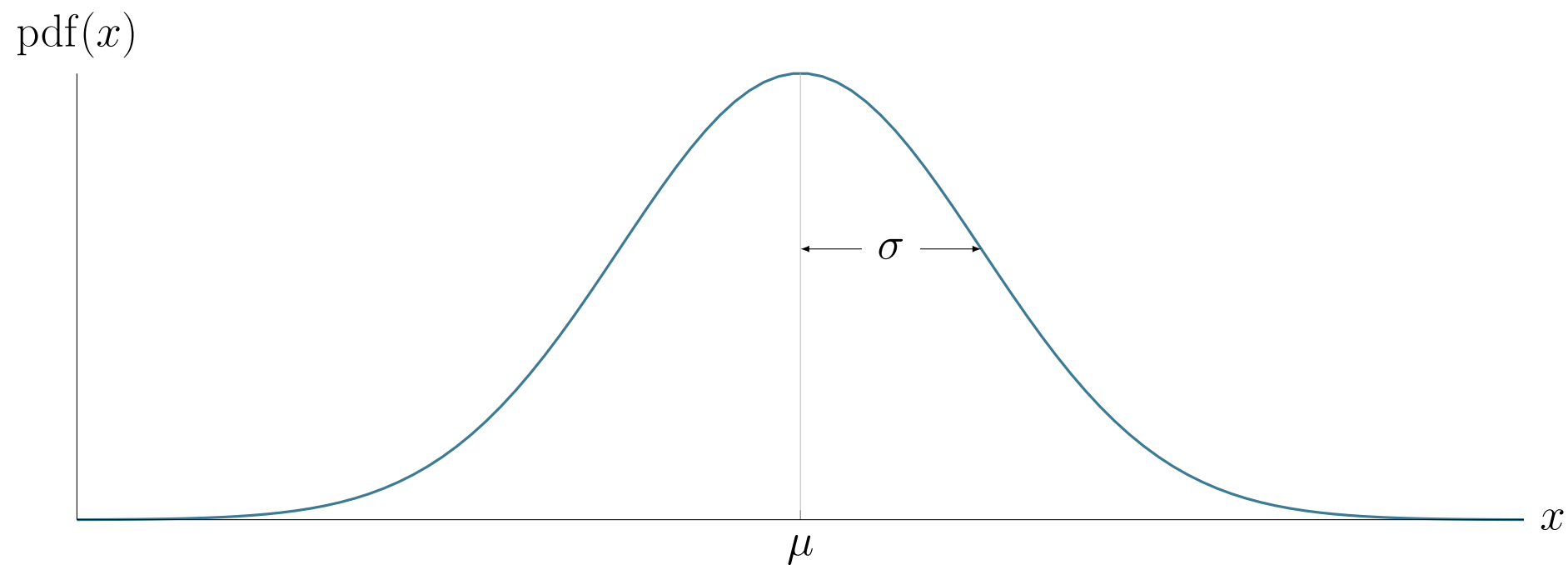
year	height
1985	138.8
1986	139.0
1987	139.0
1988	138.8
⋮	⋮
2010	138.8
2011	138.7
2012	139.2
2013	139.0
2014	139.0
2015	138.6
⋮	⋮



The data is the height of many girls over several years. We can look at the distribution of all these heights making a violin plot. In the middle of the violin you see a box plot. The body of the violin is a smoothed version of an histogram. The body of the violin is usually plotted resting on the horizontal axis. Here I want to emphasize the summary offered by the box plot, which shows in a line a location of the distribution (mean or median) and the scale of its spread (quartiles or standard deviation).

1-dimensional (single variable) Gaussian distribution

$$x \sim \mathcal{N}(\mu, \sigma^2) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \mu, \sigma \in \mathbb{R}$$
$$\mu \equiv E[x] \quad \sigma^2 \equiv E[(x - E[x])(x - E[x])]$$



The expectation operation is linear, if a and b are constants:

$$E[a + bx] = a + bE[x]$$

Numerically, the expectation is approximated by the usual arithmetic mean

$$E[x] \simeq \frac{1}{N} \sum_{i=1}^N x_i$$

where x_i are the realizations (samples) of the random variable x . For random vectors, the sum is applied to each component.

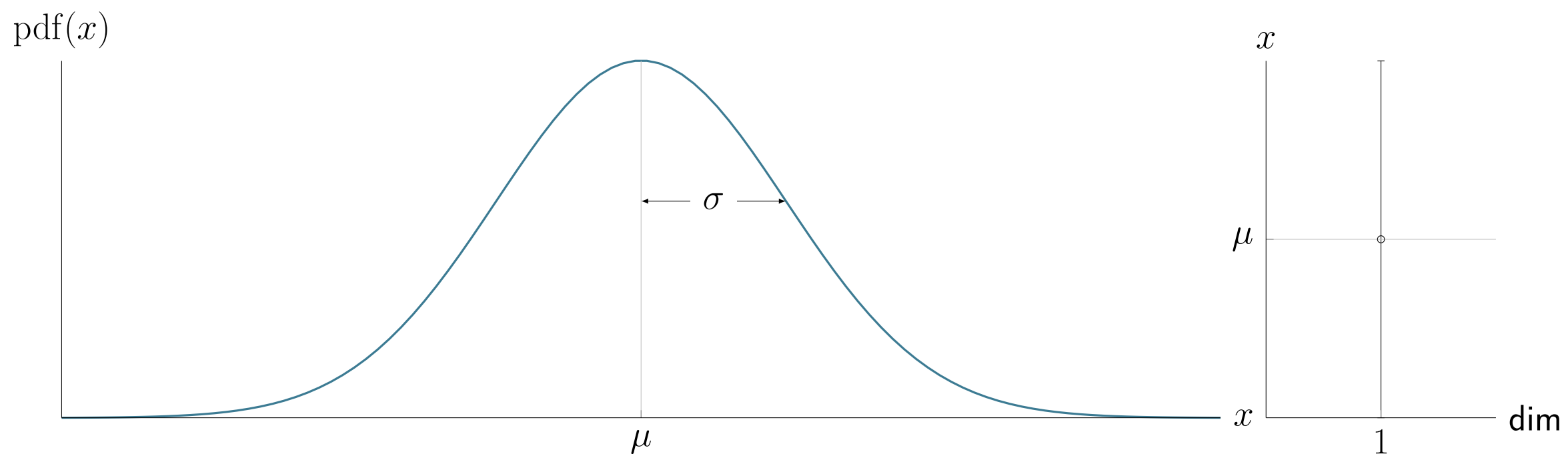
Check that the approximation is indeed linear!

The deviation from the mean (without compensation), i. e. variance, can be written as:

$$E[(x - E[x])(x - E[x])] = E[x^2 - 2xE[x] + E[x]^2] = E[x^2] - 2E[x]^2 + E[x]^2 = E[x^2] - E[x]^2$$

1-dimensional (single variable) Gaussian distribution

$$x \sim \mathcal{N}(\mu, \sigma^2) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \mu, \sigma \in \mathbb{R}$$
$$\mu \equiv E[x] \quad \sigma^2 \equiv E[(x - E[x])(x - E[x])]$$



The left panel shows the usual way of representing the body of the violin we saw before. The right panel shows the box plot view, in this case showing the mean value and the standard deviation.

The curve can indicate the frequency at which values would appear if we take large (infinite) samples from the distribution. Samples appearing more frequently in regions with higher values. It can also be used to represent our knowledge about a magnitude without the need to make a reference to sampling. The former is the frequentist interpretation, the latter is aligned with the Bayesian view.

1-dimensional (single variable) Gaussian distribution

With a linear change of variables, it maps to another gaussian:

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

$$y = ax + b$$

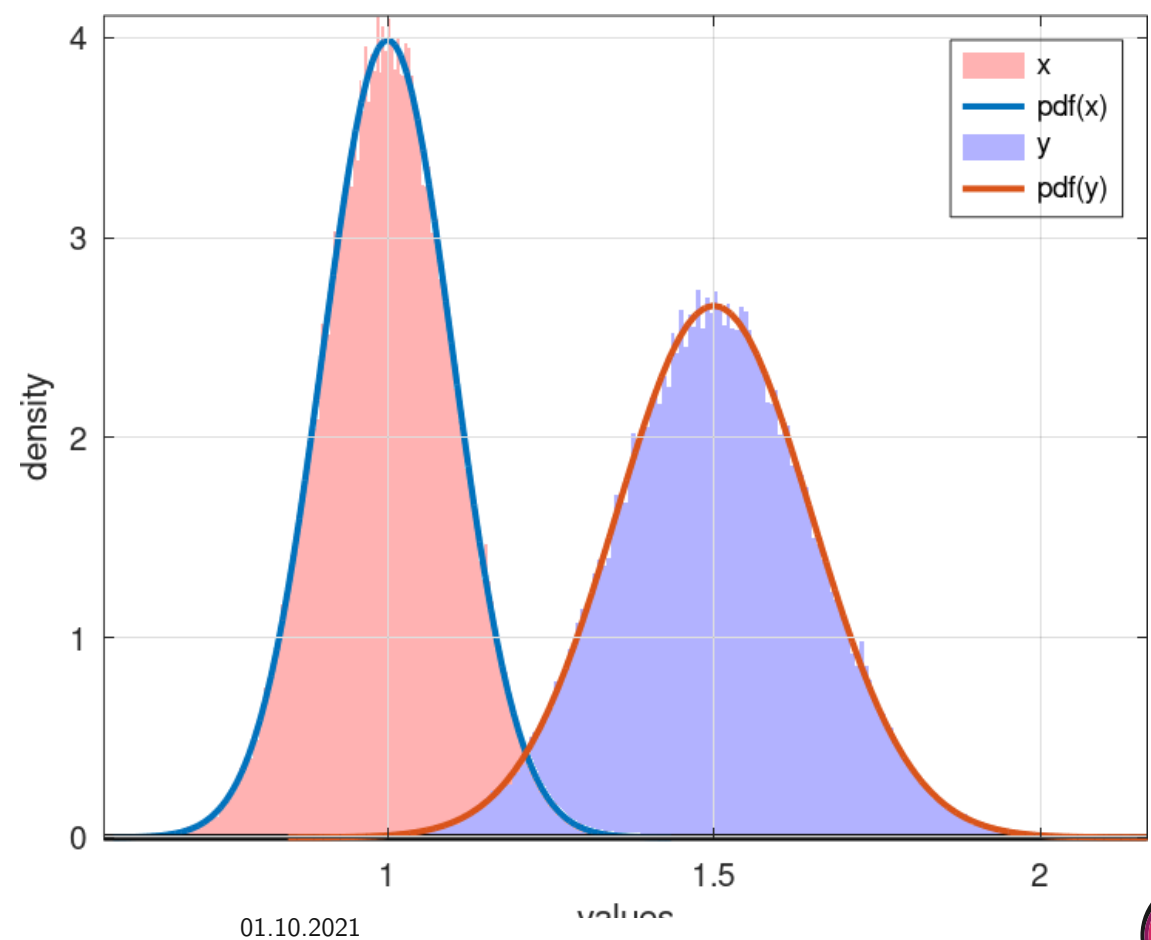
$$y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

$$\frac{-(x - \mu)^2}{2\sigma^2} = \frac{-(\frac{y-b}{a} - \mu)^2}{2\sigma^2} =$$

$$\frac{-(y - b - a\mu)^2}{2a^2\sigma^2}$$

$$\mu_y := a\mu + b, \quad \sigma_y^2 := a^2\sigma^2$$

μ transforms like the variable and σ^2 is multiplied by the squared slope

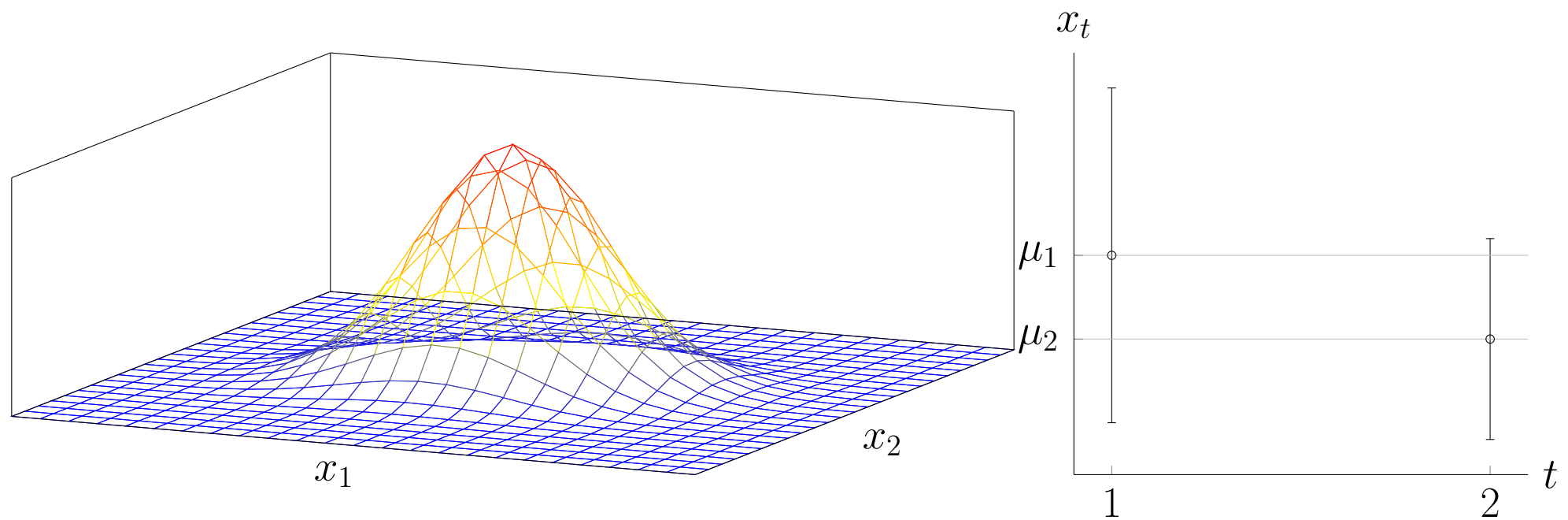


Also via the expectation operator:

$$\begin{aligned} \sigma_y^2 &= E[y^2] - E[y]^2 = E[(ax + b)^2] - E[ax + b]^2 \\ &= E[a^2x^2 + 2abx + b^2] - (a^2E[x]^2 + 2abE[x] + b^2) \\ &= a^2E[x^2] + 2abE[x] + b^2 - a^2E[x]^2 - 2abE[x] - b^2 \\ &= a^2(E[x^2] - E[x]^2) = a^2\sigma_x^2 \end{aligned}$$

2-dimensional (two variables) Gaussian distribution

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$
$$\boldsymbol{\mu} \in \mathbb{R}^{2 \times 1}, \boldsymbol{\Sigma} \in \mathbb{R}^{2 \times 2}$$



See `s_mvnormal.m`

We can still plot the joint distribution of the two variables on a piece of paper. However it can be quite difficult to read. The box plot or violin view is also useful. We lose the information about the interaction between the variables.

2-dimensional (two variables) Gaussian distribution

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{y} = A\mathbf{x} + \mathbf{b}$$

$$\mathbf{y} \sim \mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, A\boldsymbol{\Sigma}A^\top)$$

$$\text{var}(\mathbf{y}) := E[(\mathbf{y} - E[\mathbf{y}])(\mathbf{y} - E[\mathbf{y}])^\top]$$

$$\begin{aligned} \text{var}(A\mathbf{x} + \mathbf{b}) &:= E[(A\mathbf{x} + \mathbf{b} - E[A\mathbf{x} + \mathbf{b}])(A\mathbf{x} + \mathbf{b} - E[A\mathbf{x} + \mathbf{b}])^\top] \\ &= E[(A\mathbf{x} - E[A\mathbf{x}])(A\mathbf{x} - E[A\mathbf{x}])^\top] \\ &= A (E[\mathbf{x} - E[\mathbf{x}]](\mathbf{x} - E[\mathbf{x}])^\top) A^\top = A \text{var}(\mathbf{x})A^\top \end{aligned}$$

Note: this is valid for any number of dimensions

The definition of variance follows the same idea as before. The factors $\mathbf{y} - E[\mathbf{y}]$ are the deviation from the mean of each component. The product by the transpose pairs all components to each other. The diagonal terms are

$$\boldsymbol{\Sigma}_{i,i} = E[(y_i - E[y_i])(y_i - E[y_i])]$$

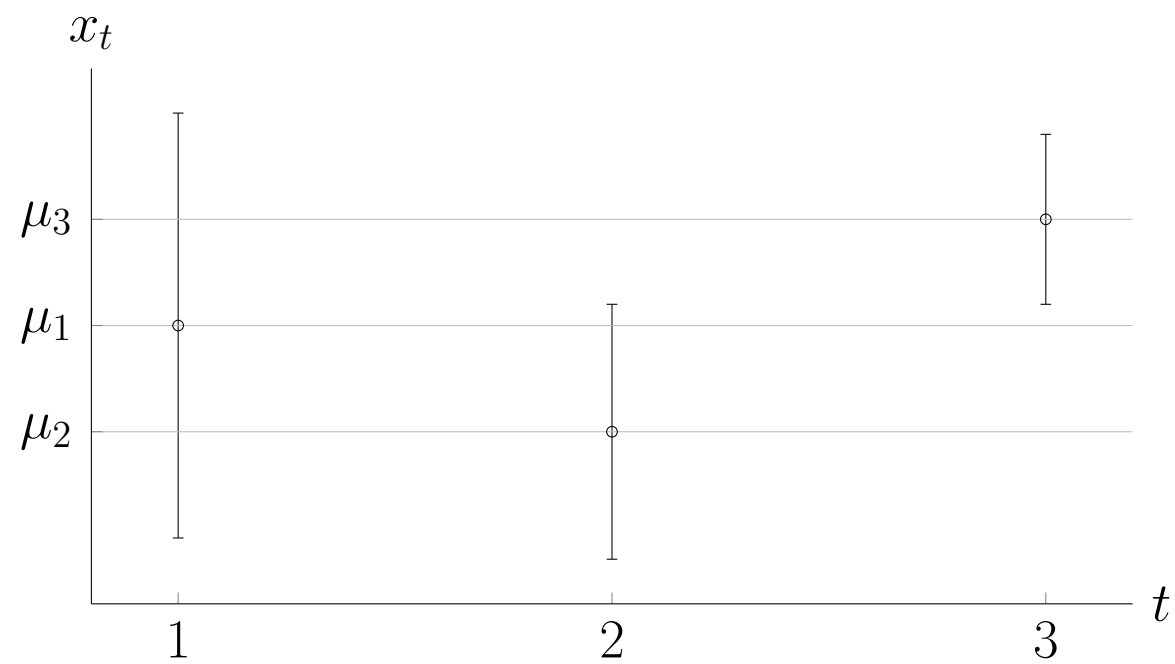
called covariance of y_i . The off-diagonal terms are

$$\boldsymbol{\Sigma}_{i,j} = E[(y_i - E[y_i])(y_j - E[y_j])]$$

called (cross-)covariances of y_i and y_j

3-dimensional Gaussian distribution

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp \left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right]$$
$$\boldsymbol{\mu} \in \mathbb{R}^{3 \times 1}, \boldsymbol{\Sigma} \in \mathbb{R}^{3 \times 3}$$



With three variables we cannot plot the joint distribution on a piece of paper. The box plot or violin view is still useful.

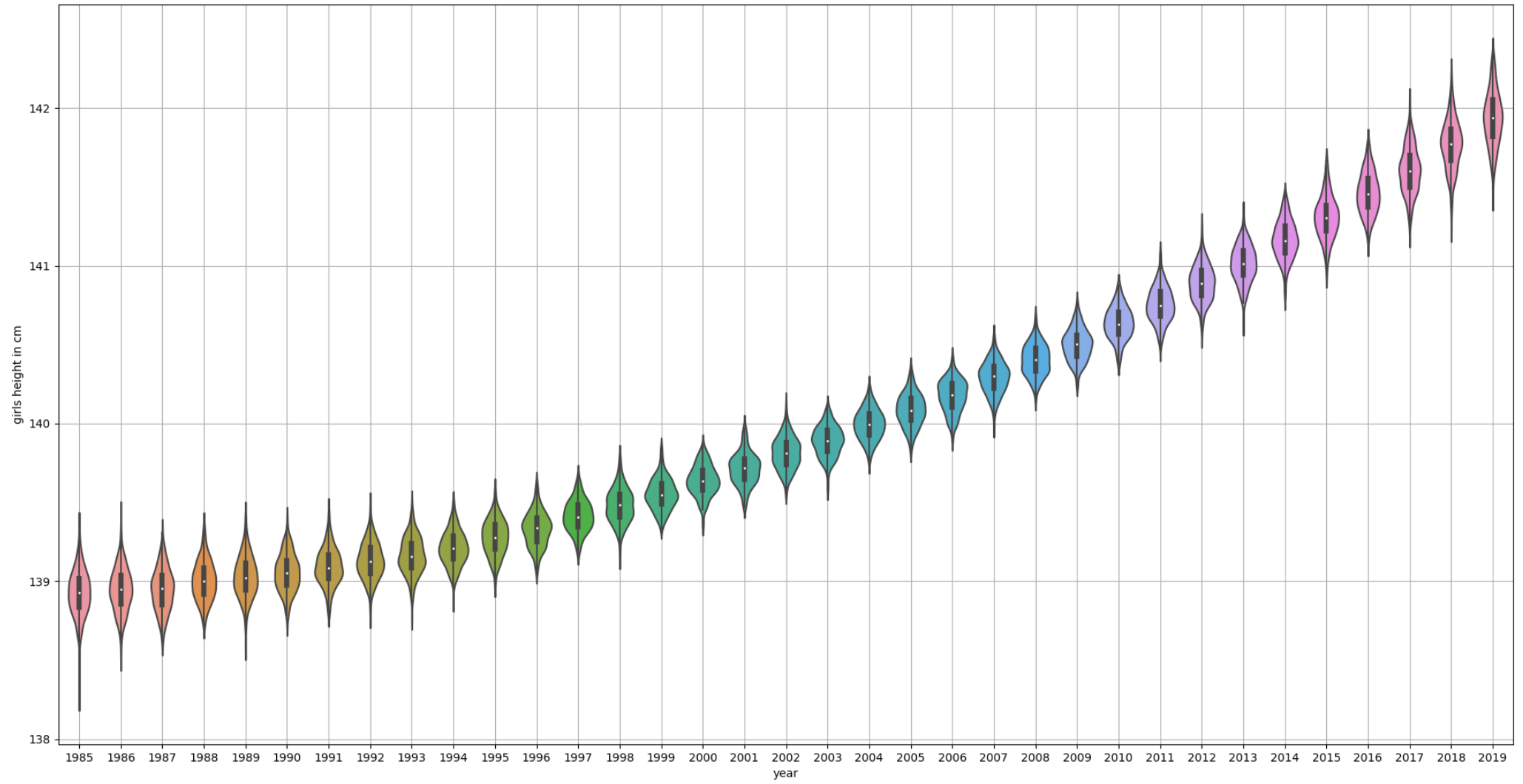
multi-dimensional distribution

year	height
1985	138.8
1986	139.0
1987	139.0
1988	138.8
⋮	⋮
2010	138.8
2011	138.7
2012	139.2
2013	139.0
2014	139.0
2015	138.6
⋮	⋮

Recall the height data, it was also indicated the year in which the measurement was done. Hence we can think of the heights of a given year as a random variable. We will have as many variables as years in our data set.

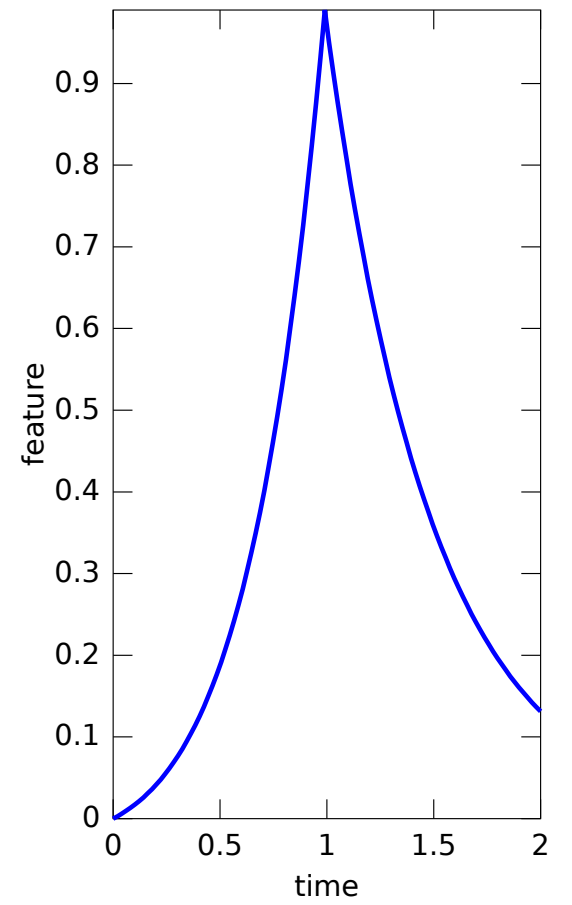
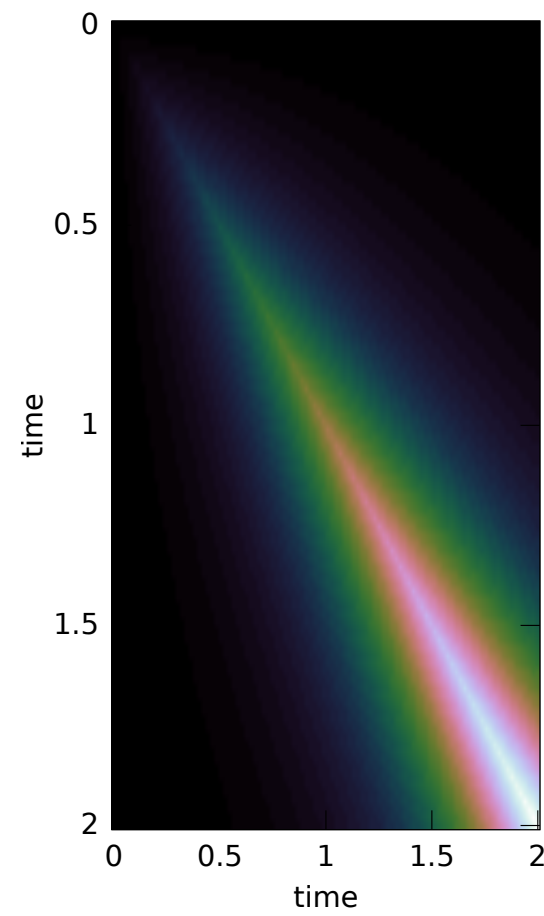
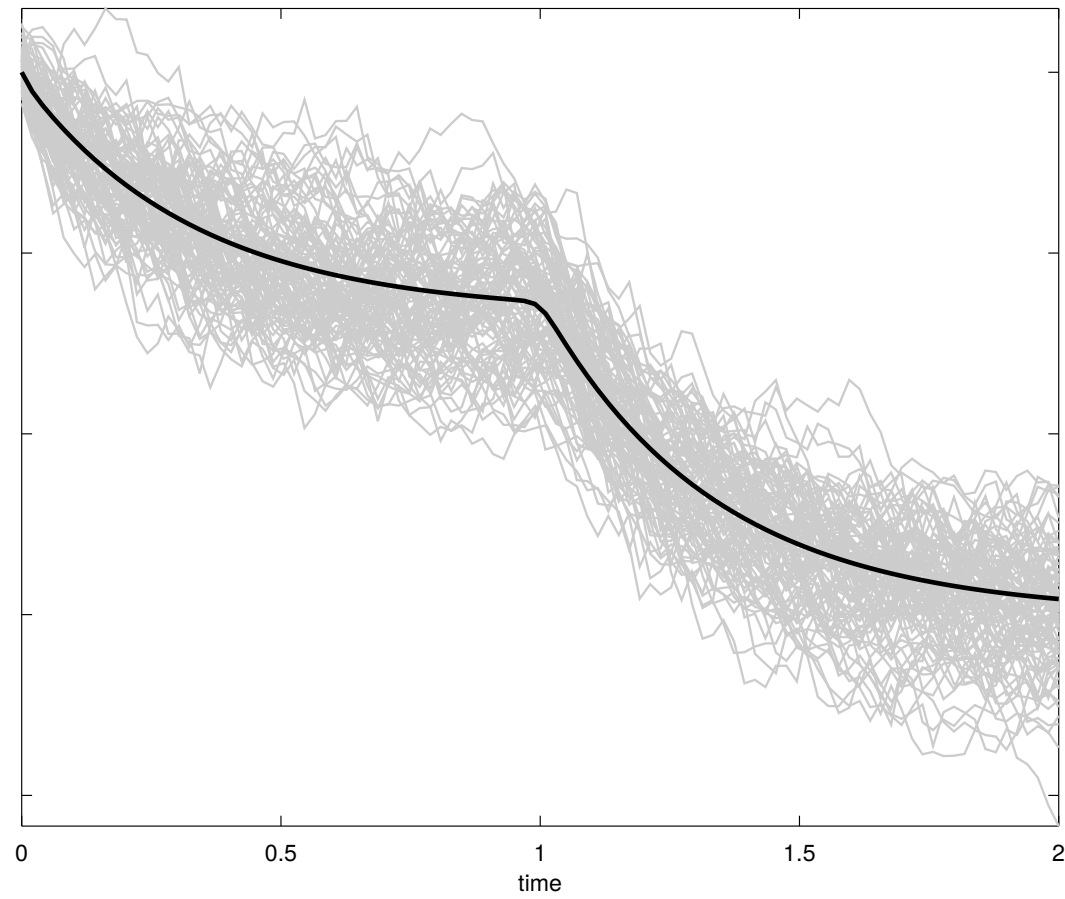
We cannot plot the joint distribution of all these variables.

multi-dimensional distribution

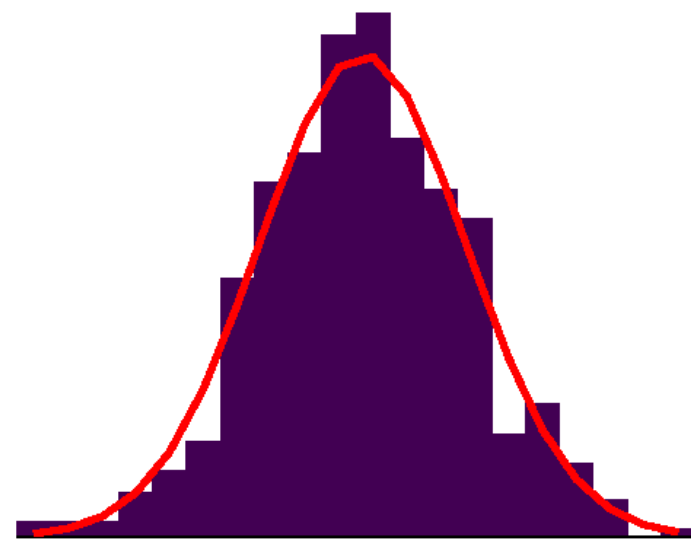
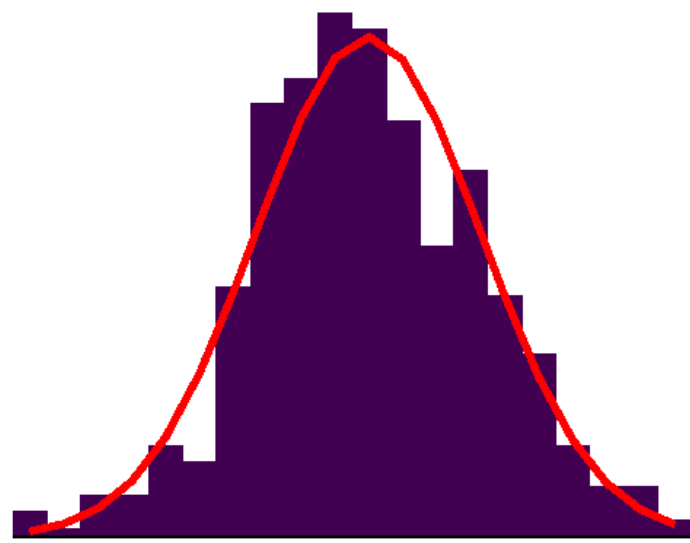
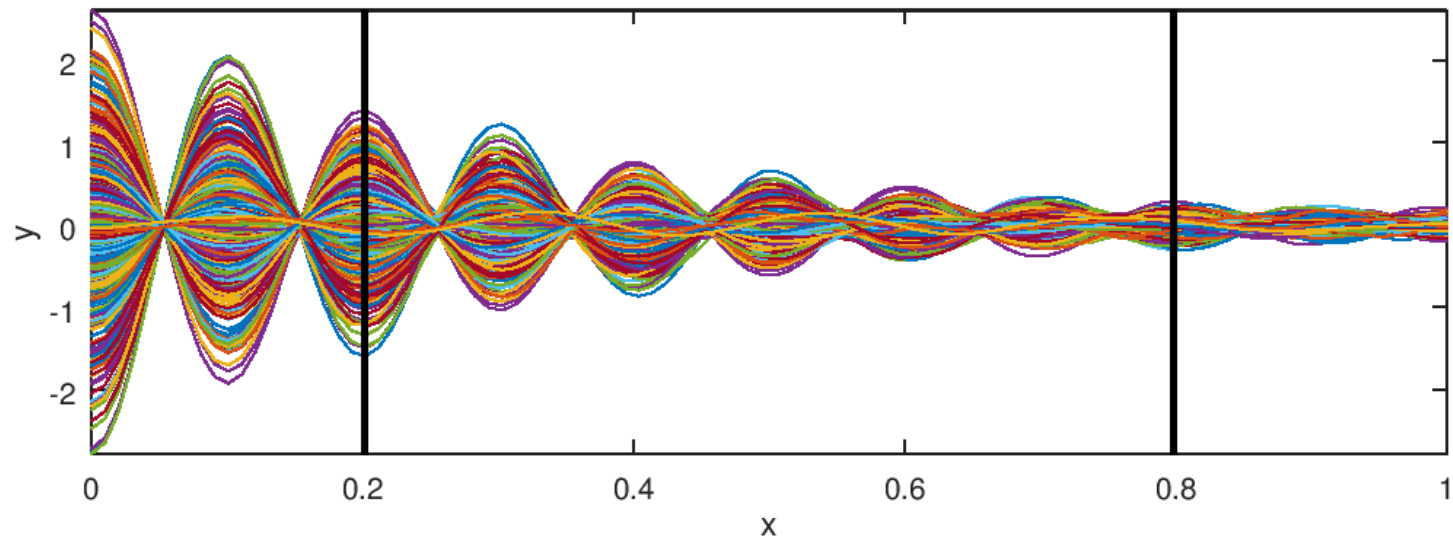


∞ -dimensional: GP

$$\begin{aligned}\mu &\rightarrow m(t) & \Sigma &\rightarrow k(t, t') \\ x(t) &\sim \mathcal{GP}(m(t), k(t, t'))\end{aligned}$$



∞ -dimensional: GP



Iterated maps revisited

Iterated map with process noise: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

What do we get if $\alpha = 1$?

What's the distribution of x if we know the value of y ?

Iterated map with process noise: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} + \epsilon_k \quad |\alpha| \leq 1, \quad \epsilon_k \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

That we draw ϵ_k from a normal distribution means that each sample is independent. If we look at the collection of ϵ_k as a random vector, this means that the covariance matrix is diagonal:

$$\Sigma_\epsilon(i, j) = \text{cov}(\epsilon_i, \epsilon_j) = \sigma_\epsilon^2 \delta_{i,j}$$

The samples at different steps are independent and therefore the (cross-)correlation between samples at different steps is zero.

Iterated map with process noise: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} + \epsilon_k \quad |\alpha| \leq 1, \quad \epsilon_k \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

0. x_0

Iterated map with process noise: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} + \epsilon_k \quad |\alpha| \leq 1, \quad \epsilon_k \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

0. x_0

1. $x_1 = \alpha x_0 + \epsilon_1$

Iterated map with process noise: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} + \epsilon_k \quad |\alpha| \leq 1, \quad \epsilon_k \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

0. x_0
1. $x_1 = \alpha x_0 + \epsilon_1$
2. $x_2 = \alpha x_1 + \epsilon_2 = \alpha^2 x_0 + \alpha \epsilon_1 + \epsilon_2$

Iterated map with process noise: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} + \epsilon_k \quad |\alpha| \leq 1, \quad \epsilon_k \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

0. x_0
1. $x_1 = \alpha x_0 + \epsilon_1$
2. $x_2 = \alpha x_1 + \epsilon_2 = \alpha^2 x_0 + \alpha \epsilon_1 + \epsilon_2$
3. $x_3 = \alpha x_2 + \epsilon_3 = \alpha^3 x_0 + \alpha^2 \epsilon_1 + \alpha \epsilon_2 + \epsilon_3$

Iterated map with process noise: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} + \epsilon_k \quad |\alpha| \leq 1, \quad \epsilon_k \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

0. x_0
1. $x_1 = \alpha x_0 + \epsilon_1$
2. $x_2 = \alpha x_1 + \epsilon_2 = \alpha^2 x_0 + \alpha \epsilon_1 + \epsilon_2$
3. $x_3 = \alpha x_2 + \epsilon_3 = \alpha^3 x_0 + \alpha^2 \epsilon_1 + \alpha \epsilon_2 + \epsilon_3$
4. ...

Iterated map with process noise: single variable

Map (mapping, transformation, etc.):

$$x = \alpha y + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Iterated map (use the result as the next value of x):

$$x_k = \alpha x_{k-1} + \epsilon_k \quad |\alpha| \leq 1, \quad \epsilon_k \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

0. x_0
1. $x_1 = \alpha x_0 + \epsilon_1$
2. $x_2 = \alpha x_1 + \epsilon_2 = \alpha^2 x_0 + \alpha \epsilon_1 + \epsilon_2$
3. $x_3 = \alpha x_2 + \epsilon_3 = \alpha^3 x_0 + \alpha^2 \epsilon_1 + \alpha \epsilon_2 + \epsilon_3$
4. ...
5. $x_k = \alpha^k x_0 + \sum_{i=1}^k \alpha^{k-i} \epsilon_i$

What happens to the "old" noise terms if $\alpha < 1$? And if $\alpha > 1$?

Take $\alpha = 10^{-1}$, $x_0 = 1$, and $\sigma_\epsilon = 10^{-2}$. We ask for what k is "likely" that $|\epsilon_k| > \alpha^k x_0$.

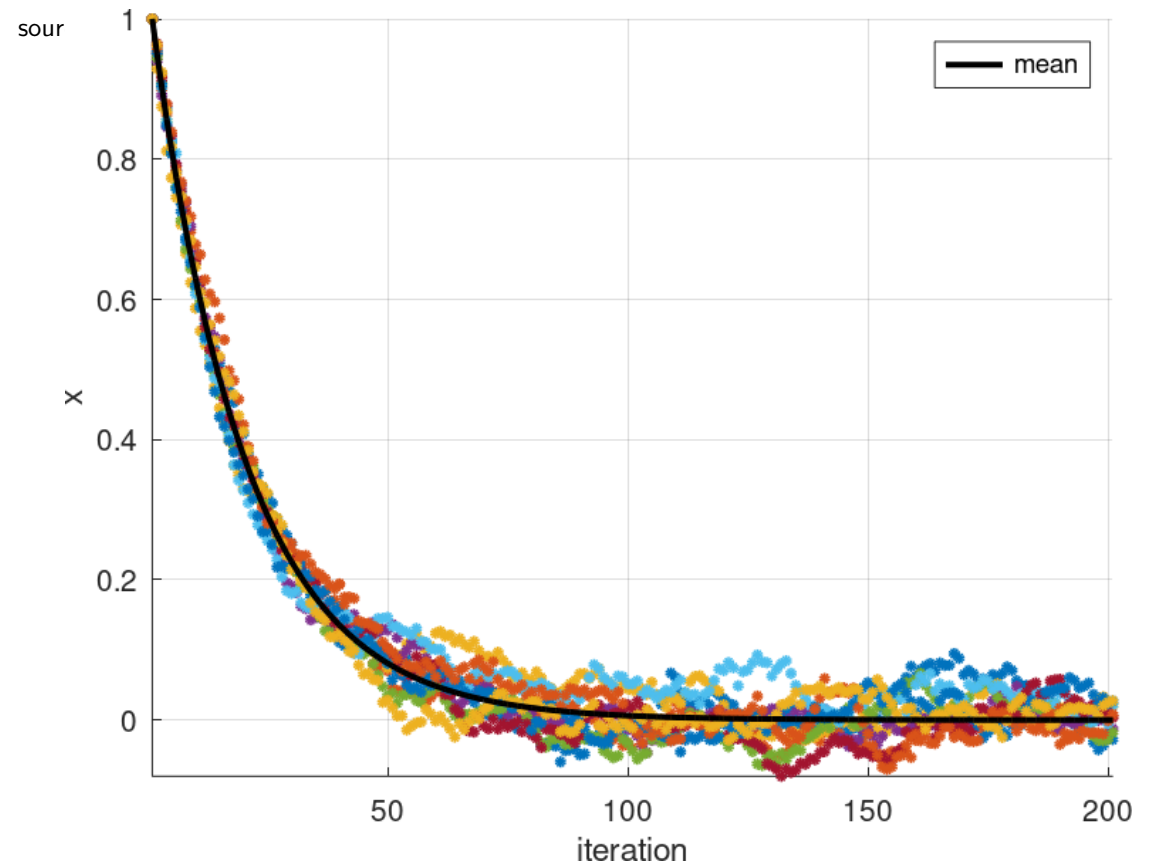
Taking "likely" as $P(|\epsilon_k| > \alpha^k x_0) > 0.9$, that is $P(|\epsilon_k| > 10^{-k})$, and using the cumulative distribution of the normal we get $k = 3$.

After 3 steps the signal is just noise. The step correlation of the noise, however, still carries information about the map.

Iterated map with process noise: single variable

$$x_k = (1 - 0.05)x_{k-1} + \epsilon_k \quad \epsilon_k \sim \mathcal{N}(0, 0.01^2)$$

- Different trajectories for the same initial value
- What's the formula for the mean trajectory (black line)?
- What's the formula for the variance of the trajectories?



The trajectories do not repeat, even for the same initial value. Each simulation is "unique". Running the map several times from the same initial value shows a distribution of trajectories (also called paths). We can compute the mean trajectory by averaging the ensemble of trajectories at each step. We can also compute the variance of the trajectories at each step.

Can we find a map for the mean trajectory?

Can we find a map for the variance?

Iterated map with process noise: single variable

$$x_k = \alpha^k x_0 + \sum_{i=1}^k \alpha^{k-i} \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

At each time step it is a linear mapping, the result is Gaussian.

$$\begin{aligned} E[x_k] &= E\left[\alpha^k x_0 + \sum_{i=1}^k \alpha^{k-i} \epsilon_i\right] = \alpha^k E[x_0] + \sum_{i=1}^k \alpha^{k-i} \underbrace{E[\epsilon_i]}_{=0} \\ &= \alpha^k E[x_0] \end{aligned}$$

The mean trajectory is the one of the map without noise

We directly compute the mean using the expectation operator.

Iterated map with process noise: single variable

$$x_k = \alpha^k x_0 + \sum_{i=1}^k \alpha^{k-i} \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

At each time step it is a linear mapping, the result is Gaussian.

$$r_k := x_k - E[x_k] = \alpha^k (x_0 - E[x_0]) + \overbrace{\sum_{i=1}^k \alpha^{k-i} \epsilon_i}^{\phi_k}$$

$$\begin{aligned} E[r_k^2] &= E[\alpha^{2k} (x_0 - E[x_0])^2 + 2\alpha^k (x_0 - E[x_0]) \phi_k + \phi_k^2] \\ &= \alpha^{2k} \underbrace{E[(x_0 - E[x_0])^2]}_{\sigma_0^2} + 2\alpha^k \underbrace{E[(x_0 - E[x_0]) \phi_k]}_{=0} + E[\phi_k^2] \\ &= \alpha^{2k} \sigma_0^2 + E\left[\left(\sum_{i=1}^k \alpha^{k-i} \epsilon_i\right)^2\right] \end{aligned}$$

In the last term all crossed terms are zero $E[\epsilon_i \epsilon_j] = \delta_{ij}$ (uncorrelated noise).

To compute the variance we first define the residual r_k at each step. Then apply the expectation operator as usual. The initial condition could have $\sigma_0 = 0$, I just kept it because it doesn't hurt.

Iterated map with process noise: single variable

$$x_k = \alpha^k x_0 + \sum_{i=1}^k \alpha^{k-i} \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

At each time step it is a linear mapping, the result is Gaussian.

$$r_k := x_k - E[x_k] = \alpha^k (x_0 - E[x_0]) + \overbrace{\sum_{i=1}^k \alpha^{k-i} \epsilon_i}^{\phi_k}$$

$$\begin{aligned} E[r_k^2] &= \alpha^{2k} \sigma_0^2 + E \left[\left(\sum_{i=1}^k \alpha^{k-i} \epsilon_i \right)^2 \right] = \alpha^{2k} \sigma_0^2 + \sum_{i=1}^k \alpha^{2(k-i)} E[\epsilon_i^2] \\ &= (\alpha^2)^k \sigma_0^2 + \sum_{i=1}^k (\alpha^2)^{k-i} \sigma_\epsilon^2 \end{aligned}$$

Same as the state iteration but with α^2 and noise σ_ϵ^2 .

$$\sum_{i=1}^k (\alpha^2)^{k-i} \stackrel{s=k-i}{=} \sum_{s=0}^{k-1} (\alpha^2)^s \xrightarrow{k \rightarrow \infty} \frac{1}{1 - \alpha^2}$$

Iterated map: two variables

Map (mapping, transformation, etc.)

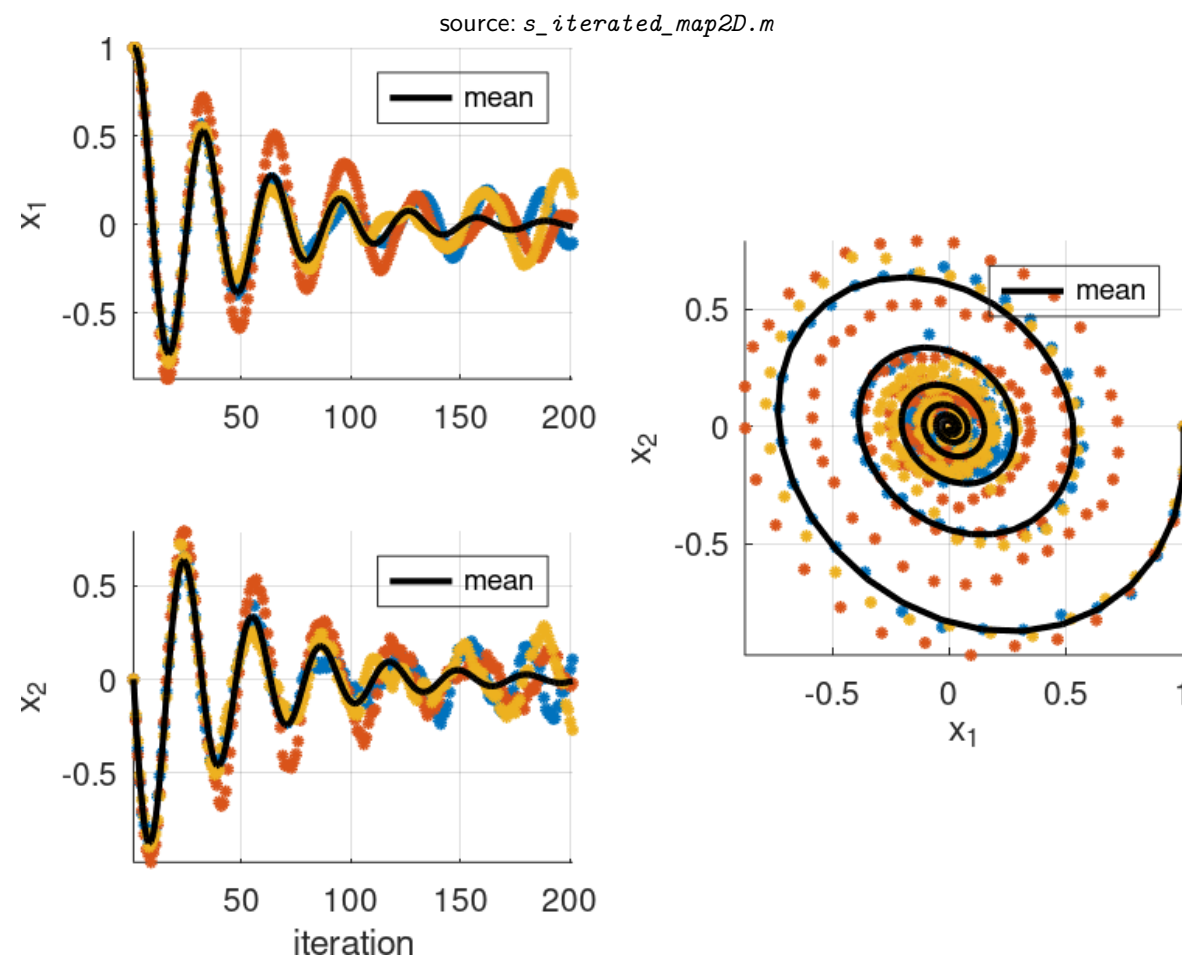
$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$$

Iterated map:

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \boldsymbol{\epsilon}_k \quad \boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$$

Iterated map: two variables

$$\mathbf{x}_k = \left(\mathbf{I} + 0.2 \begin{bmatrix} 0 & 1 \\ -1 & -0.4 \end{bmatrix} \right) \mathbf{x}_{k-1} + \boldsymbol{\epsilon}_k \quad \boldsymbol{\epsilon}_k \sim \mathcal{N} \left(0, \begin{bmatrix} \approx 0 & 0 \\ 0 & 0.03^2 \end{bmatrix} \right)$$



Iterated map: two variables

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \boldsymbol{\epsilon}_k \quad \boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$$

$$\boldsymbol{\mu}_k := E[\mathbf{x}_k] = \mathbf{A}E[\mathbf{x}_{k-1}] = \mathbf{A}\boldsymbol{\mu}_{k-1}$$

$$\begin{aligned} \mathbf{r}_k &:= \mathbf{x}_k - \boldsymbol{\mu}_k = \mathbf{A}(\mathbf{x}_{k-1} - \boldsymbol{\mu}_{k-1}) + \boldsymbol{\epsilon}_k = \mathbf{A}\mathbf{r}_{k-1} + \boldsymbol{\epsilon}_k \\ \boldsymbol{\Sigma}_k &:= E[\mathbf{r}_k\mathbf{r}_k^\top] = E[(\mathbf{A}\mathbf{r}_{k-1} + \boldsymbol{\epsilon}_k)(\mathbf{A}\mathbf{r}_{k-1} + \boldsymbol{\epsilon}_k)^\top] = \\ &= E[\mathbf{A}\mathbf{r}_{k-1}\mathbf{r}_{k-1}^\top\mathbf{A}^\top + \mathbf{A}\mathbf{r}_{k-1}\boldsymbol{\epsilon}_k^\top + \boldsymbol{\epsilon}_k\mathbf{r}_{k-1}^\top\mathbf{A}^\top + \boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k^\top] = \\ &= \mathbf{A}E[\mathbf{r}_{k-1}\mathbf{r}_{k-1}^\top]\mathbf{A}^\top + \mathbf{A}E[\mathbf{r}_{k-1}\boldsymbol{\epsilon}_k^\top] + E[\boldsymbol{\epsilon}_k\mathbf{r}_{k-1}^\top]\mathbf{A}^\top + E[\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k^\top] = \\ &= \mathbf{A}\boldsymbol{\Sigma}_{k-1}\mathbf{A}^\top + \boldsymbol{\Sigma}_\epsilon \end{aligned}$$

$$E[\mathbf{r}_{k-1}\boldsymbol{\epsilon}_k^\top] = E[(\mathbf{x}_k - \boldsymbol{\mu}_k)\boldsymbol{\epsilon}_k^\top] = E[\mathbf{x}_k\boldsymbol{\epsilon}_k^\top] - \boldsymbol{\mu}_k E[\boldsymbol{\epsilon}_k^\top] = 0$$

As before we directly compute the mean and variance by application of the expectation. To simplify the algebra in the variance we use the residuals \mathbf{r}_k . The cross-covariance between residuals and noise at different steps $E[\mathbf{r}_{k-1}\boldsymbol{\epsilon}_k^\top]$ are zero because the noise has zero mean and is not correlated with the state.

Iterated map: important results

The iteration represent a dynamical model of a process:

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \boldsymbol{\epsilon}_k \quad \boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$$

Mean trajectory and variance:

$$\boldsymbol{\mu}_k = \mathbf{A}\boldsymbol{\mu}_{k-1} \quad \boldsymbol{\Sigma}_k = \mathbf{A}\boldsymbol{\Sigma}_{k-1}\mathbf{A}^\top + \boldsymbol{\Sigma}_\epsilon$$

In the literature usually are written:

$$\mathbf{m}_k = \mathbf{A}\mathbf{m}_{k-1} \quad \mathbf{P}_k = \mathbf{A}\mathbf{P}_{k-1}\mathbf{A}^\top + \mathbf{Q}_{k-1}$$

and they are called the **prediction step**.

The notation used in many books on Bayesian filtering and smoothing is shown. Note that the covariance of the noise can also be step dependent, as long as it does not depend on the state of the system.

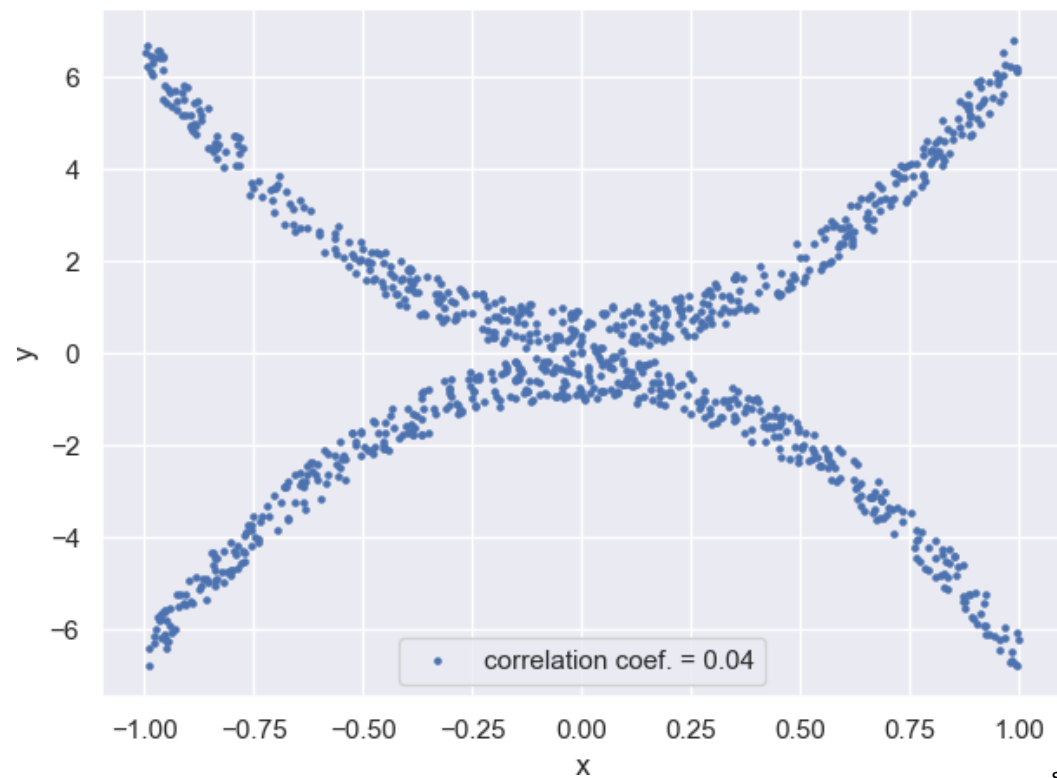
So far we refer to $\boldsymbol{\epsilon}$ as "the noise", in the literature it is called *process* noise, because it feeds into the evolution of the process. It models uncertainties and/or perturbations in the dynamics.

Stochastic modelling

Correlation and Causation

$$x \sim \mathcal{U}(-1, 1)$$

$$y \leftarrow a \frac{\epsilon}{|\epsilon|} x^2 + \epsilon \quad \epsilon \sim \mathcal{U}(-1, 1)$$

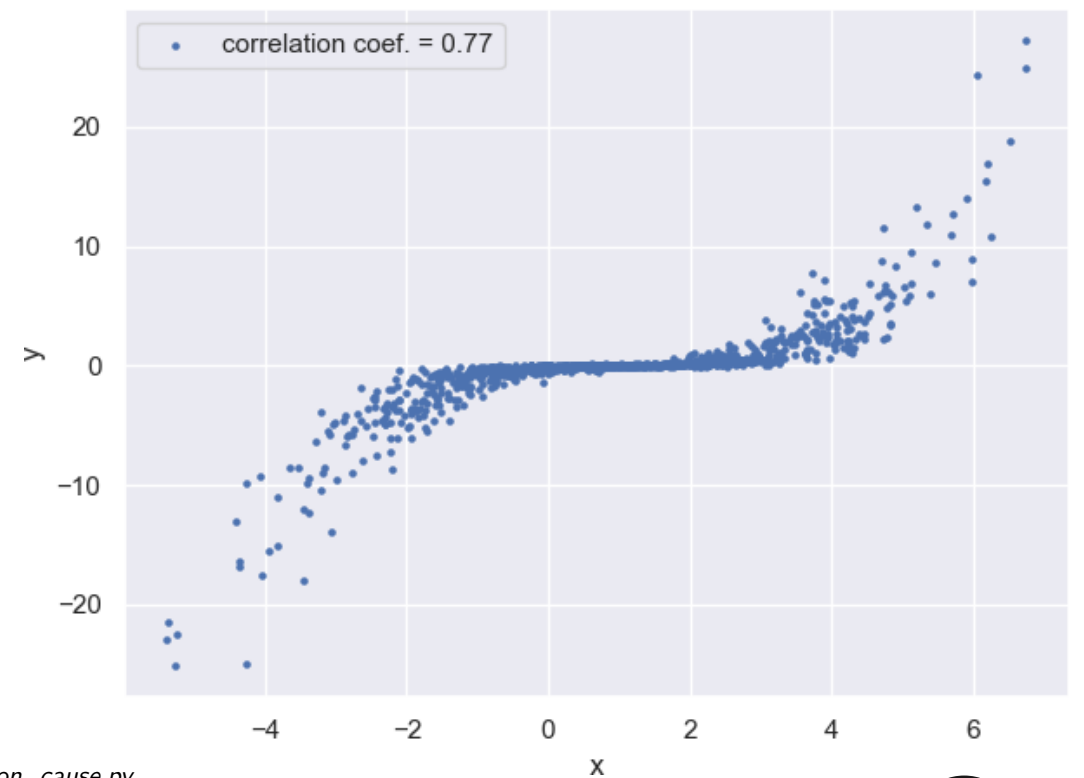


source: s_common_cause.py

$$c \sim \mathcal{N}(0, 1)$$

$$x \sim \mathcal{N}(2c - 1, 0.2)$$

$$y \sim \mathcal{N}(c^3, 0.1)$$



X vs. Y plots have different interpretations. On the left panel we see a "functional" or "casual" plot, in which changing the values of x will affect the values of y . This is because the variables are connected by a function, y being a function of x . Note that the correlation, however, is negligible.

On the right panel we see a plot that would seem to also be "causal", saying that values of x affect y , however if we inspect the process generating the data we see that x does not affect y . The variables are correlated and are statistically dependent, but they are not causally dependent.

Correlation and Causation

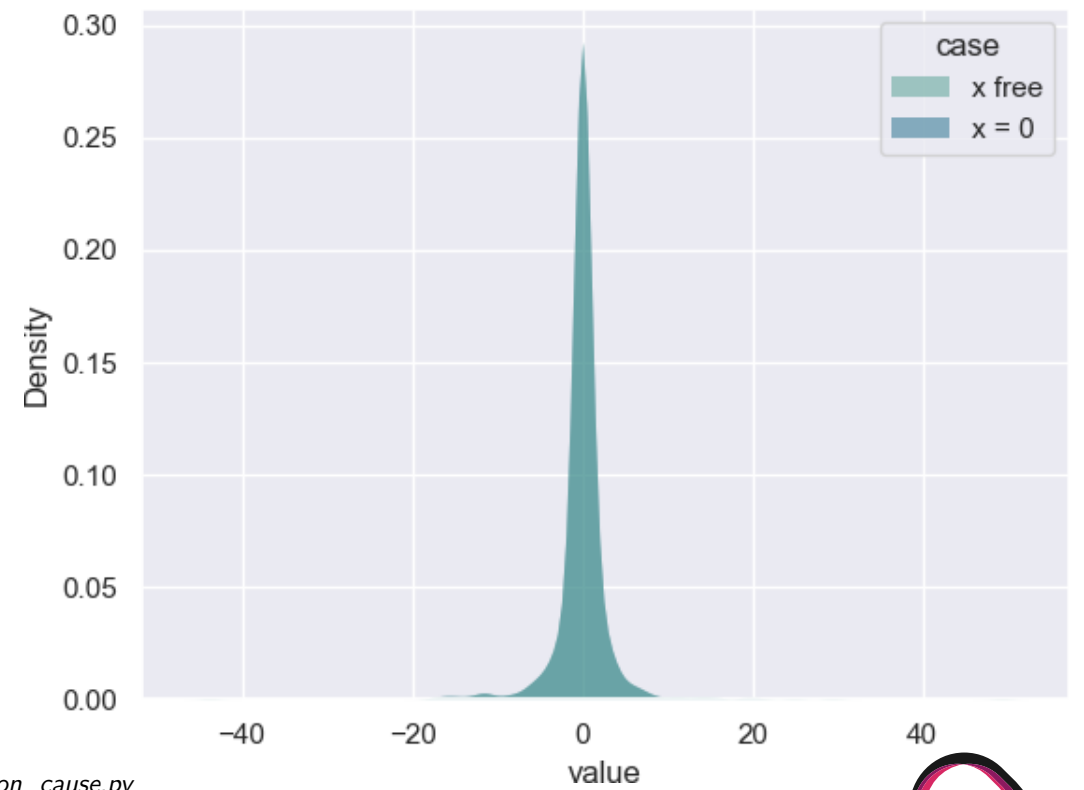
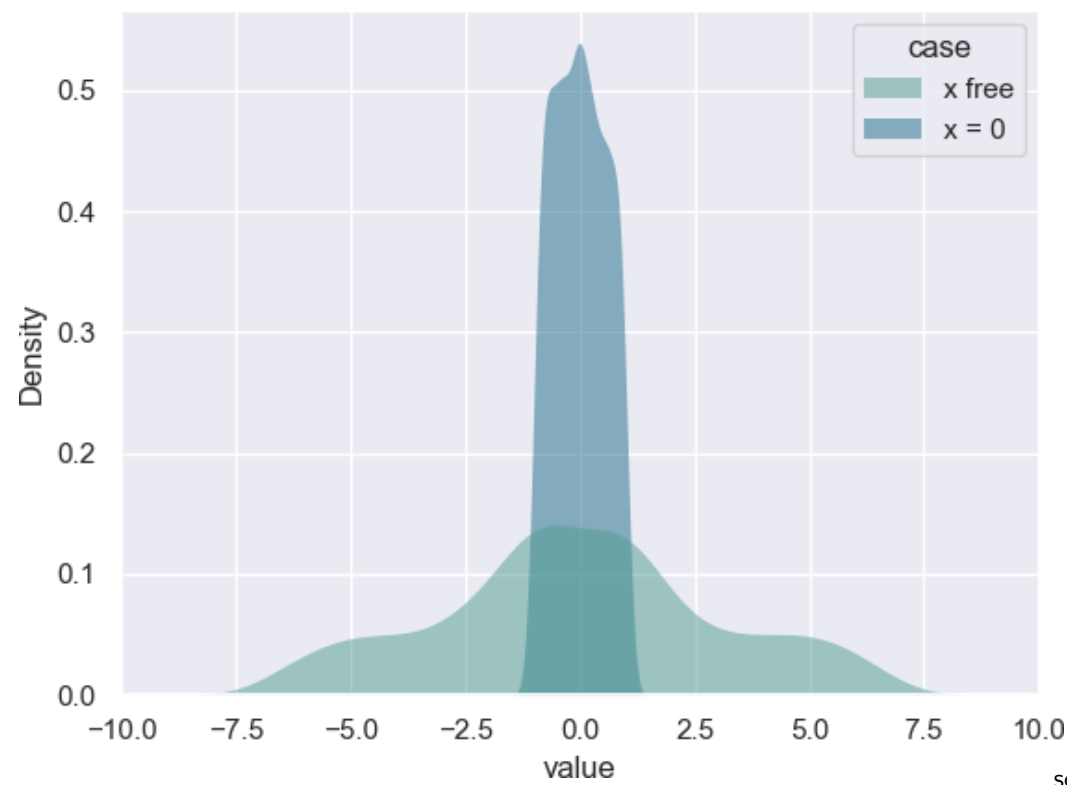
$$x \leftarrow 0 \sim \mathcal{U}(-1, 1)$$

$$y \leftarrow a \frac{\epsilon}{|\epsilon|} x^2 + \epsilon \quad \epsilon \sim \mathcal{U}(-1, 1)$$

$$c \sim \mathcal{N}(0, 1)$$

$$x \leftarrow 0 \sim \mathcal{N}(2c - 1, 0.2)$$

$$y \sim \mathcal{N}(c^3, 0.1)$$



38 Introduction to recursive probabilistic learning

source: [s common_cause.py](#)
01.10.2021



The situation becomes evident when we intervene the values of x . That is, we make an experiment and set x to a value (in this case $x \leftarrow 0$).

In the right panel, it is evident that the distribution of y after the intervention is the same as before the intervention, x wasn't affecting y , although they were correlated. In the left panel however, the intervention on x radically changed the distribution of y .

These types of distributions are called interventional distributions. We will see that they are conceptually different from conditional distributions.

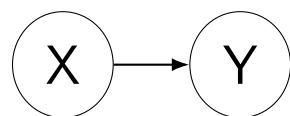
For more information refer to

- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press.
- Pearl, Judea, and Dana Mackenzie. 2019. The Book of Why. Harlow, England: Penguin Books.

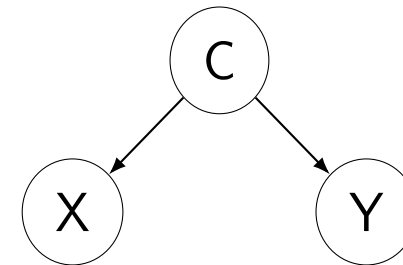
Graphical models

Graphic models are a tool to visualize dependencies between (random) variables. The nodes of a graph represent the variables and arrows indicate direction of influence.

$$x \sim \mathcal{U}(-1, 1)$$
$$y \leftarrow a \frac{\epsilon}{|\epsilon|} x^2 + \epsilon \quad \epsilon \sim \mathcal{U}(-1, 1)$$



$$c \sim \mathcal{N}(0, 1)$$
$$x \sim \mathcal{N}(2c - 1, 0.2)$$
$$y \sim \mathcal{N}(c^3, 0.1)$$



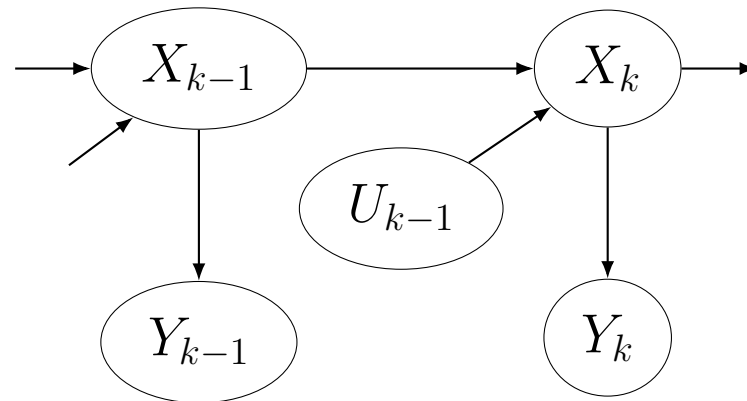
There are also undirected graphic models. See https://en.wikipedia.org/wiki/Graphical_model for an overview.

Iterated map

For an iterated map

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_{k-1} + \boldsymbol{\epsilon}_x$$

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \boldsymbol{\epsilon}_y$$



Wrap-up

Iterated map: important results

The iteration represent a dynamical model of a process:

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \boldsymbol{\epsilon}_k \quad \boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$$

Mean trajectory and variance:

$$\boldsymbol{\mu}_k = \mathbf{A}\boldsymbol{\mu}_{k-1} \quad \boldsymbol{\Sigma}_k = \mathbf{A}\boldsymbol{\Sigma}_{k-1}\mathbf{A}^\top + \boldsymbol{\Sigma}_\epsilon$$

In the literature usually are written:

$$\mathbf{m}_k = \mathbf{A}\mathbf{m}_{k-1} \quad \mathbf{P}_k = \mathbf{A}\mathbf{P}_{k-1}\mathbf{A}^\top + \mathbf{Q}_{k-1}$$

and they are called the **prediction step**.

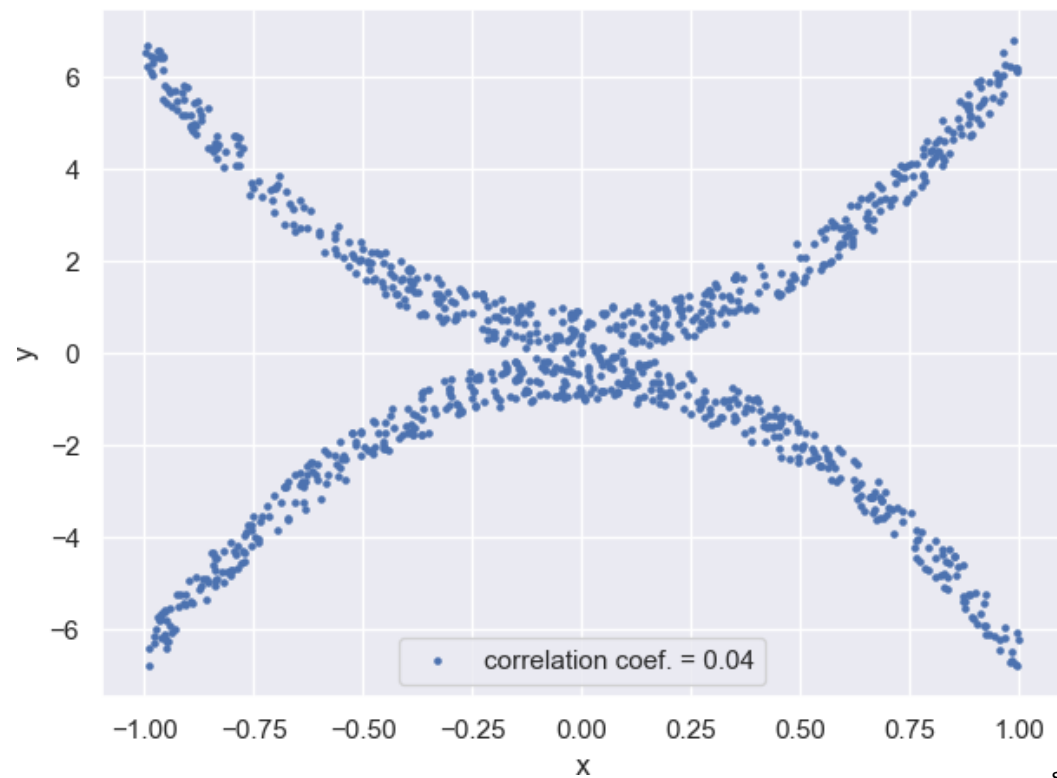
The notation used in many books on Bayesian filtering and smoothing is shown. Note that the covariance of the noise can also be step dependent, as long as it does not depend on the state of the system.

So far we refer to $\boldsymbol{\epsilon}$ as "the noise", in the literature it is called *process* noise, because it feeds into the evolution of the process. It models uncertainties and/or perturbations in the dynamics.

Correlation and Causation

$$x \sim \mathcal{U}(-1, 1)$$

$$y \leftarrow a \frac{\epsilon}{|\epsilon|} x^2 + \epsilon \quad \epsilon \sim \mathcal{U}(-1, 1)$$

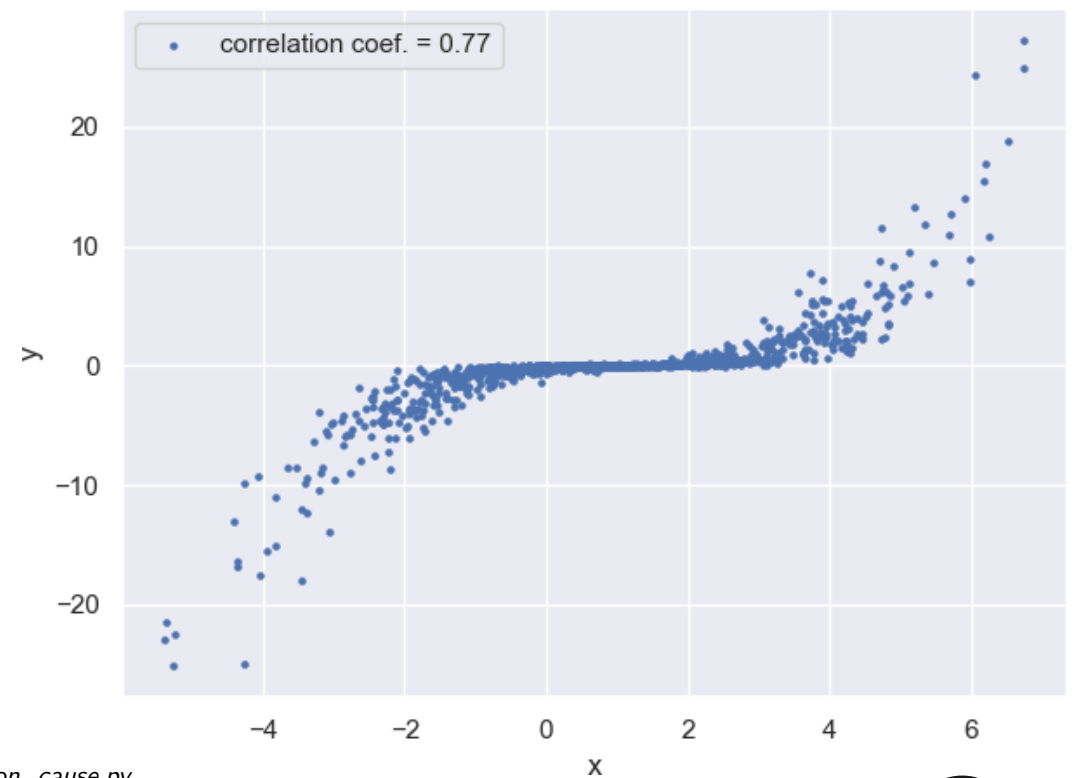


source: s_common_cause.py

$$c \sim \mathcal{N}(0, 1)$$

$$x \sim \mathcal{N}(2c - 1, 0.2)$$

$$y \sim \mathcal{N}(c^3, 0.1)$$



X vs. Y plots have different interpretations. On the left panel we see a "functional" or "casual" plot, in which changing the values of x will affect the values of y . This is because the variables are connected by a function, y being a function of x . Note that the correlation, however, is negligible.

On the right panel we see a plot that would seem to also be "causal", saying that values of x affect y , however if we inspect the process generating the data we see that x does not affect y . The variables are correlated and are statistically dependent, but they are not causally dependent.