IAEA
International Atomic Energy Agency
Atoms for Peace and Development

Probabilistic methods for nuclear data

(An introduction to Gaussian process regression)

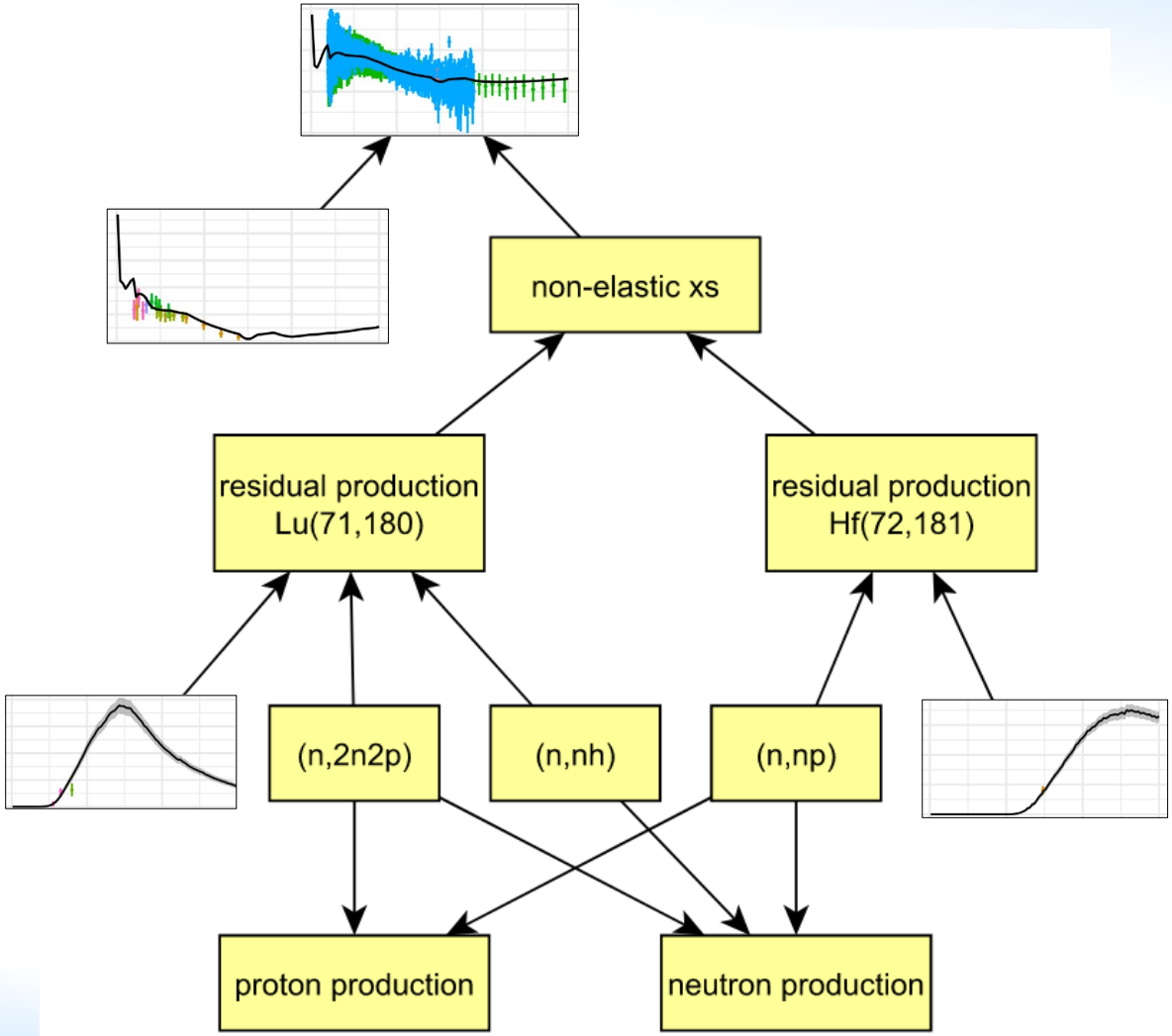
Georg Schnabel

Nuclear Data Section

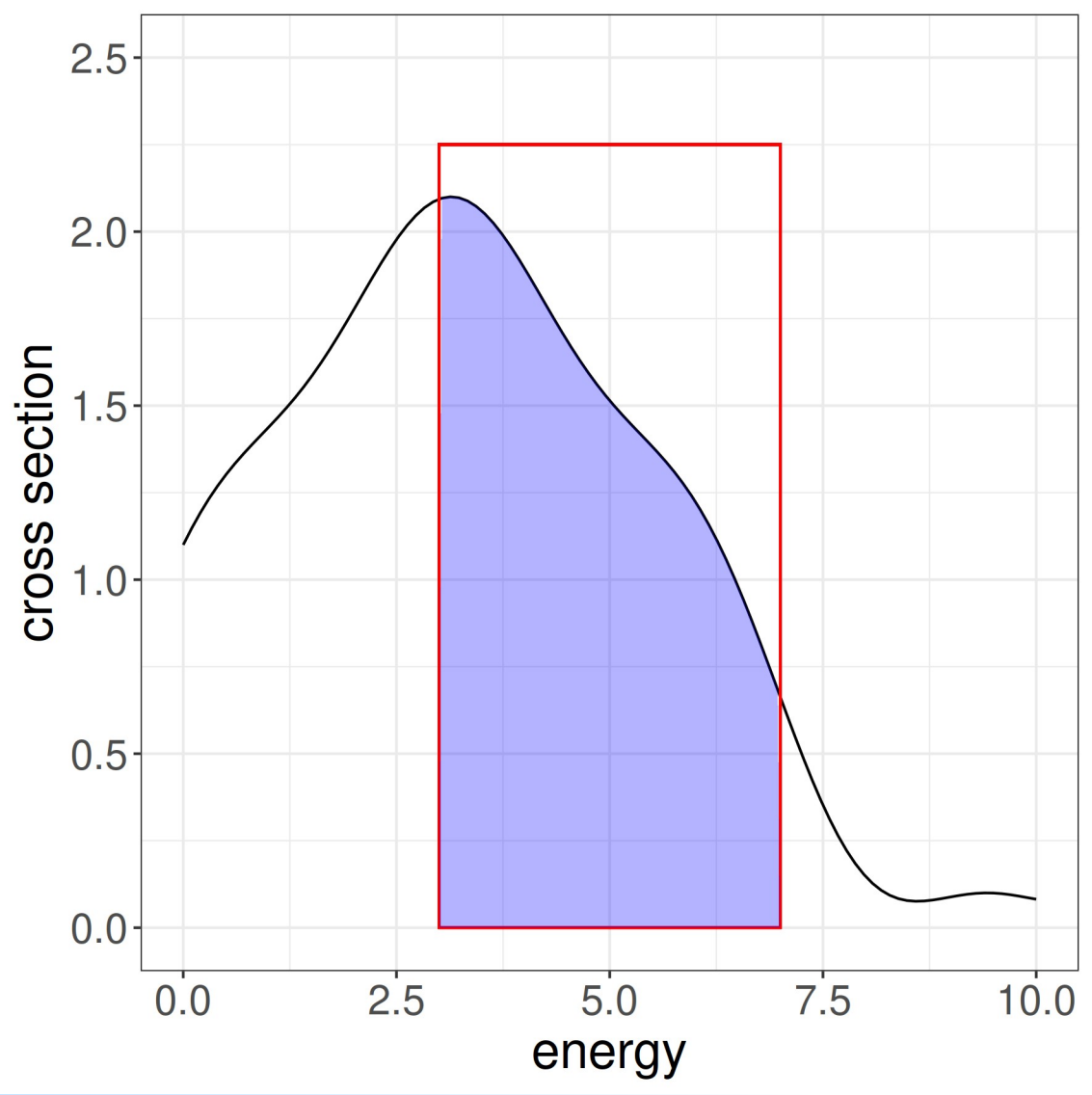
NAPC/IAEA

Joint ICTP/IAEA School on AI for Nuclear, Plasma and Fusion Science
Trieste, 2023

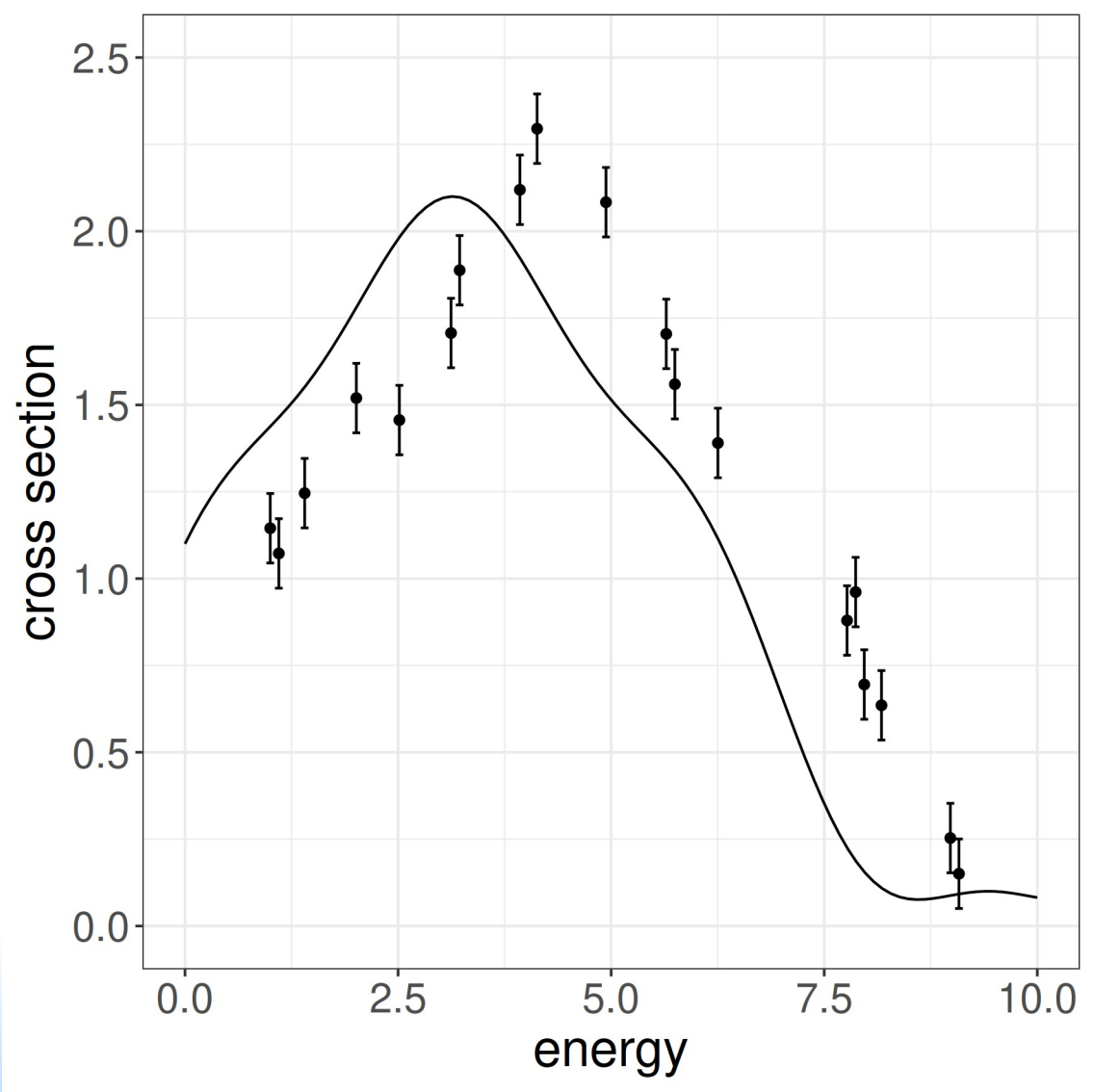
System of excitation functions



Measurement with finite resolution



Measurement with uncertain energy calibration



General setting

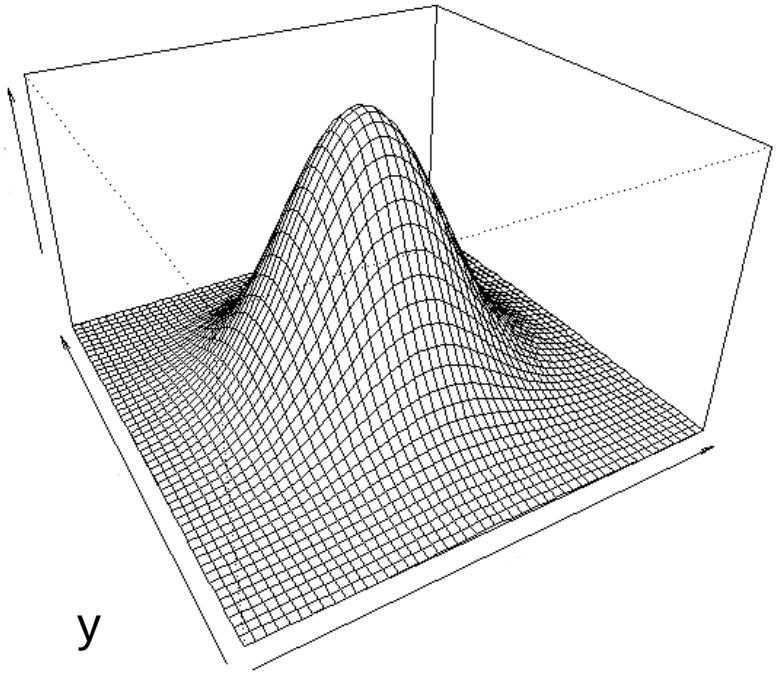
- System of unknown functions, values, parameters linked together by some relationships (e.g., sum of some functions should yield another function)
- Partial observations (noisy, finite resolution, not precisely known x-value)

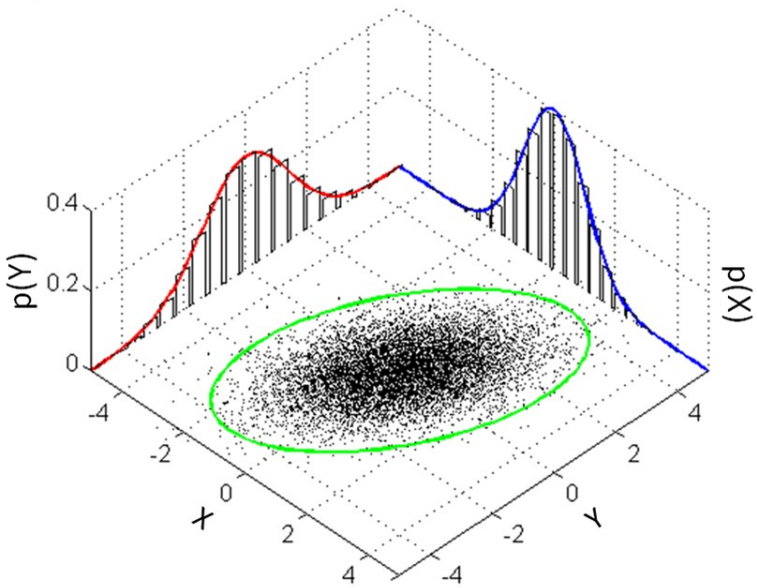
Specific focus of this lecture

- System of unknown functions, values, parameters linked together by some relationships (e.g., sum of some functions should yield another function)
- Partial observations (noisy, finite resolution, not precisely known x-value)

Multivariate normal distribution

$$\mathcal{N}(\vec{x}\,|\,\vec{x}_0, \mathbf{A}_0) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{A}_0|}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{x}_0)^T \mathbf{A}_0^{-1} (\vec{x} - \vec{x}_0) \right)$$





$$\vec{x} = \begin{pmatrix} \text{altitude} \\ \text{temperature} \end{pmatrix}$$

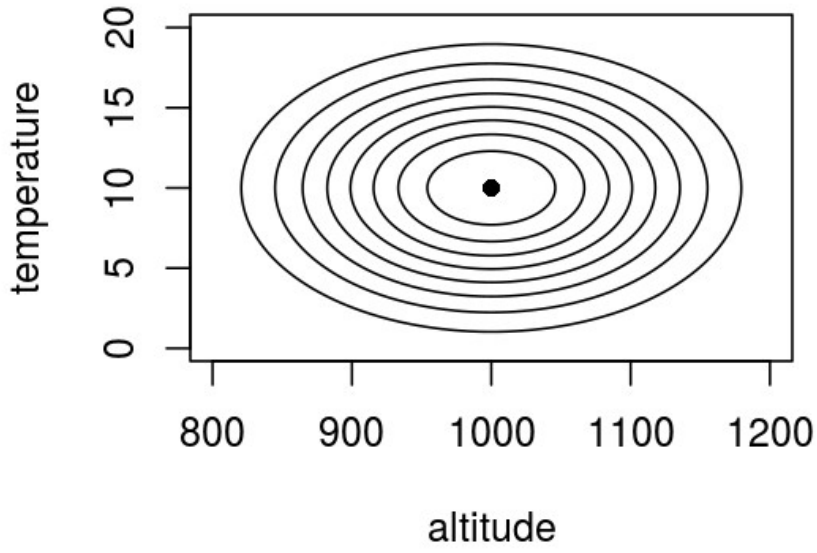
$$\vec{x}_0 = \begin{pmatrix} 1000 \\ 10 \end{pmatrix}$$

Covariance matrix

$$\vec{x} = \begin{pmatrix} \text{altitude} \\ \text{temperature} \end{pmatrix}$$

Variance / uncertainty squared

$$\mathbf{A}_0 = \begin{pmatrix} 100^2 & 0 \\ 0 & 5^2 \end{pmatrix}$$

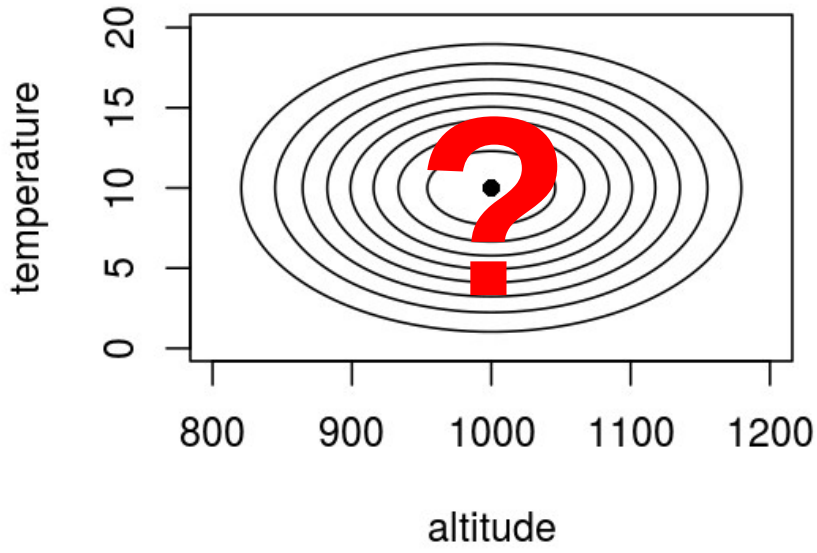


$$\rho(\text{temperature}, \text{altitude})$$

Covariance matrix

$$\vec{x} = \begin{pmatrix} \text{altitude} \\ \text{temperature} \end{pmatrix}$$

$$\mathbf{A}_0 = \begin{pmatrix} 100^2 & 0 \\ 0 & 0 \end{pmatrix}$$



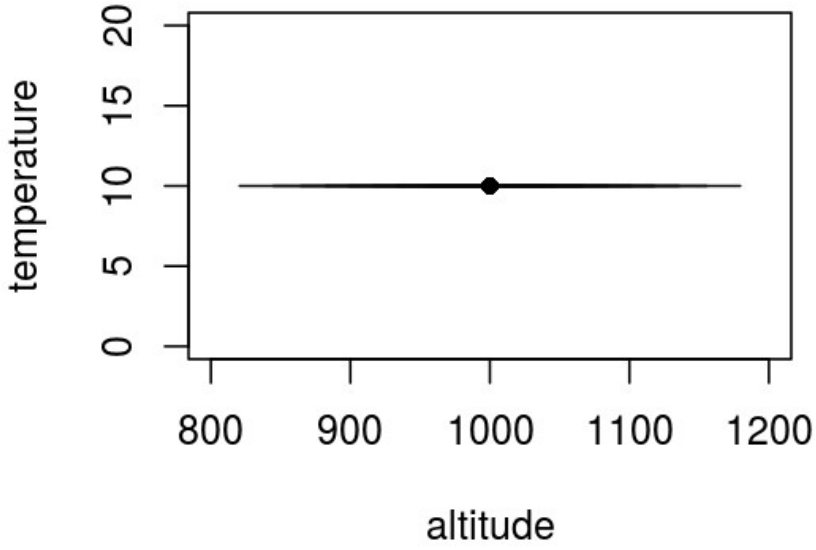
$$\rho(\text{temperature}, \text{altitude})$$

Covariance matrix

$$\vec{x} = \begin{pmatrix} \text{altitude} \\ \text{temperature} \end{pmatrix}$$

Variance / uncertainty squared

$$\mathbf{A}_0 = \begin{pmatrix} 100^2 & 0 \\ 0 & 0 \end{pmatrix}$$

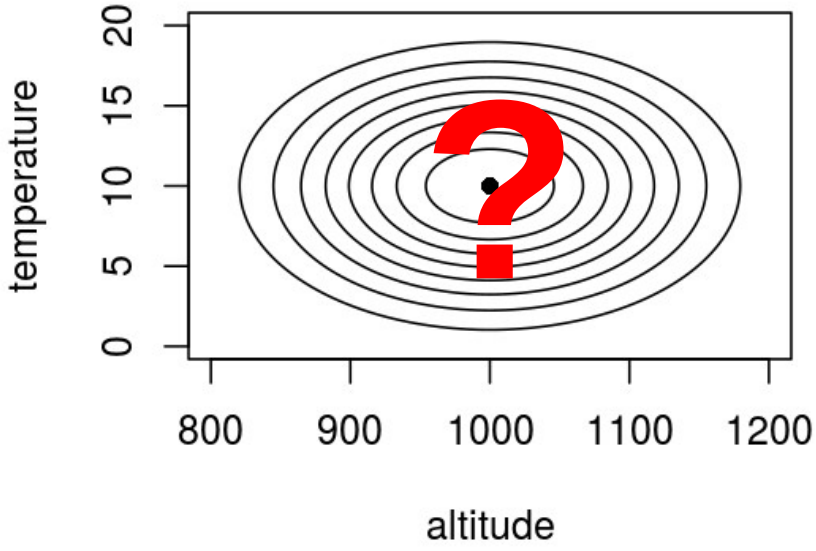


$$\rho(\text{temperature}, \text{altitude})$$

Off-diagonal elements

$$\vec{x} = \begin{pmatrix} \text{altitude} \\ \text{temperature} \end{pmatrix}$$

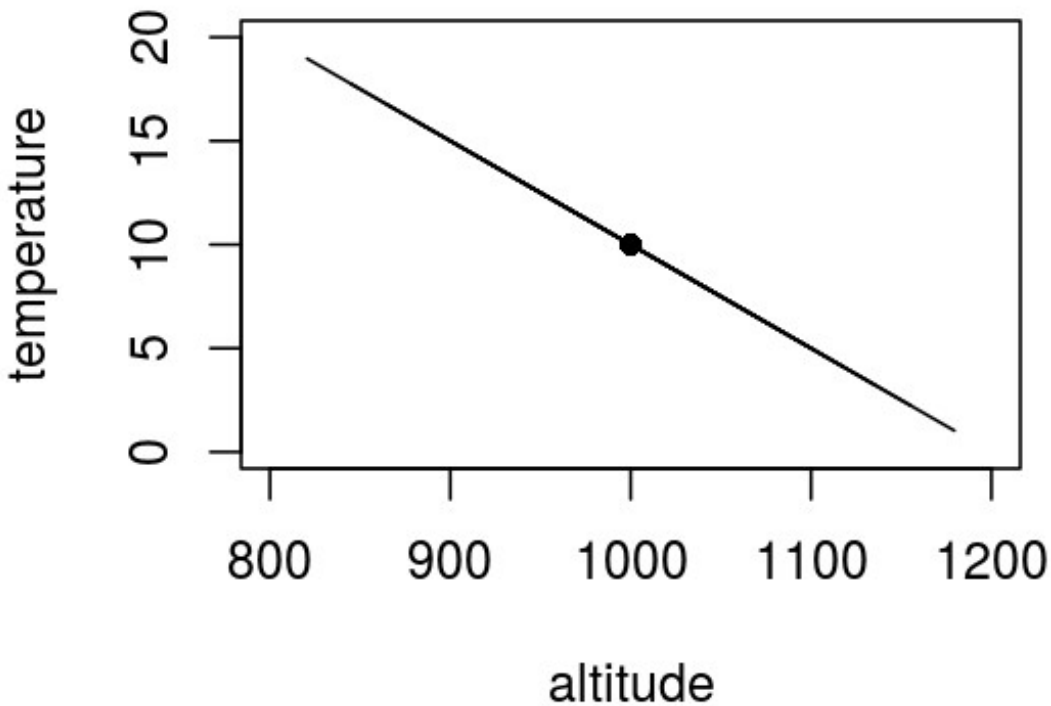
$$\mathbf{A}_0 = \begin{pmatrix} 100^2 & -100 \cdot 5 \\ -100 \cdot 5 & 5^2 \end{pmatrix}$$



Off-diagonal elements

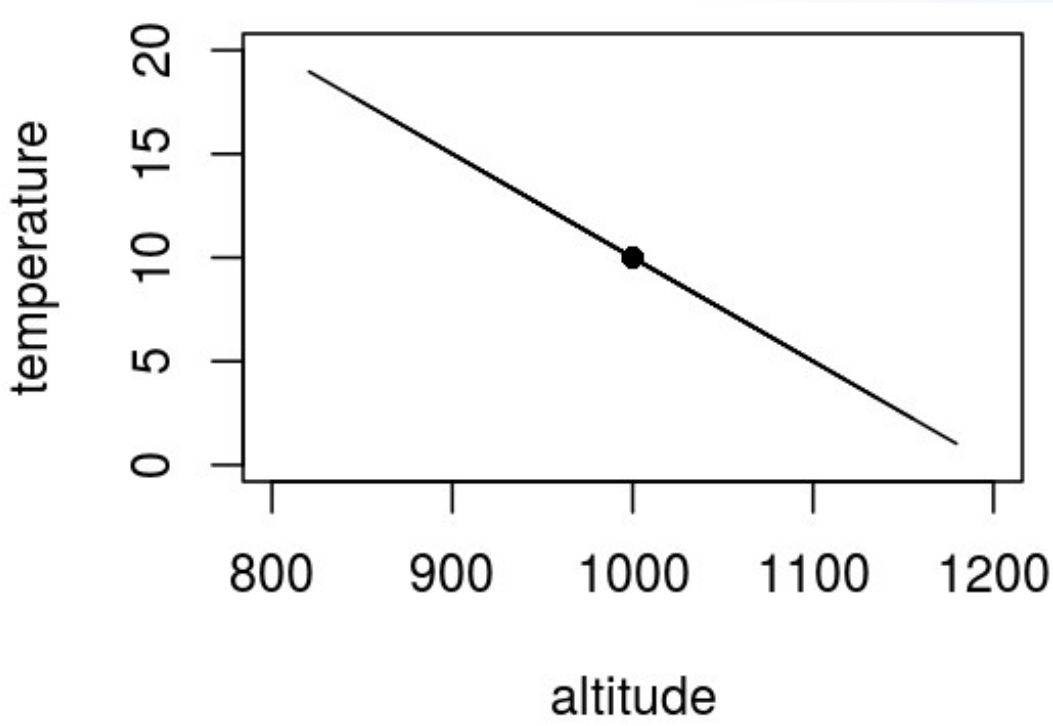
$$\vec{x} = \begin{pmatrix} \text{altitude} \\ \text{temperature} \end{pmatrix}$$

$$\mathbf{A}_0 = \begin{pmatrix} 100^2 & -100 \cdot 5 \\ -100 \cdot 5 & 5^2 \end{pmatrix}$$



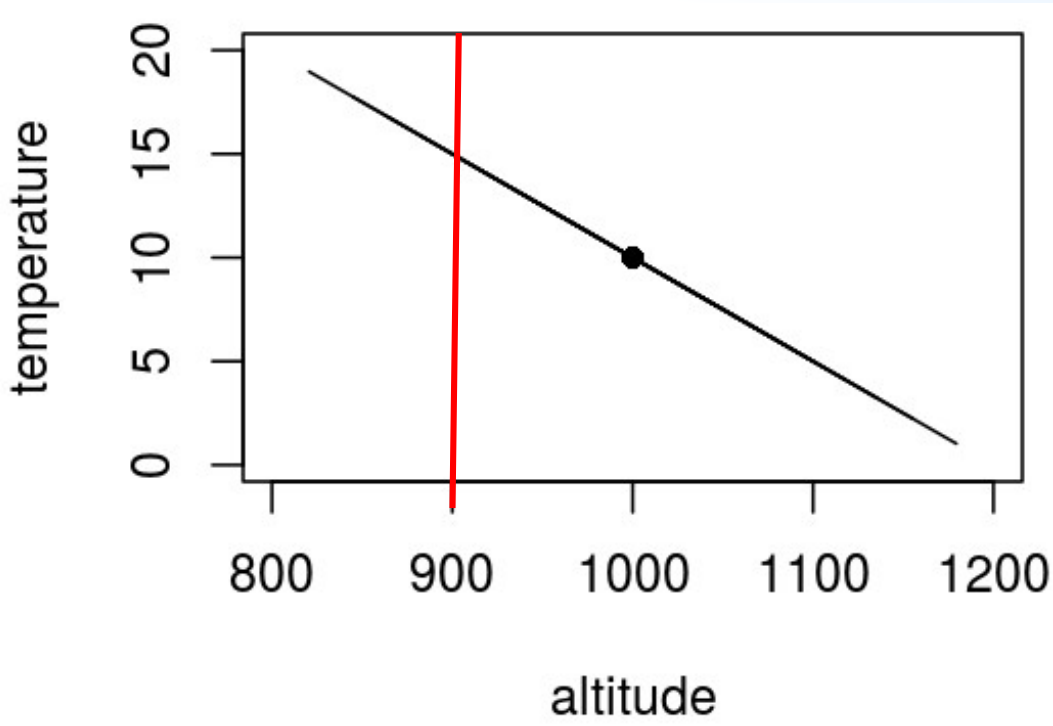
$$|\mathrm{Cov}[X, Y]| \leq \sqrt{\mathrm{Var}[X] * \mathrm{Var}[Y]}$$

Conditional probability density function



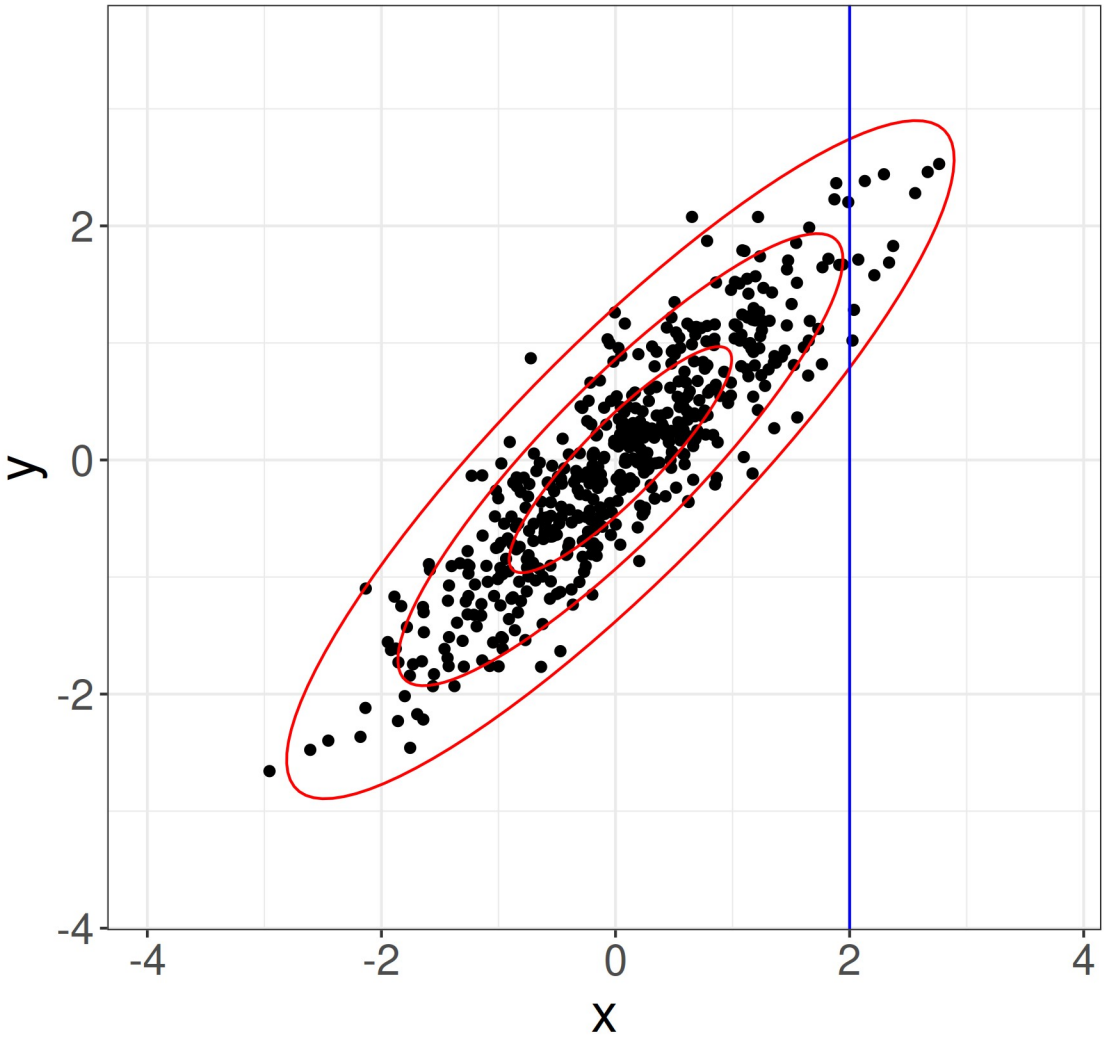
$$\rho(\text{temperature} \,|\, \text{altitude} = 900)$$

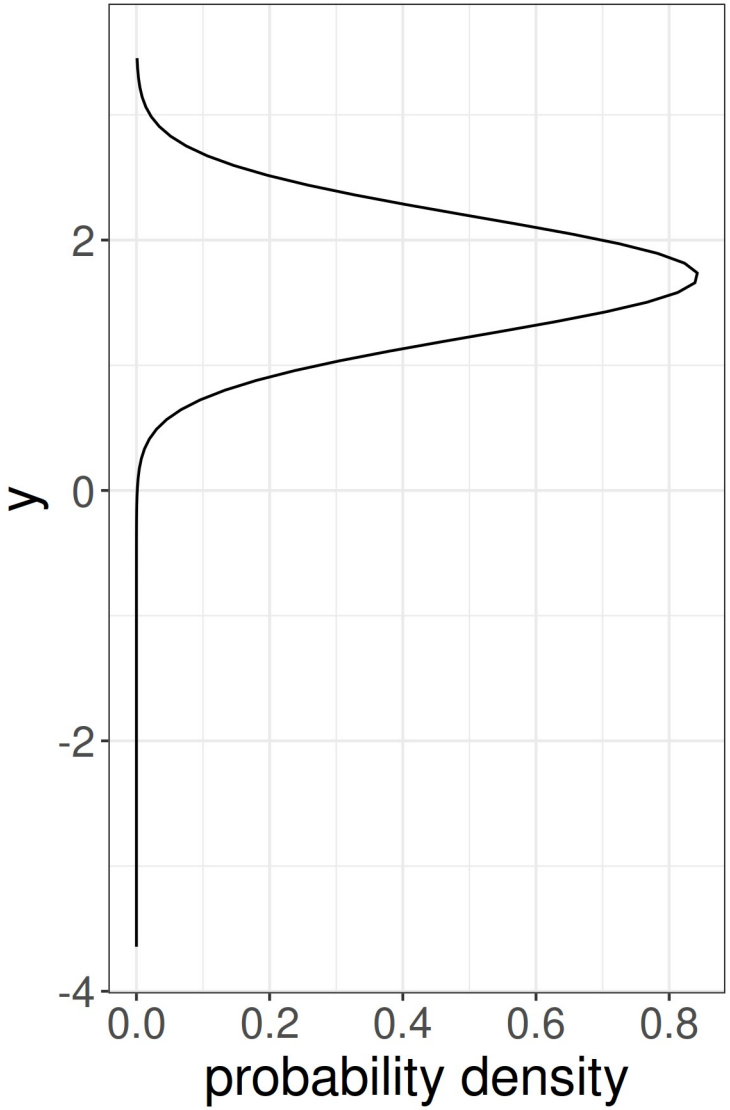
Conditional probability density function



$$\rho(\text{temperature} \,|\, \text{altitude} = 900)$$

Correlated non-degenerate case





Conditional pdf of multivariate normal distribution

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Distribution parameters for $\mathbf{x_1}$ if we know $\mathbf{x_2=a}$

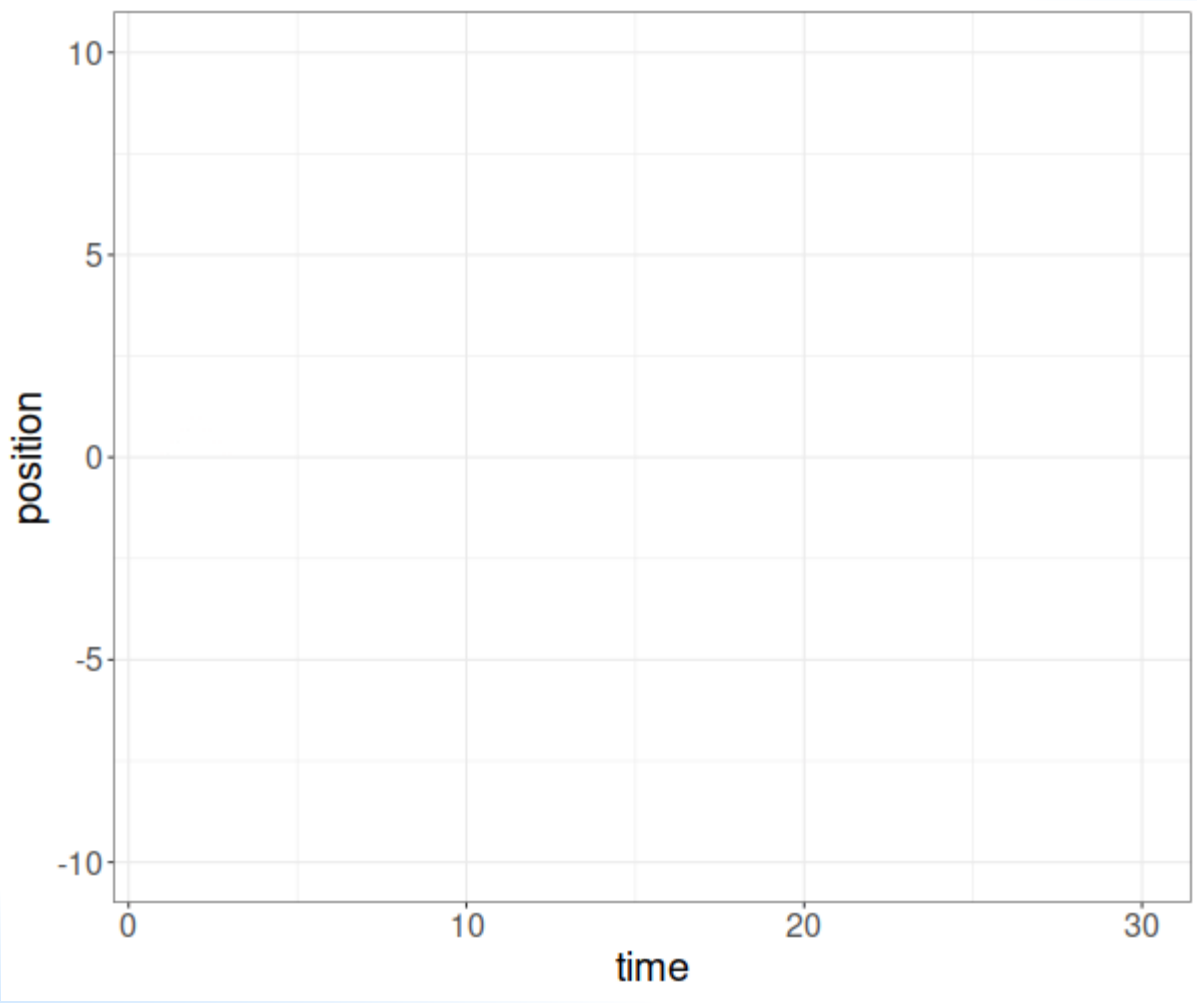
$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{a} - \boldsymbol{\mu}_2\right)$$

$$\overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

$$\rho(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{a}) = \mathcal{N}(\mathbf{x}_1 \,|\, \overline{\mu}, \boldsymbol{\Sigma})$$

Stochastic process

e.g. Brownian motion



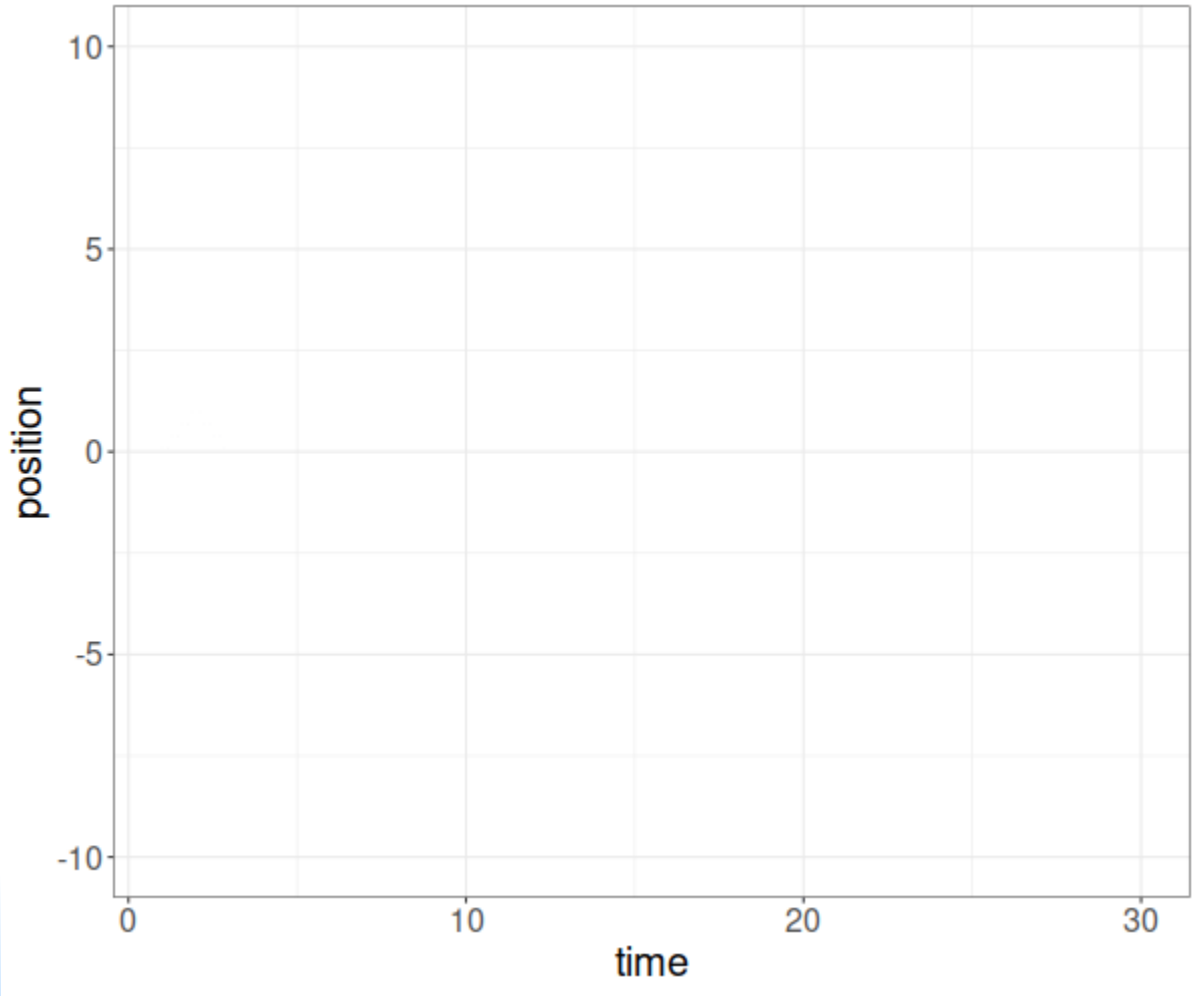
one jump/second

jumps: +1 / -1

stochastic process
is associated with a
random function

Stochastic process

e.g. Brownian motion

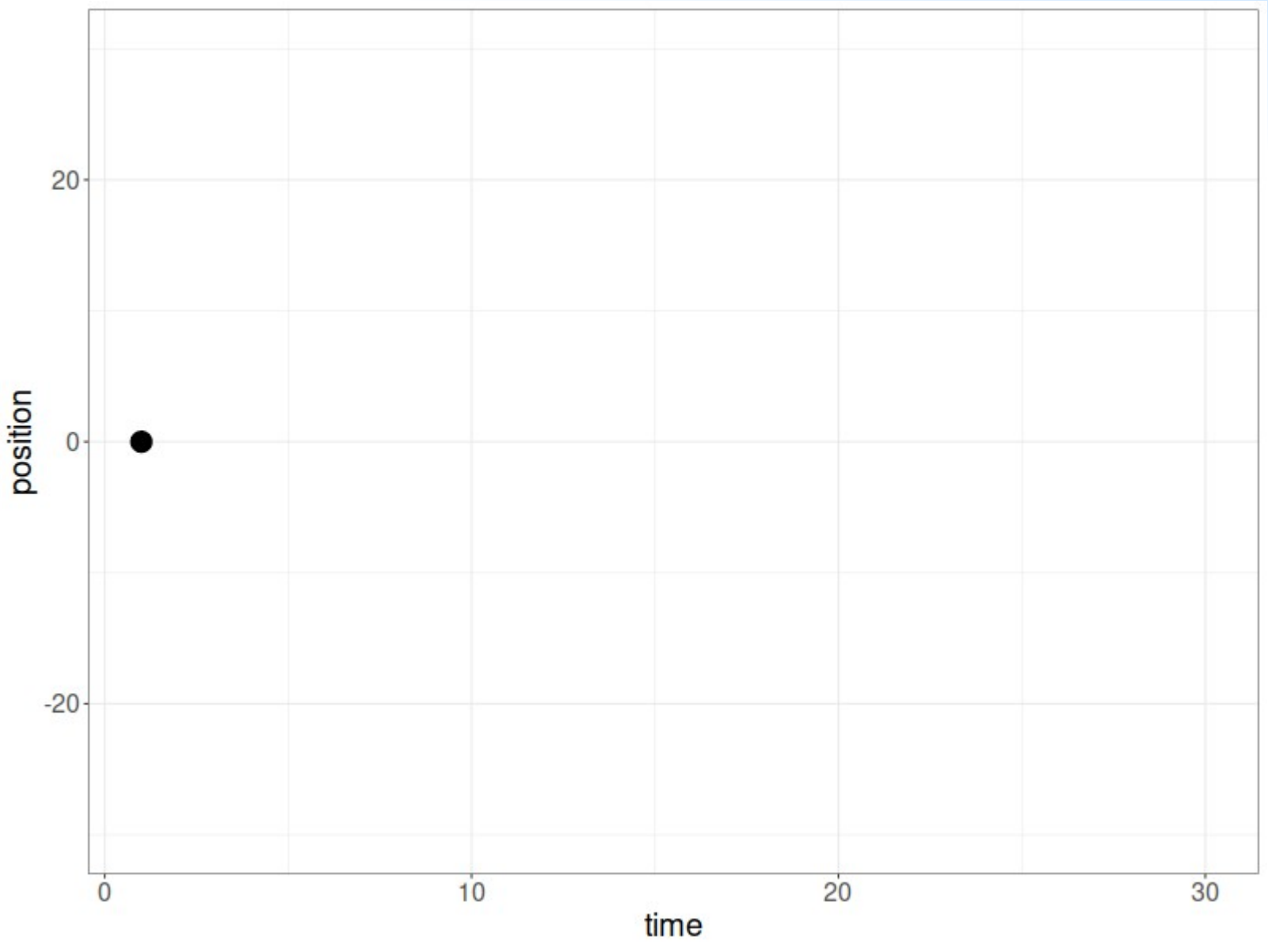


one jump/second

jumps: +1 / -1

stochastic process
is associated with a
random function

Brownian motion



Spatial distribution after some time

- At time **t**, we had a number of t jumps (+1 or -1)
- N_p is number of +1 steps
- N_m is number of -1 steps
- $x(t) = N_p - N_m = 2*N_p - t$
- +1 and -1 jumps are equally likely

Spatial distribution after some time

- At time **t**, we had a number of t jumps (+1 or -1)
- N_p number of +1 steps
- N_m number of -1 steps
- $x(t) = N_p - N_m = 2*N_p - t$
- +1 and -1 jumps are equally likely

$$f(k,n,p) = \Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$n = t$
$k = (x + t) / 2$
$p = 0.5$

Normal approximation (N>>k, k>>0

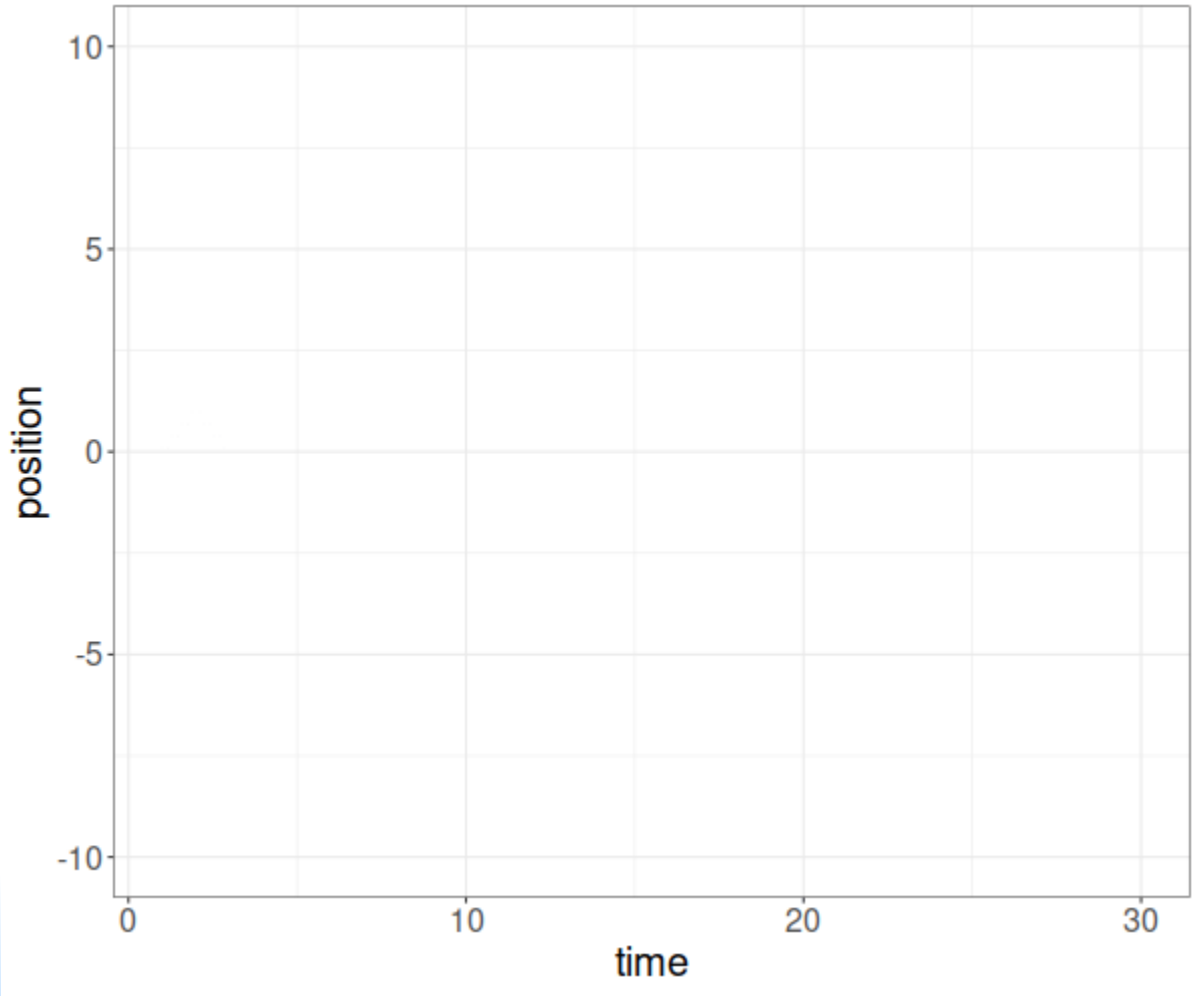
$$f(k,n,p) = \Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

n = t
k = (x + t) / 2
p = 0.5

$$\hat{k} = np \qquad \mathrm{Var}(k) = np(1-p)$$

$$\boxed{\hat{x}_t = 0 \qquad \mathrm{Var}(x_t) = t}$$

What about observations?



one jump/second

jumps: +1 / -1

stochastic process
is associated with a
random function

Conditional distribution

Starting with $x_0 = 0$

$$\hat{x}_t = 0 \qquad \mathrm{Var}(x_t) = t$$

Evolution after we observed position x_{obs} at t_{obs}

$$\hat{x}_t = x_{\mathrm{obs}} \qquad \mathrm{Var}(x_t) = (t - t_{\mathrm{obs}})$$

Conditional distribution

Starting at $x_0 = 0$

$$\hat{x}_t = 0 \qquad \mathrm{Var}(x_t) = t$$

$$\rho(x_t)$$

Evolution after we observed position x_{obs} at t_{obs}

$$\hat{x}_t = x_{\mathrm{obs}} \qquad \mathrm{Var}(x_t) = (t - t_{\mathrm{obs}})$$

$$\rho(x_t | x_{t_{\mathrm{obs}}})$$

Covariance between x_t and x_{t_obs}

$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

$$\rho(x_t, x_{t_{\mathrm{obs}}}) = \rho(x_t | x_{t_{\mathrm{obs}}}) \rho(x_{t_{\mathrm{obs}}})$$

$$\mathrm{Cov}(x_t, x_{t_{\mathrm{obs}}}) = \int\int x_t \, x_{t_{\mathrm{obs}}} \, \rho(x_t, x_{t_{\mathrm{obs}}}) \, dx_t \, dx_{t_{\mathrm{obs}}}$$

Covariance between x_t and x_{t_obs}

$$\mathrm{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

$$\rho(x_t, x_{t_{\mathrm{obs}}}) = \rho(x_t | x_{t_{\mathrm{obs}}}) \rho(x_{t_{\mathrm{obs}}})$$

$$\mathrm{Cov}(x_t, x_{t_{\mathrm{obs}}}) = \int\int x_t \, x_{t_{\mathrm{obs}}} \, \rho(x_t, x_{t_{\mathrm{obs}}}) \, dx_t \, dx_{t_{\mathrm{obs}}}$$

$$\boxed{\mathrm{Cov}(x_t, x_{t_{\mathrm{obs}}}) = \mathrm{Var}(x_{t_{\mathrm{obs}}}) = t_{\mathrm{obs}}}$$

Covariance matrix for our process

$$\mathrm{Var}(x_t) = t$$

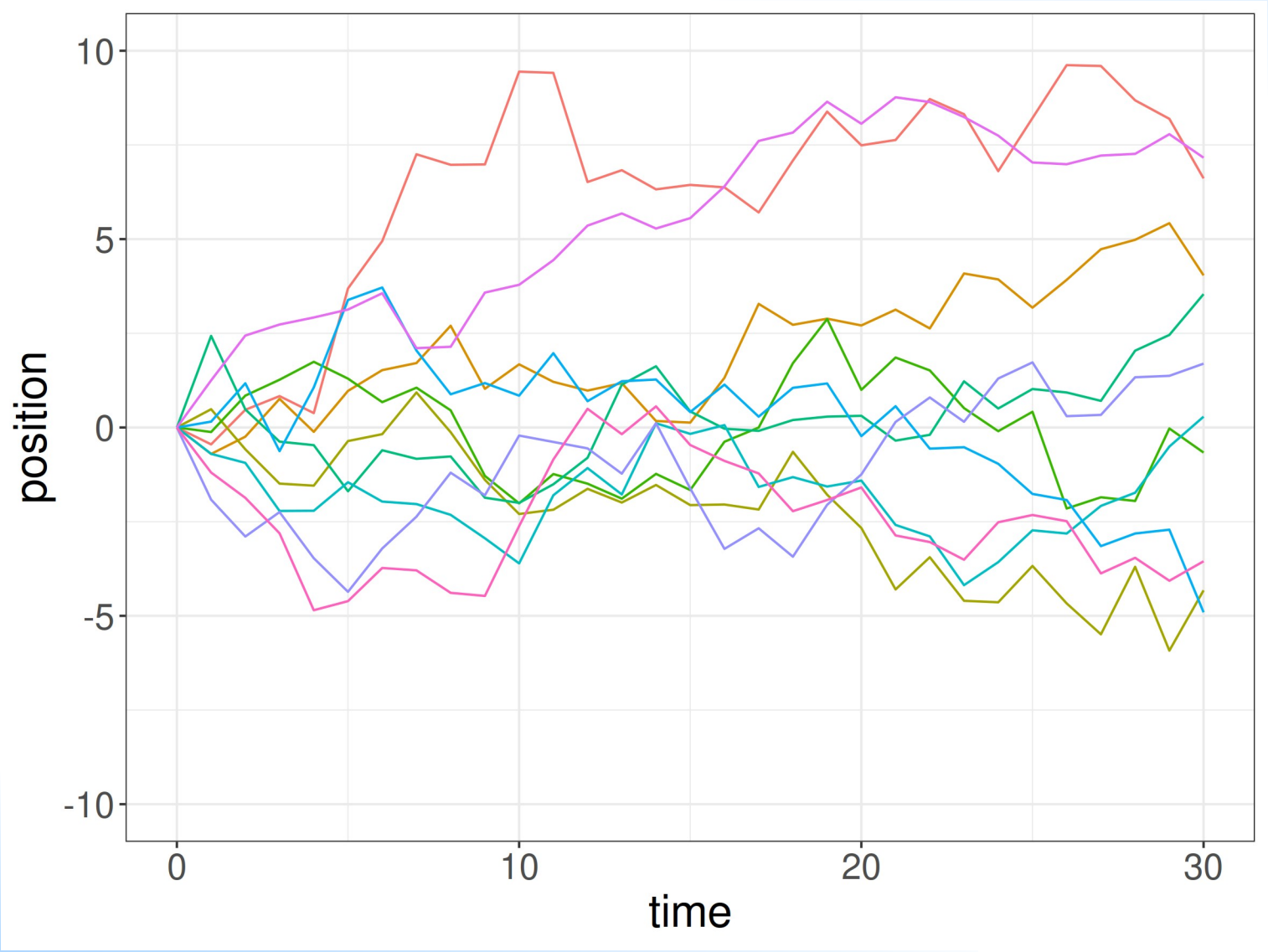
$$\mathrm{Cov}(x_t, x_{t_{\mathrm{obs}}}) = \mathrm{Var}(x_{t_{\mathrm{obs}}}) = t_{\mathrm{obs}}$$

$$\mathbf{A}_0 = \begin{pmatrix} t_{\mathrm{obs}} & t_{\mathrm{obs}} \\ t_{\mathrm{obs}} & t \end{pmatrix}$$

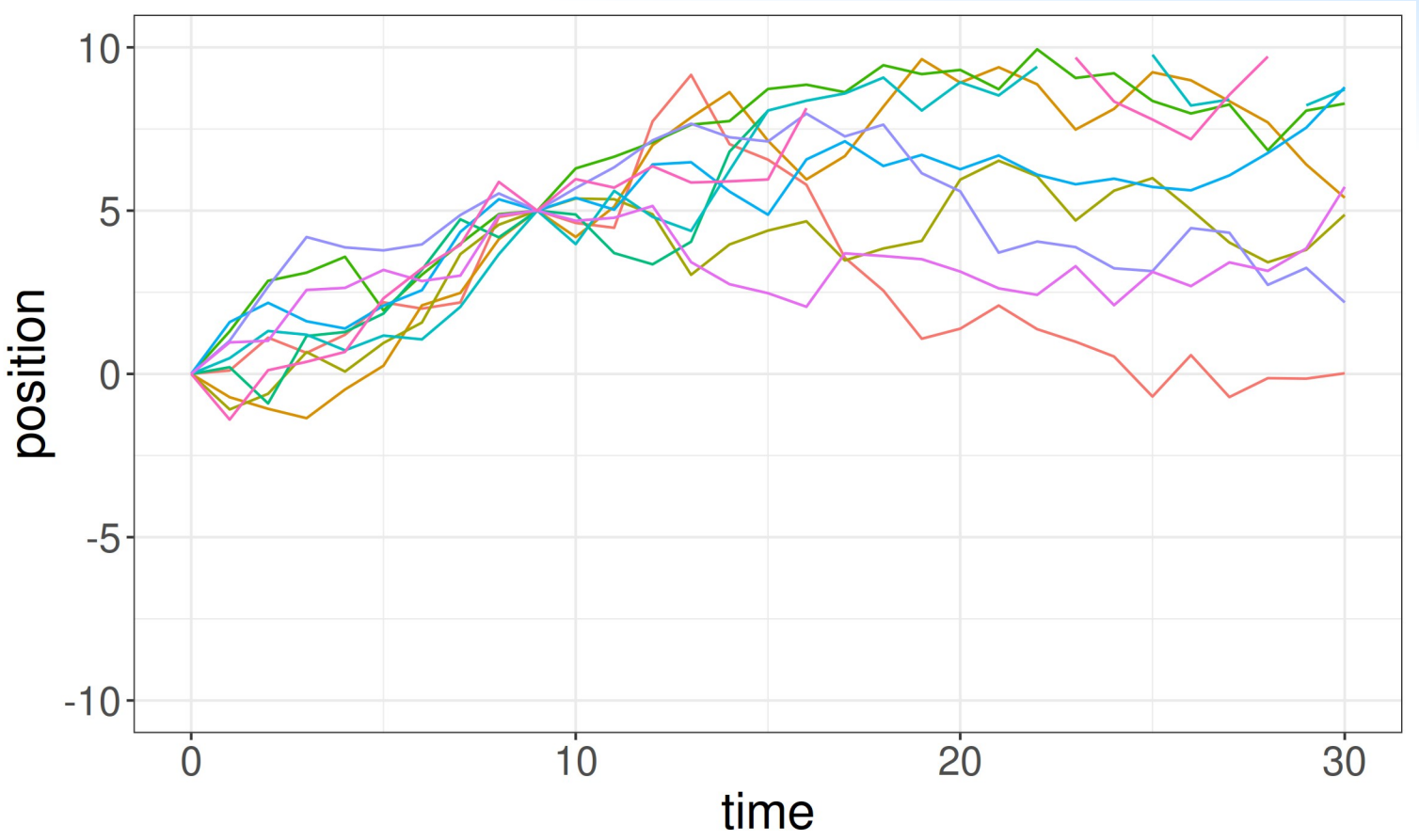
Full covariance matrix up to t=5

$$\mathbf{A}_0 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 & 2 & 2 \\ 0 & 1 & 2 & 3 & 3 & 3 \\ 0 & 1 & 2 & 3 & 4 & 4 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

Sampling from the multivariate normal distribution

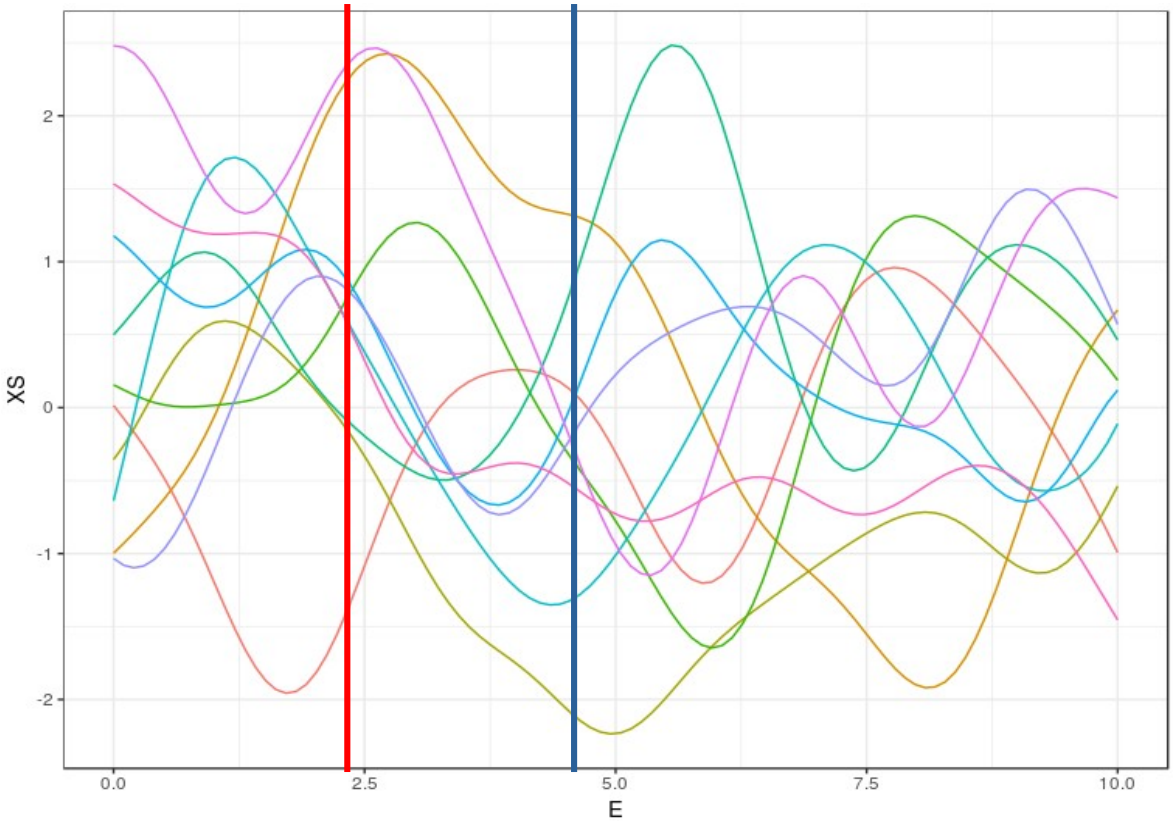


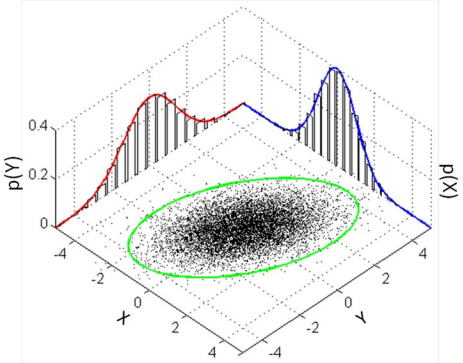
Constraining with observation



$$\rho(x_1, x_2, \ldots, x_{30} \,|\, x_{10} = 5) = \rho(\mathbf{x} \,|\, x_{10} = 5)$$

Characterisation of Gaussian process





All finite sets of function values at different locations are distributed according to a multivariate normal distribution
→ <u>Gaussian process</u>

Specification of a Gaussian process

Specify a function that yields the covariance between function values f($\mathbf{x}_1$) and f($\mathbf{x}_2$) for any possible pair $\mathbf{x}_1$ and $\mathbf{x}_2$. This function is called a **covariance function**.

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \delta^2 \exp\left(-\frac{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^2}{2\lambda^2}\right)$$

Specify another function μ(**x**) that yields the center value of the process at location **x**. This function is called a **mean function**.

$$\mu(\mathbf{x}) = 0$$

Some features of Gaussian process regression

- Predictions can be calculated one at a time

$$\rho(x_1, x_2, \ldots, x_{30} \,|\, x_{10} = 5)$$

$$\rho(x_1 \,|\, x_{10} = 5) \quad \rho(x_2 \,|\, x_{10} = 5)$$

- Usually (depending) on covariance matrix, GP regression amounts to fitting infinitely many coefficients

$$f(x) = \sum_{k=1}^{\infty} a_k f_k(x) \qquad a_k \sim \mathcal{N}(\mu_k, \sigma_k)$$

Sampling from a Gaussian process

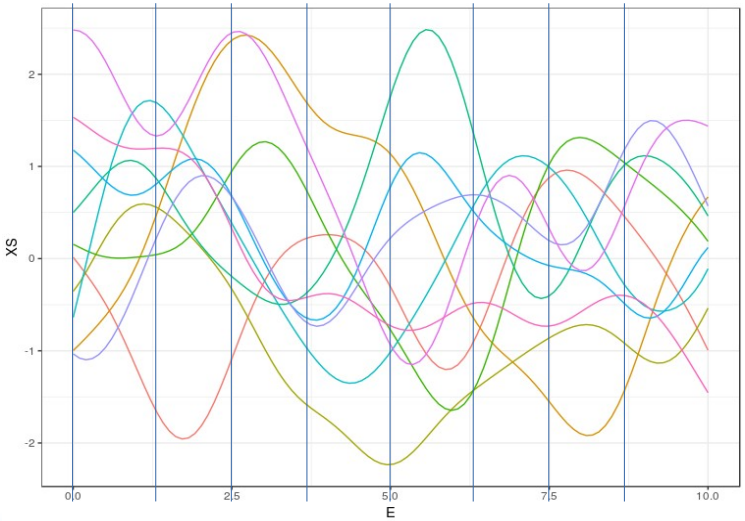
$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \delta^2 \exp\left(-\frac{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^2}{2\lambda^2}\right)$$

$$\mu(\mathbf{x}) = 0$$

$$\mathbf{K} = \begin{pmatrix} \kappa(x_1, x_1) & \kappa(x_1, x_2) & \kappa(x_1, x_3) & \dots \\ \kappa(x_2, x_1) & \kappa(x_2, x_2) & \kappa(x_2, x_3) & \dots \\ \kappa(x_3, x_1) & \kappa(x_3, x_2) & \kappa(x_3, x_3) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$\vec{m} = \begin{pmatrix} \mu(x_1) \\ \mu(x_2) \\ \mu(x_3) \\ \vdots \end{pmatrix}$$

$$\mathbf{y} \sim \mathcal{N}(\vec{m}, \mathbf{K})$$



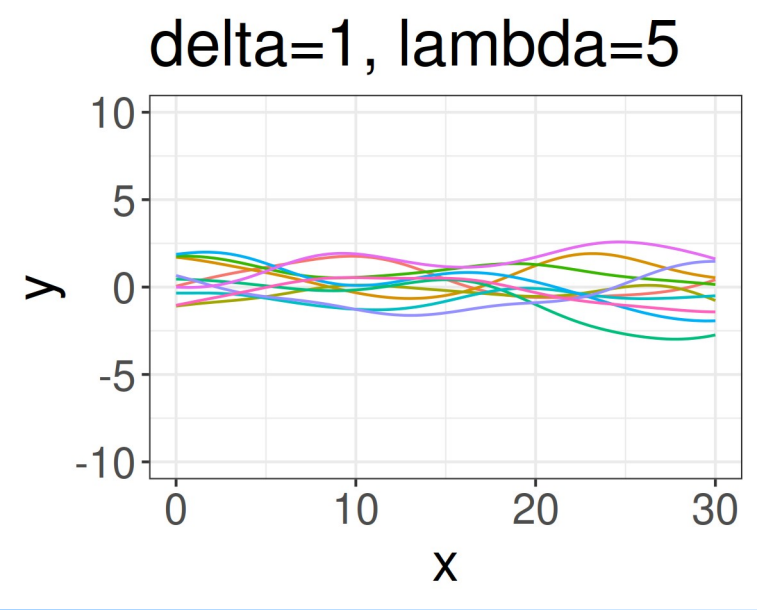
X_1 X_2 X_3 ...

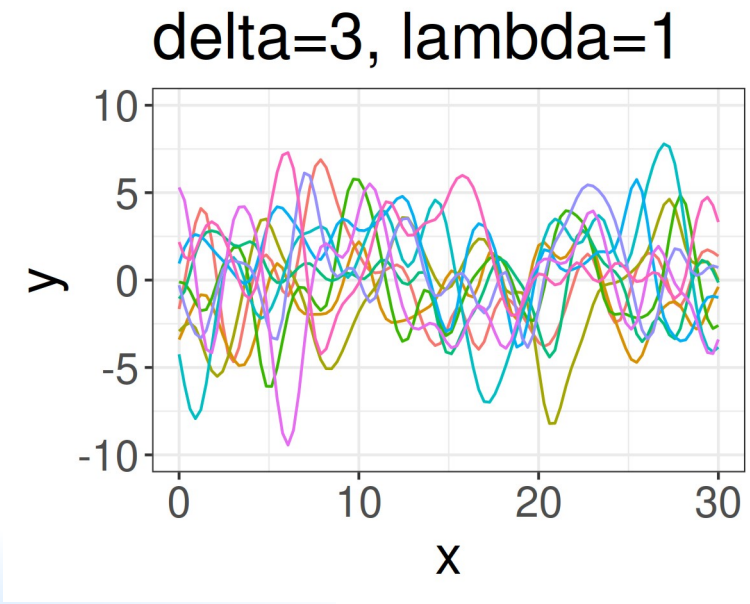
Hyperparameters

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \delta^2 \exp\left(-\frac{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^2}{2\lambda^2}\right)$$

amplitude

length scale

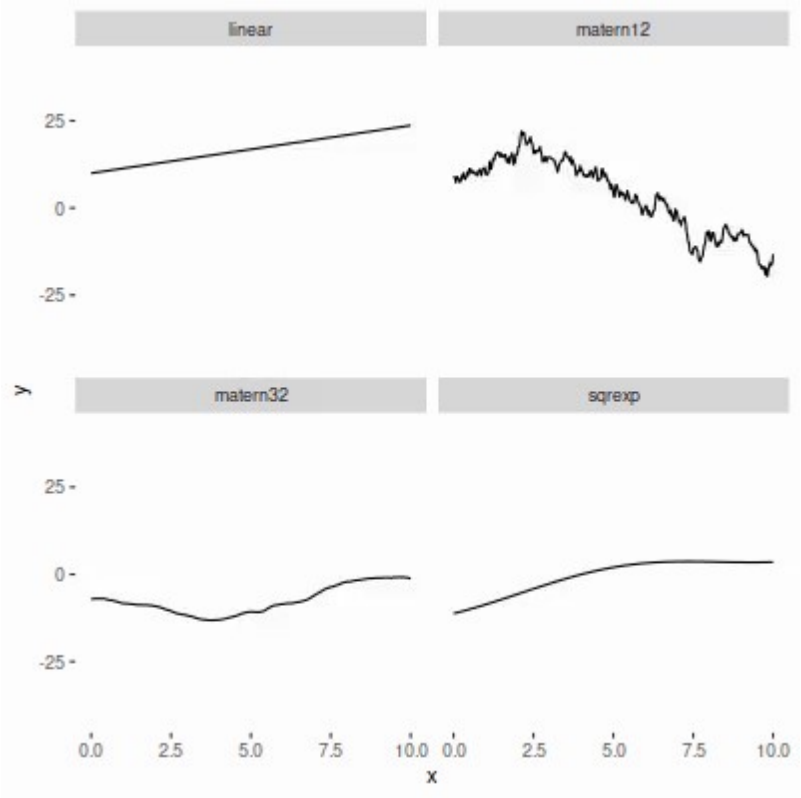




Examples of Gaussian processes

$$\kappa_{\text{lin}}(x_1, x_2) = x_1 x_2 {\delta_k}^2 + {\delta_d}^2$$

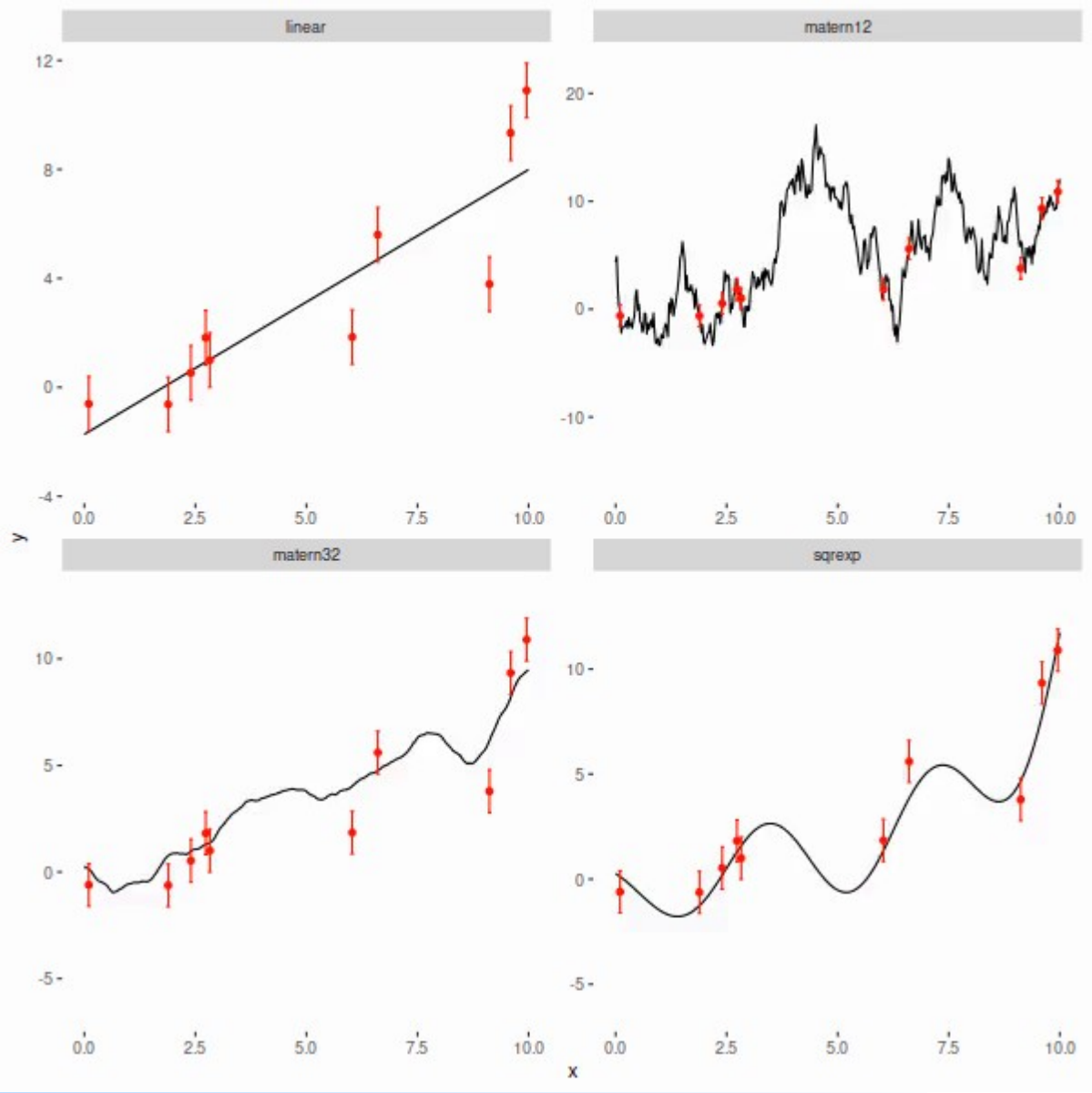
$$\kappa_{\text{mat},1/2}(x_1, x_2) = \delta^2 \exp\left(-\frac{|x_1 - x_2|}{\rho}\right)$$



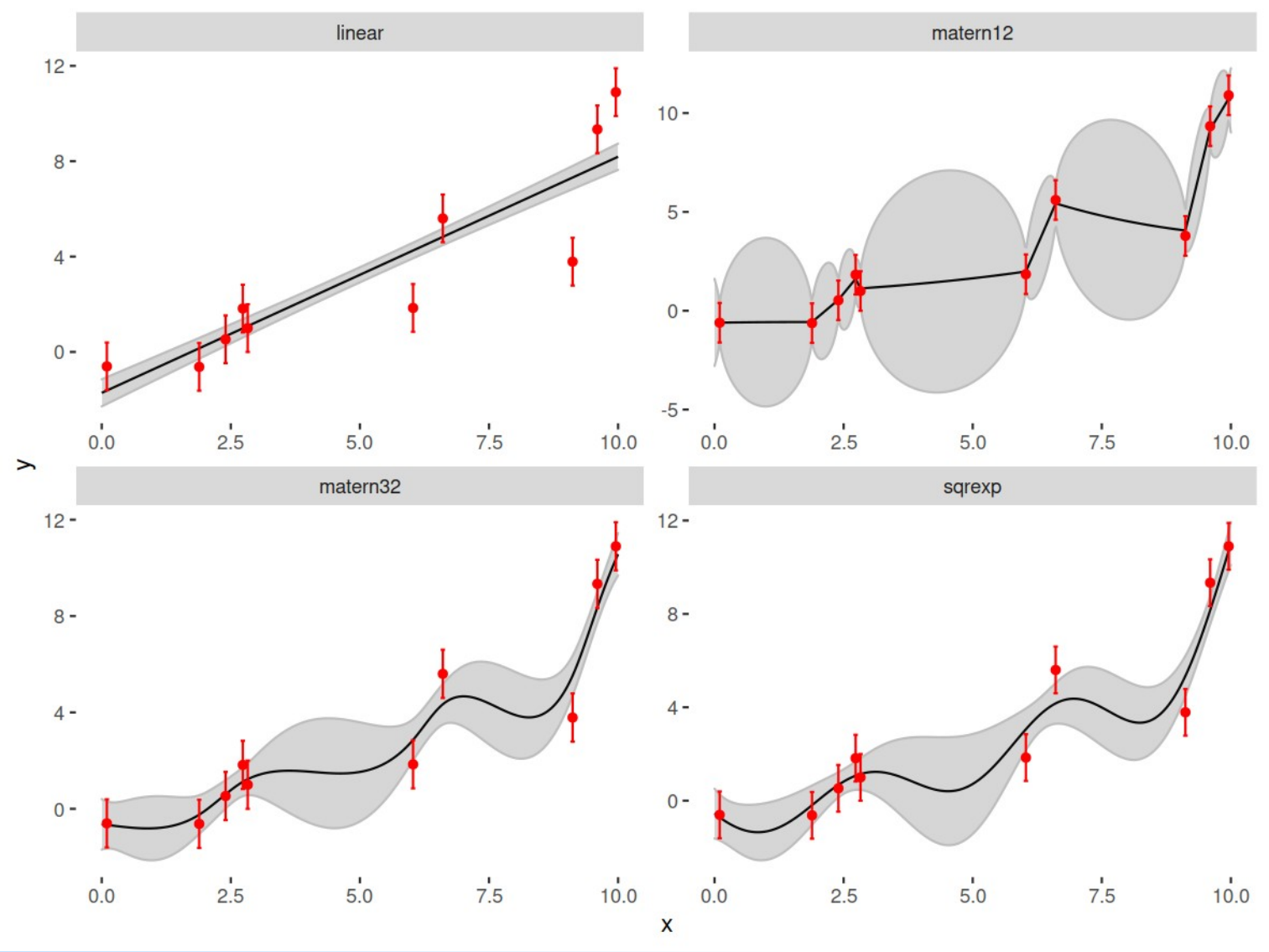
$$\kappa_{\text{mat},3/2}(x_1, x_2) = \delta^2 \left(1 + \frac{\sqrt{3}|x_1 - x_2|}{\rho}\right) \exp\left(-\frac{\sqrt{3}|x_1 - x_2|}{\rho}\right)$$

$$\kappa_{\text{sqrexp}}(x_1, x_2) = \delta^2 \exp\left(-\frac{1}{2\lambda^2}(x_1 - x_2)^2\right)$$

Conditioning on observations



Conditioning on observations



Covariance brewing

Combination rules for covariance functions

$$\kappa_{1+2}(x_1, x_2) = \kappa_1(x_1, x_2) + \kappa_2(x_1, x_2)$$

$$\kappa_{1\times 2}(x_1, x_2) = \kappa_1(x_1, x_2) \times \kappa_2(x_1, x_2)$$

Squared exponential covariance functions

$$\kappa_1(x_1, x_2) = {\delta_1}^2 \exp\left(\frac{1}{2{\lambda_1}^2}(x_1 - x_2)^2)\right)$$

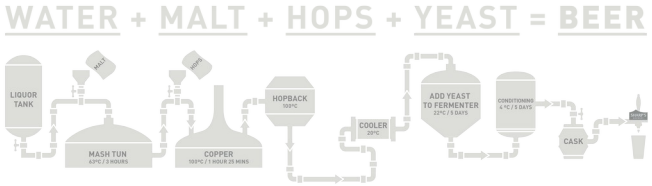
$$\kappa_2(x_1, x_2) = {\delta_2}^2 \exp\left(\frac{1}{2{\lambda_2}^2}(x_1 - x_2)^2)\right)$$

Transition kernel

$$\tau_1(x_1, x_2) = \sigma(x_1)\sigma(x_2)$$

$$\tau_2(x_1, x_2) = (1 - \sigma(x_1)(1 - \sigma(x_2))$$

$$\sigma(x) = \frac{1}{1 + \exp(-k(x - x_0))}$$

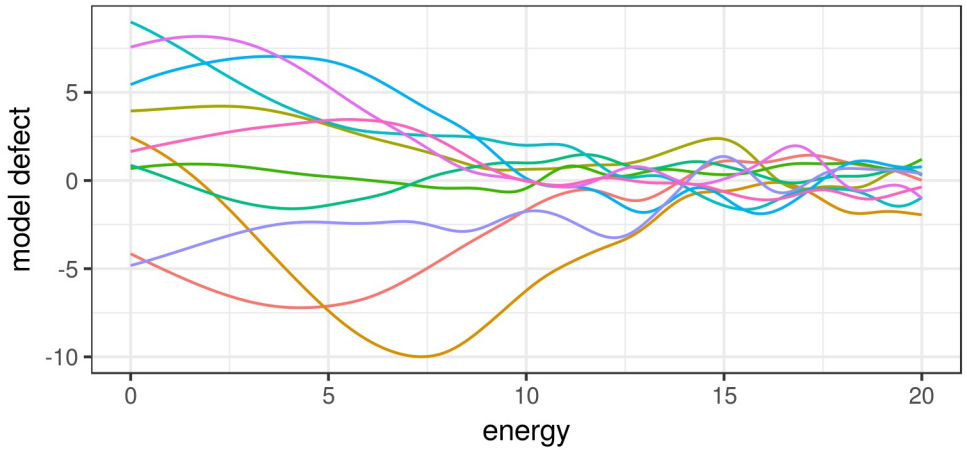


$$\boxed{\kappa_{\text{comp}}(x_1, x_2) = \tau_1(x_1, x_2) \times \kappa_1(x_1, x_2) + \tau_2(x_1, x_2) \times \kappa_2(x_1, x_2)}$$

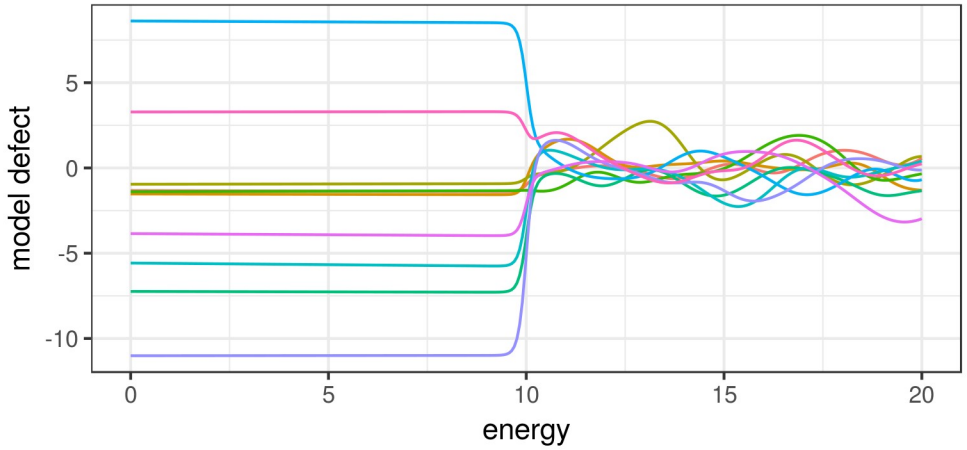
Example of compound covariance function

$$\kappa_{\text{comp}}(x_1, x_2) = \tau_1(x_1, x_2) \times \kappa_1(x_1, x_2) + \tau_2(x_1, x_2) \times \kappa_2(x_1, x_2)$$

$\delta_1 = 5$, $\lambda_1 = 5$, $\delta_2 = 1$, $\lambda_2 = 1$, $k = 0.5$, $x_0 = 10$



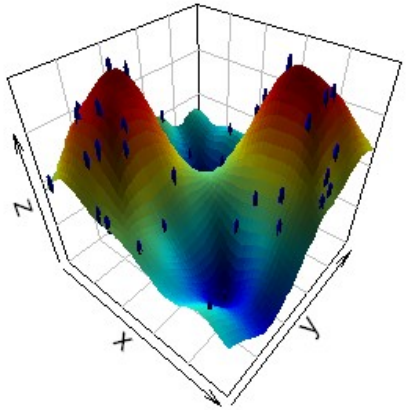
$\delta_1 = 5$, $\lambda_1 = 500$, $\delta_2 = 1$, $\lambda_2 = 1$, $k = 10$, $x_0 = 10$



ML view on GPR

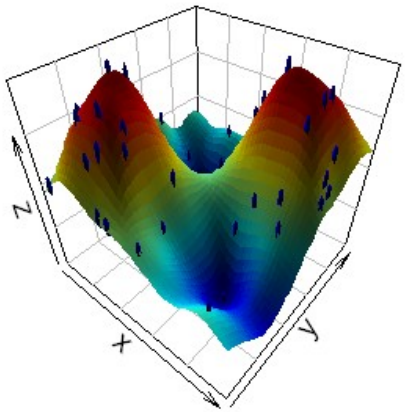
- Gaussian process regression is a method to learn a non-linear function **f(x)** from observations $\{(x_i, y_i)\}_{i=1..N}$
- It can be used for regression and classification problems

It is an alternative to other ML techniques, e.g., random forests and neural networks with its own pros & cons

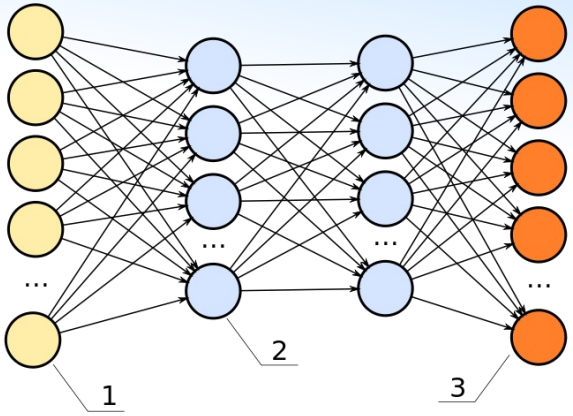


Comparison to neural networks

GP processes



artificial neural networks



<u>Both approaches</u> …

… are methods for classification and regression

… are universal function approximators

<u>Neural networks</u> …

… scale better to large data sets

… are able to capture non-local features

… are difficult to interpret

<u>GP processes</u> …

… are statistical methods from the ground up (uncertainties)

… facilitate the incorporation of prior assumptions

… interface well with existing nuclear data evaluation methods

Useful resources

- C.E. Rasmussen, C. K. I. Williams, “Gaussian processes for machine learning”
- D. Duvenaud, “The Kernel Cookbook”, https://www.cs.toronto.edu/~duvenaud/cookbook/