
UKRI
Science and Technology Facilities Council
Hartree Centre
Welcome

Hartree Centre

INTELLECTUAL PROPERTY RIGHTS NOTICE:

For further details, please email us: hartreetraining@stfc.a.c.uk

Science and Technology Facilities Council

Hartree Centre

Big Data Week 10

Bayesian approaches

Dr Simon Goodchild
Data Science group leader, Hartree Centre

Science and Technology Facilities Council

Hartree Centre

Summary

- Recap probability and probabilistic models from Lecture 7
- Conditional probability
- Bayes' theorem
- Bayesian inference – general principles
- Sequential Bayesian inference
- Markov chains
- Bayesian graphical models
- *Break*
- Kalman filter
- Hidden Markov models

Science and Technology Facilities Council

Hartree Centre

Conditional probability and Bayes' theorem

Science and Technology Facilities Council

Hartree Centre

Probabilistic models: a recap

A probabilistic model is a set of probability distributions for each point of the sample space.

Bayesian methods make this explicit by giving a probability distribution for the model parameters.

Conditional probability

$\mathrm{P}(A \cap B)$ is the probability that events A and B both occur.

A and B are *independent* if $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$

The *conditional probability* is the probability that A occurs given that B already has. It is defined as

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

Note that if A and B are independent, the conditional probability is just $\mathrm{P}(A|B) = P(A)$

Bayes' theorem

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}$$

$\mathrm{P}(A|B)$ *posterior probability:* the probability of A happening given B

$\mathrm{P}(A)$ *prior probability:* probability of A happening before observing any other events.

$\mathrm{P}(B)$ *normalisation*: the overall probability of B happening.

If there are a number of different possible outcomes A_i of the experiment, then

$$\mathrm{P}(B) = \sum_i \mathrm{P}(B|A_i)\mathrm{P}(A_i)$$

Derivation:

By definition, $\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\mathrm{P}(B) = \mathrm{P}(B|A)\mathrm{P}(A)$

and re-arrange to obtain Bayes' theorem.

Bayes' theorem: examples

A disease has a 1% prevalence in the population.
A test for this disease has 90% true positive rate and 5% false positive rate.
If a person tests positive, what is the probability that they have the disease?

$$P(\text{disease}|\text{positive}) = \frac{P(\text{positive}|\text{disease})P(\text{disease})}{P(\text{positive}|\text{disease})P(\text{disease}) + P(\text{positive}|\text{no disease})P(\text{no disease})}$$

$$\frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.05 \times 0.99} = \frac{18}{117} \approx 0.15$$

Bayes' theorem: examples

A chemist has discovered a new radioactive element, and wants to know how rapidly it decays.
Radioactive decays are a *Poisson process* which means the time between decays follows an exponential distribution.

$$\Delta T \sim \mathrm{Exp}(\lambda) \qquad P(\Delta T) = \lambda \mathrm{e}^{-\lambda \Delta T}$$

Can also use an exponential distribution as the prior probability (as it covers the whole possible range)

$$\lambda \sim \mathrm{Exp}(\beta)$$

If a decay interval ΔT_0 is observed then the posterior distribution is now

$$\mathrm{P}(\lambda|\Delta T_0) \propto \mathrm{P}(\Delta T_0|\lambda)\mathrm{P}(\lambda) = \lambda \mathrm{e}^{-\lambda \Delta T_0} \beta \mathrm{e}^{-\lambda \beta} = \lambda \beta \mathrm{e}^{-\lambda(\beta + \Delta T_0)}$$

The normalisation term (to make sure the area under the whole curve is 1) is a gamma function $\Gamma(2)$: this distribution is called the *gamma distribution.*

$$\mathrm{P}(\lambda|\Delta T_0) = \frac{\lambda(\beta + \Delta T_0)^2 \mathrm{e}^{-\lambda(\beta + \Delta T_0)}}{\Gamma(2)}$$

Gamma distribution:

$$x \sim \mathrm{Gamma}(\alpha, \beta) \text{ means } \mathrm{p}(x) = \frac{\beta^\alpha x^{\alpha - 1} \mathrm{e}^{-\beta x}}{\Gamma(\alpha)}$$

Science and Technology Facilities Council

Hartree Centre

Bayesian inference

Science and Technology Facilities Council

Hartree Centre

Bayesian inference

Bayes' theorem in the form $\mathrm{P}(A|B) = \dfrac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}$ is a theorem – a proven result assuming the axioms of probability.

Bayesian inference uses it to describe a hypothesis, as

$$\mathrm{P}(\text{hypothesis}|\text{evidence}) = \frac{\mathrm{P}(\text{evidence}|\text{hypothesis})\mathrm{P}(\text{hypothesis})}{\mathrm{P}(\text{evidence}|\text{hypothesis})\mathrm{P}(\text{hypothesis}) + \mathrm{P}(\text{evidence}|\text{false hypothesis})}$$

Sequential inference

Bayes' theorem calculates a posterior probability based on the prior and the observations.

If a new set of observations is taken, this posterior probability can be used as the prior in a new Bayesian calculation.

Conjugate priors

Remember the example of the radioactive element and the time between decays:

$$\mathrm{P}(\lambda|\Delta T_0) = \frac{\lambda(\beta+\Delta T_0)^2 \mathrm{e}^{-\lambda(\beta+\Delta T_0)}}{\Gamma(2)}$$ was the posterior probability after observing decay interval ΔT_0

If another observation is made, using sequential inference, the new posterior is

$$\mathrm{P}(\lambda|\Delta T_1) = \frac{\mathrm{P}(\Delta T_1|\lambda)\mathrm{P}(\lambda)}{\mathrm{P}(\Delta T_1)}$$

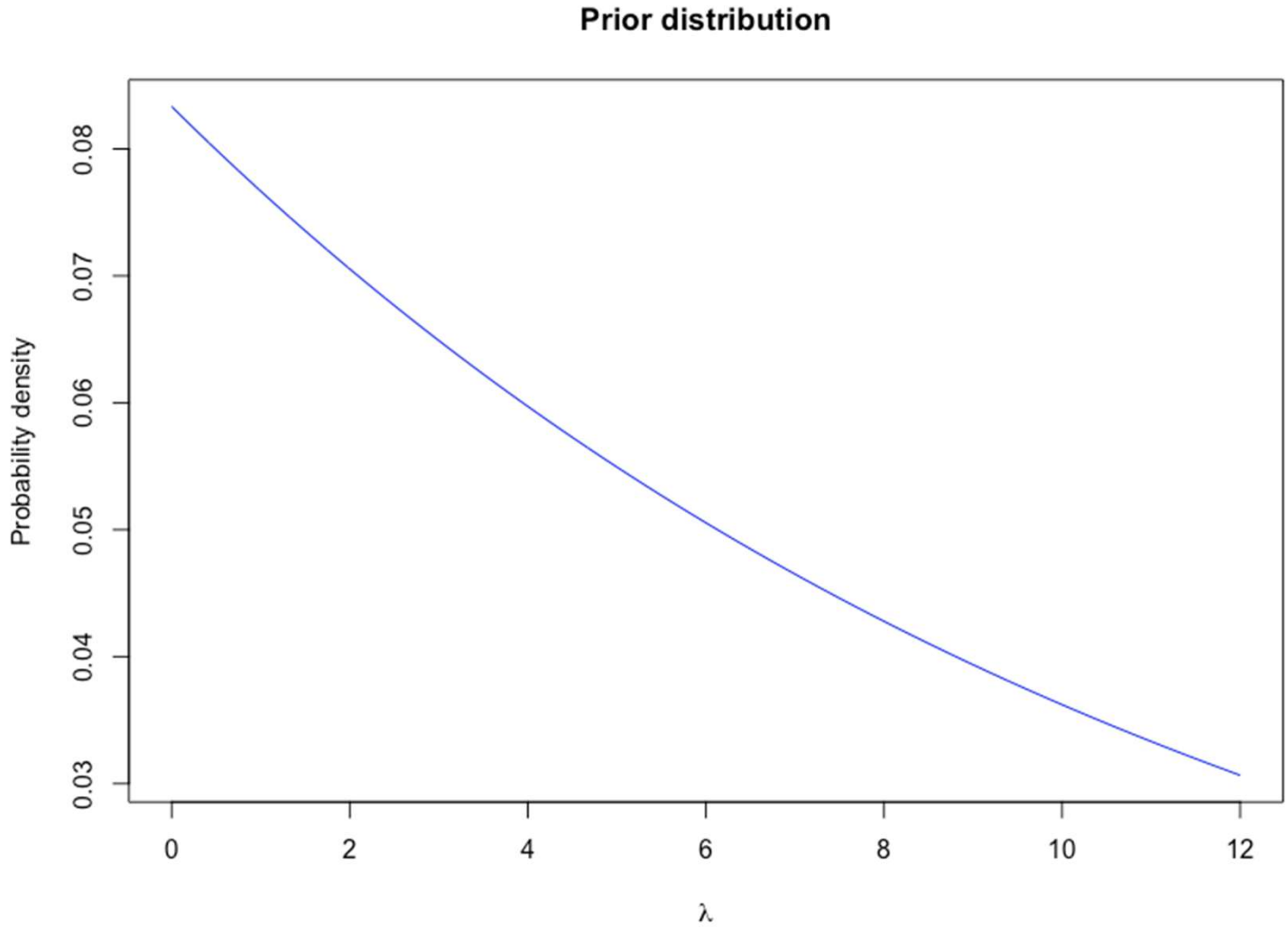
$$\propto \frac{\lambda(\beta+\Delta T_0)^2 \mathrm{e}^{-\lambda(\beta+\Delta T_0)}}{\Gamma(2)} \lambda \mathrm{e}^{\lambda \Delta T_1} = \frac{\lambda^2(\beta+\Delta T_0)^2 \mathrm{e}^{-\lambda(\beta+\Delta T_0+\Delta T_1)}}{\Gamma(2)} \xrightarrow{\text{normalise}} \frac{\lambda^2(\beta+\Delta T_0+\Delta T_1)^3 \mathrm{e}^{-\lambda(\beta+\Delta T_0+\Delta T_1)}}{\Gamma(3)}$$

This is another gamma distribution, with the new parameters $\alpha' = 3$ and $\beta' = \beta + \Delta T_0 + \Delta T_1$

If more observations are made, the posterior distribution is always a gamma, with new parameters.
This is called a *conjugate prior*: where the posterior distribution is the same form as the prior, with new parameters.
Conjugate priors are very useful when making Bayesian calculations.

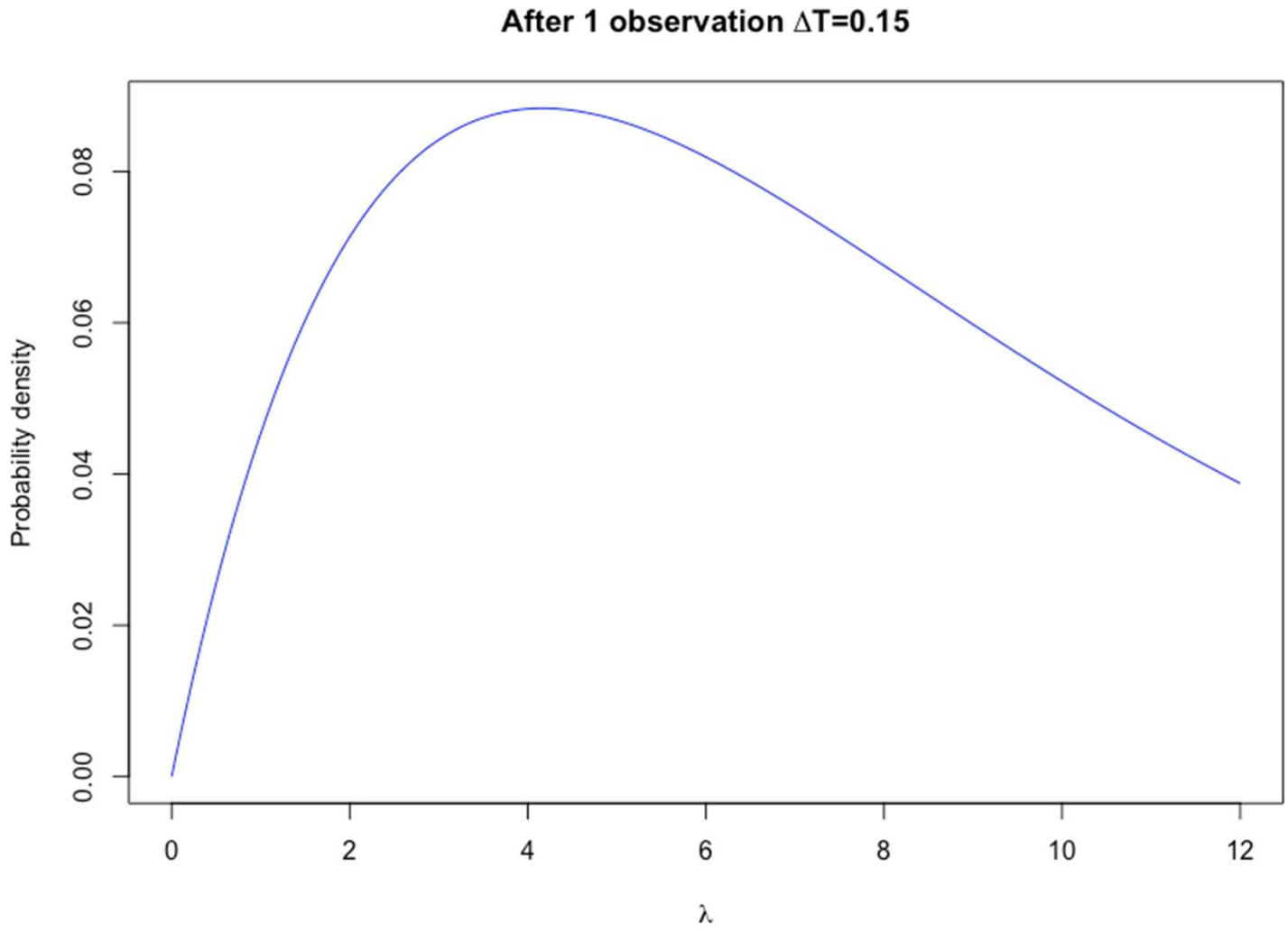
Sequential updates

$\lambda \sim \mathrm{Exp}(\beta), \beta = 12$ hours



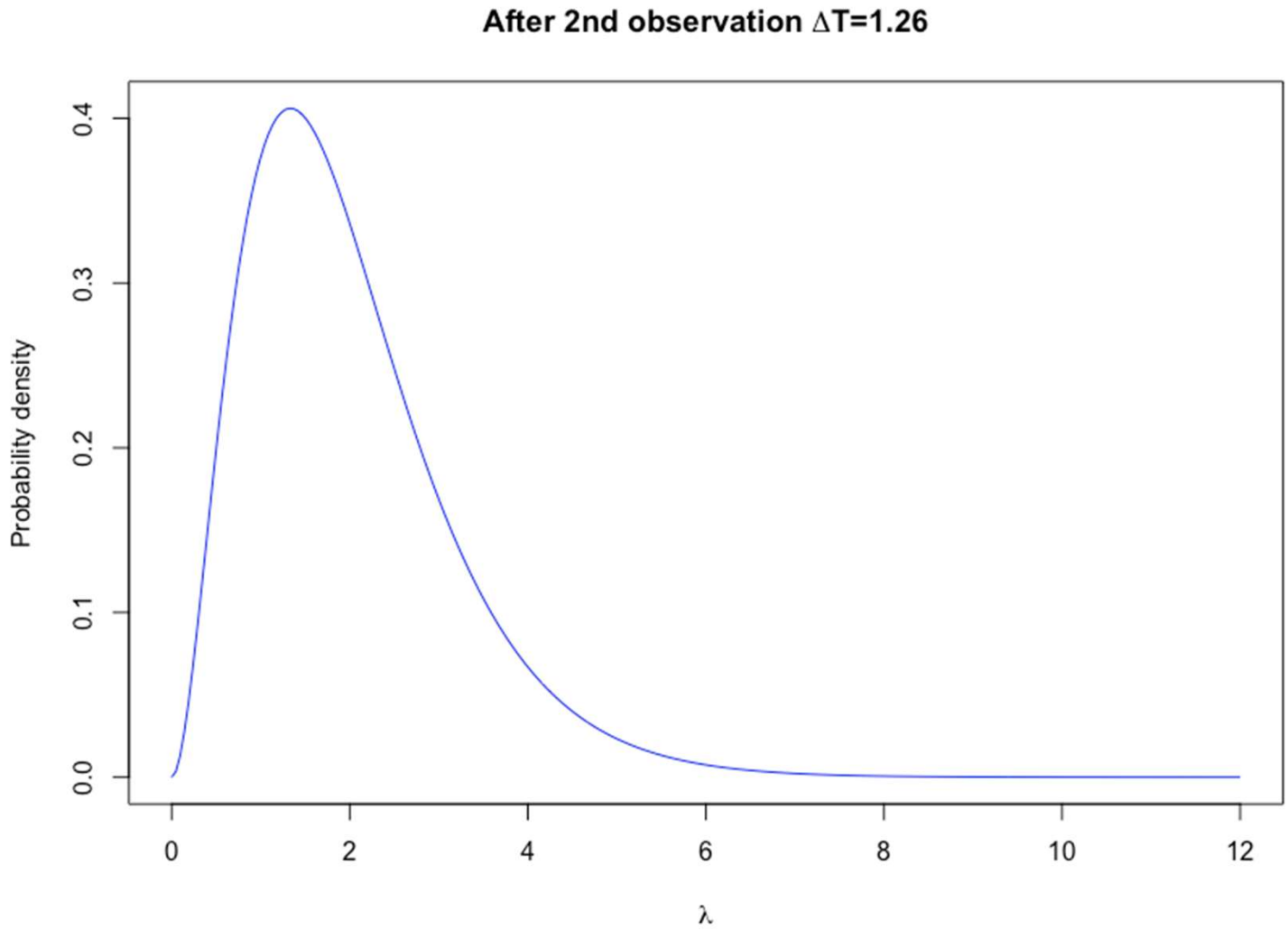
Science and Technology Facilities Council
Hartree Centre

Sequential updates



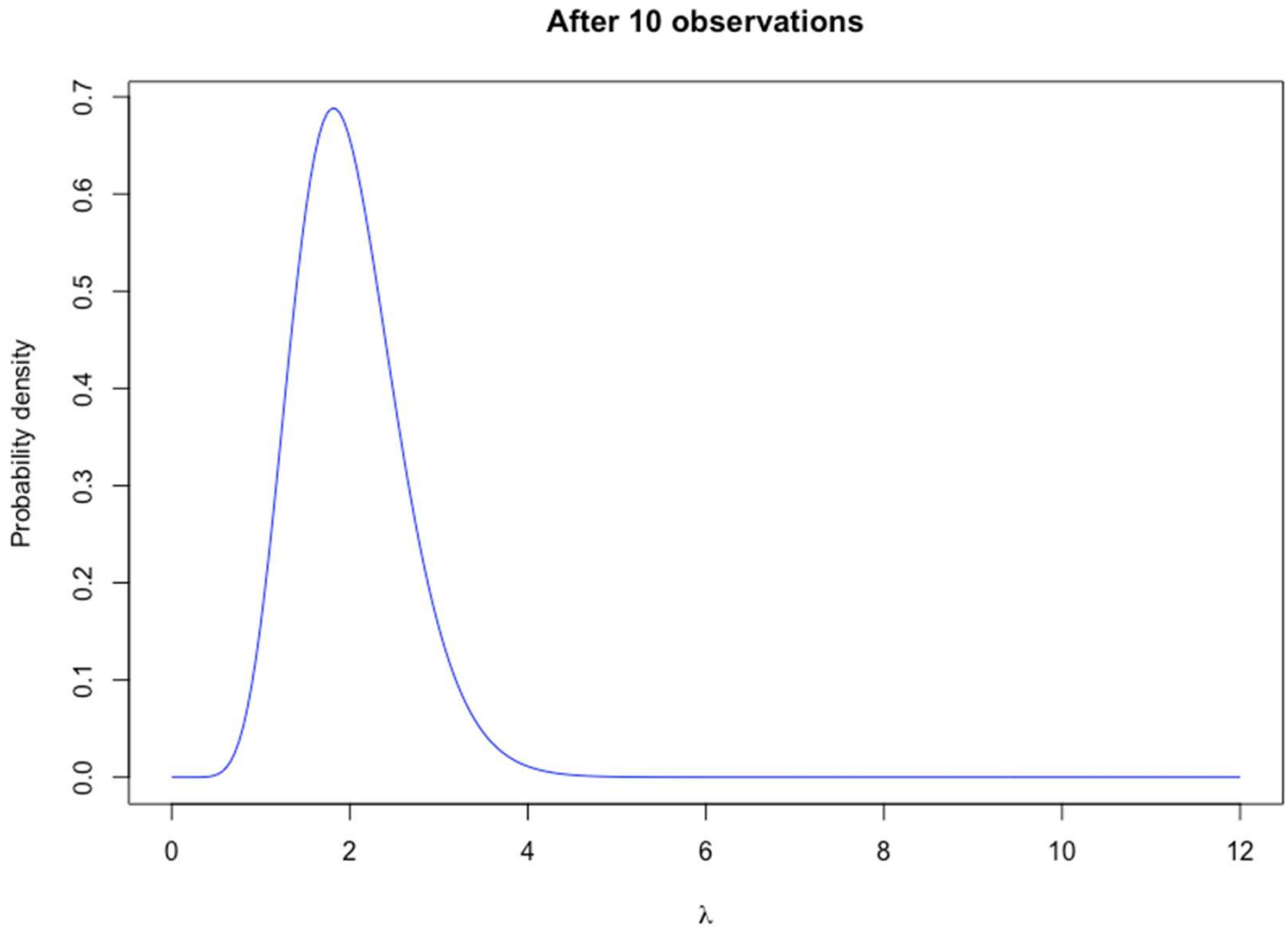
Science and Technology Facilities Council
Hartree Centre

Sequential updates



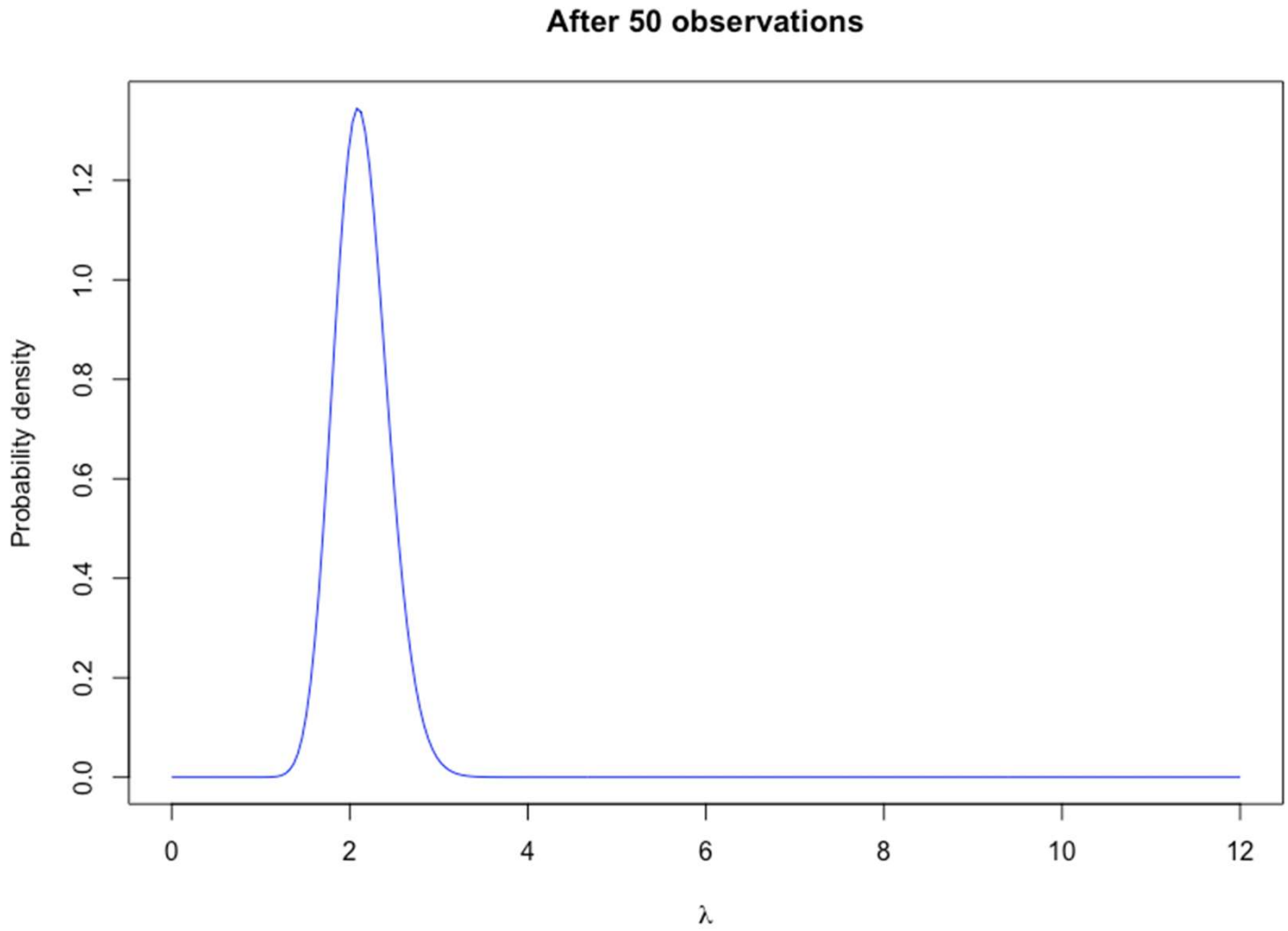
Science and Technology Facilities Council
Hartree Centre

Sequential updates



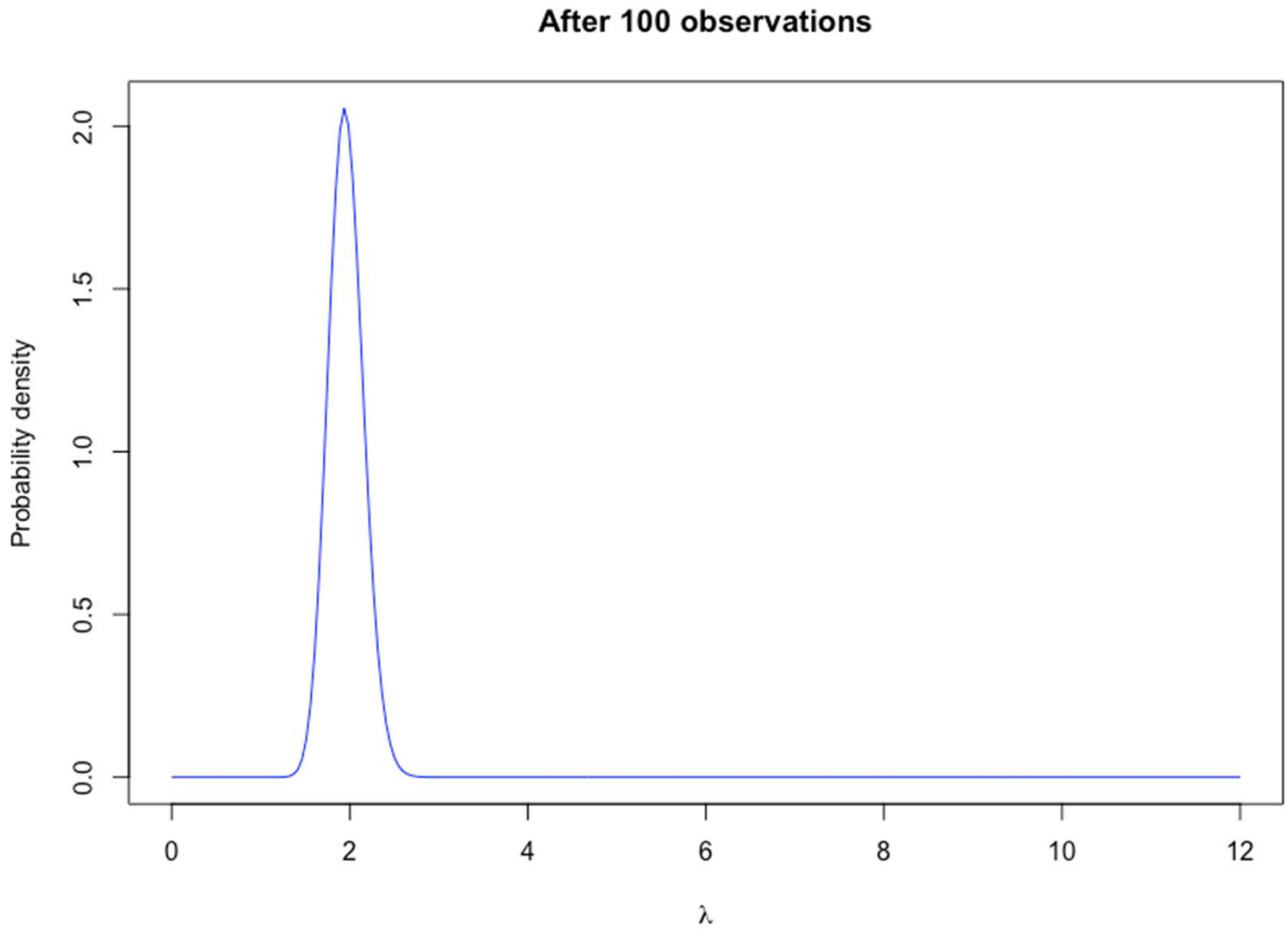
Science and Technology Facilities Council
Hartree Centre

Sequential updates

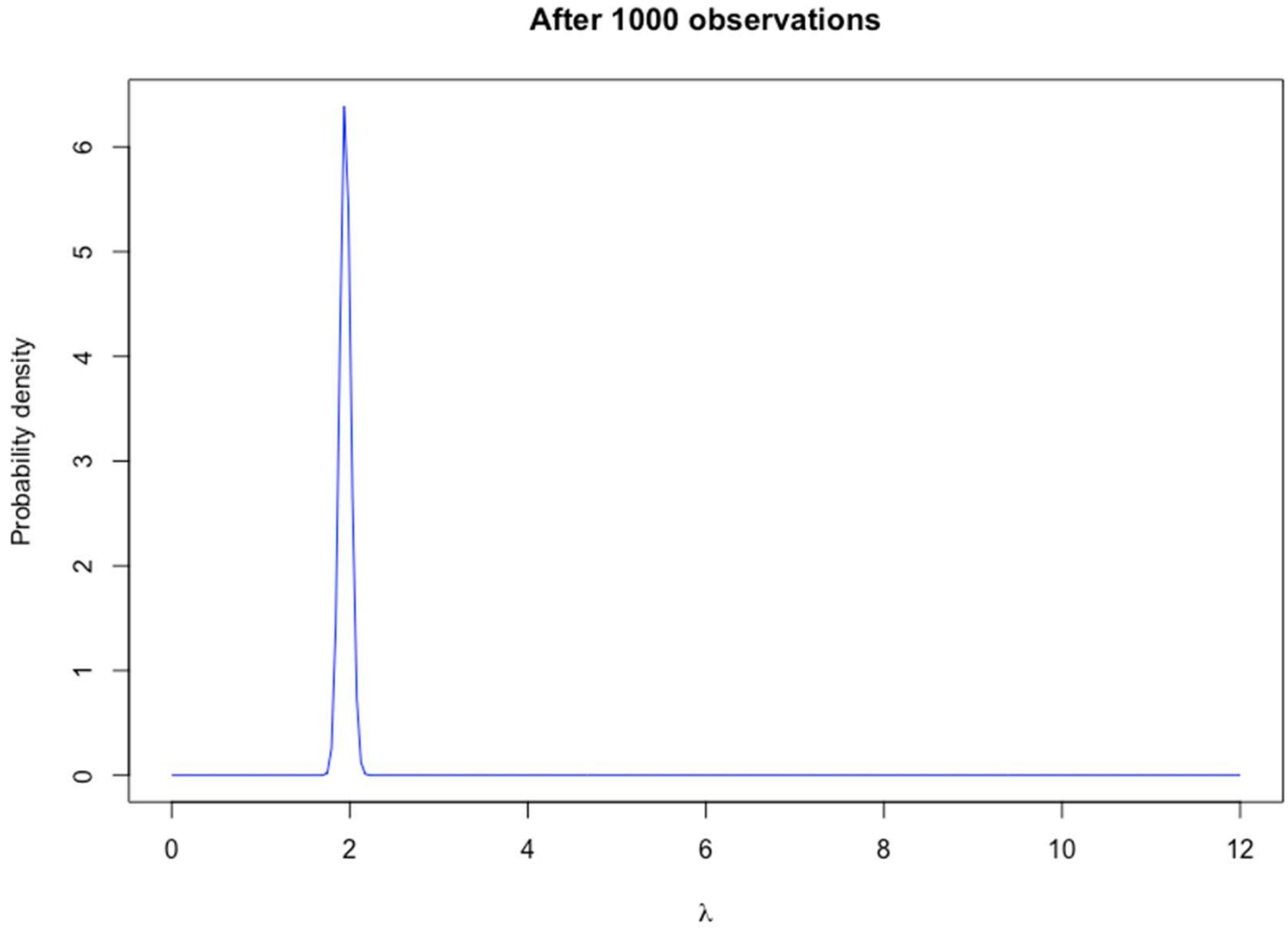


Science and Technology Facilities Council
Hartree Centre

Sequential updates



Sequential updates



Science and Technology Facilities Council

Hartree Centre

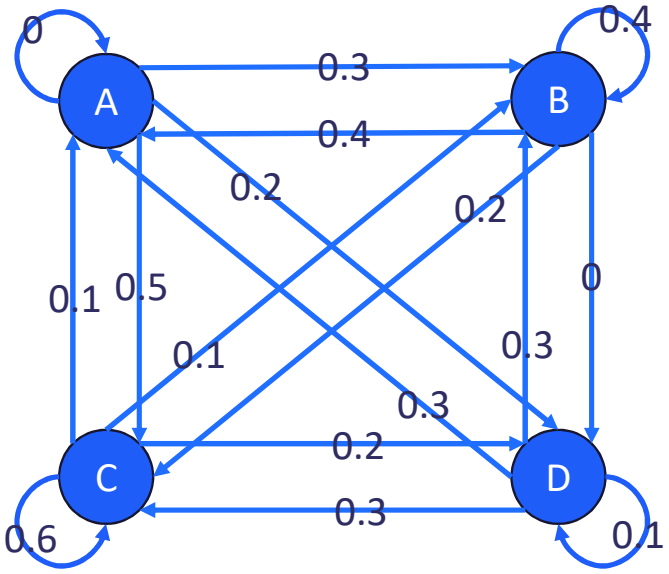
Markov Chains and Bayesian networks

Science and Technology Facilities Council

Hartree Centre

Markov chains

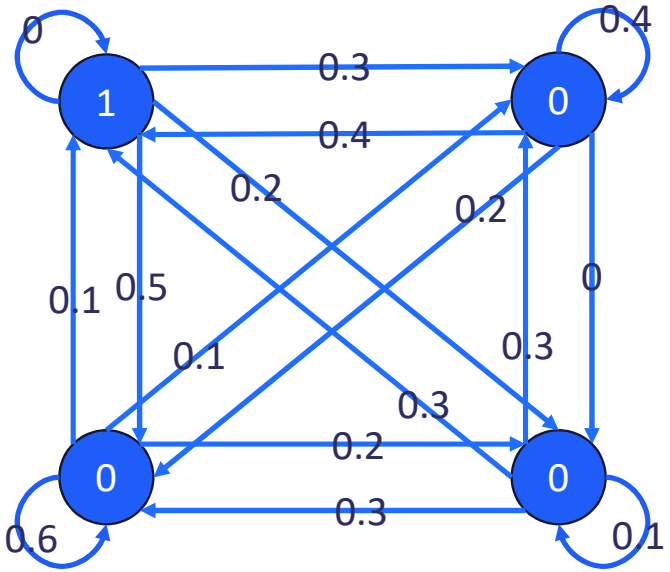
A *Markov chain* is a model of a system which has a set of states.
At each new time step, the state changes according to a fixed set of probabilities called the *transition matrix.*
T_{ij} gives the probability of transitioning to state j from state i.



$$\begin{bmatrix} 0 & 0.3 & 0.5 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix}$$

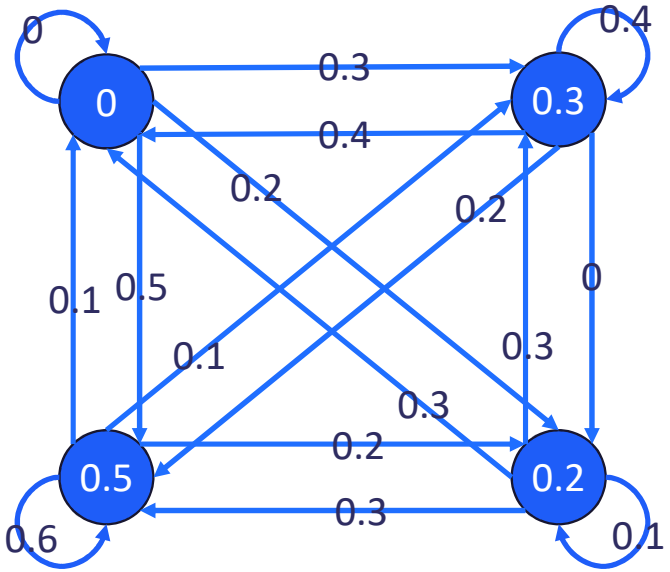
The *Markov property* is that the state only depends on the state of the previous timestep, and not on any earlier states.

Markov chains



$$[1 \quad 0 \quad 0 \quad 0]\begin{bmatrix} 0 & 0.3 & 0.5 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix} = [0 \quad 0.3 \quad 0.5 \quad 0.2]$$

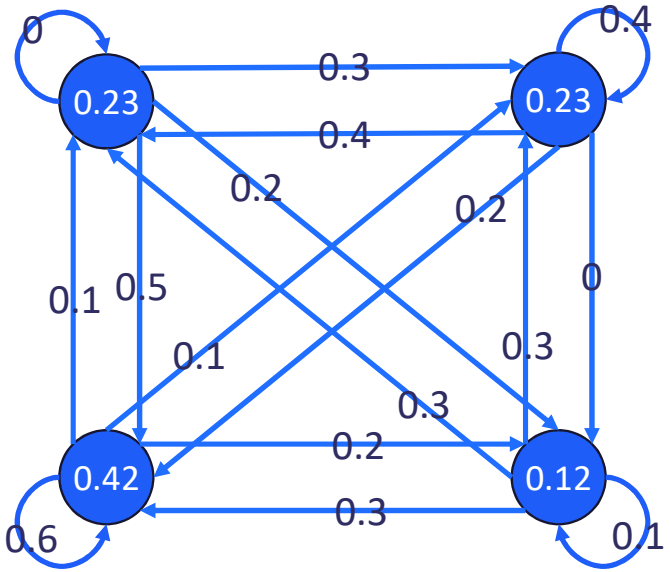
Markov chains



$$[1 \quad 0 \quad 0 \quad 0]\begin{bmatrix} 0 & 0.3 & 0.5 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix} = [0 \quad 0.3 \quad 0.5 \quad 0.2]$$

$$[0 \quad 0.3 \quad 0.5 \quad 0.2]\begin{bmatrix} 0 & 0.3 & 0.5 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix} = [0.23 \quad 0.23 \quad 0.42 \quad 0.12]$$

Markov chains



$$[1 \quad 0 \quad 0 \quad 0]\begin{bmatrix} 0 & 0.3 & 0.5 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix} = [0 \quad 0.3 \quad 0.5 \quad 0.2]$$

$$[0 \quad 0.3 \quad 0.5 \quad 0.2]\begin{bmatrix} 0 & 0.3 & 0.5 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix} = [0.23 \quad 0.23 \quad 0.42 \quad 0.12]$$

Conditional independence

The definition of conditional probability can be extended to larger sets of events. With 3 events

$$\mathrm{P}(A|B \cap C) = \frac{\mathrm{P}(A \cap B \cap C)}{\mathrm{P}(B \cap C)}$$ since $B \cap C$ is just another event in the probability space.

The joint probability can then be decomposed $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A \mid B \cap C)\mathrm{P}(B \cap C)$
then $\mathrm{P}(B \cap C)$ can be decomposed as before so $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A \mid B \cap C)\mathrm{P}(B \mid C)\mathrm{P}(C)$

This process can be repeated for longer chains of events
$\mathrm{P}(A \cap B \cap C \cap D) = \mathrm{P}(A \mid B \cap C \cap D)\mathrm{P}(B \mid C \cap D)\mathrm{P}(C \cap D)\mathrm{P}(D)$

If $\mathrm{P}(A|B \cap C) = \mathrm{P}(A|C)$ then A and B are *conditionally independent* given C.

Then, for example, $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A \mid B \cap C)\mathrm{P}(B \cap C) = \mathrm{P}(A|C)\mathrm{P}(B|C)\mathrm{P}(C)$

Conditional independence

The conditional independence expression can also be written
$\mathrm{P}(A \cap B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C)$

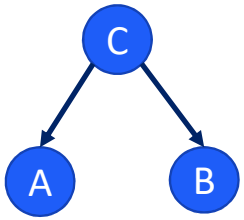
This is another way of obtaining
$\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A \cap B|C)\mathrm{P}(C) = \mathrm{P}(A|C)\mathrm{P}(B|C)\mathrm{P}(C)$

Also note that if B and C are independent,
$\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A \mid B \cap C)\mathrm{P}(B)\mathrm{P}(C)$

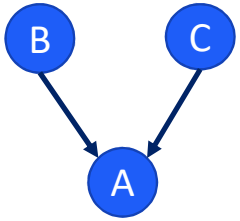
Bayesian networks

These results can also be pictured graphically:

$$\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A \cap B|C)\mathrm{P}(C) = \mathrm{P}(A|C)\mathrm{P}(B|C)\mathrm{P}(C)$$



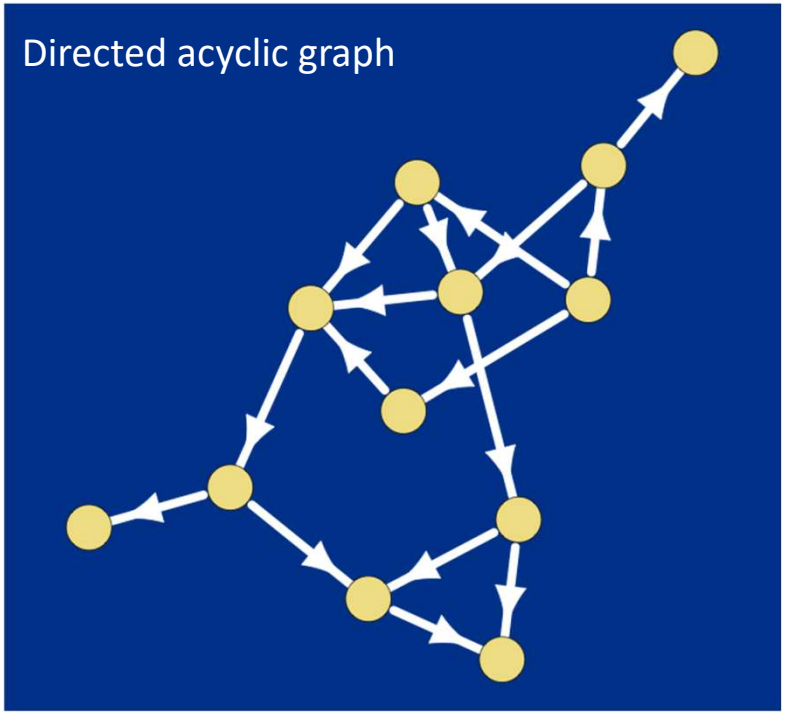
$$\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A \mid B \cap C)\mathrm{P}(B)\mathrm{P}(C)$$



Bayesian networks

This introduces the idea of a *Bayesian network.* This is a way of representing a joint probability distribution.

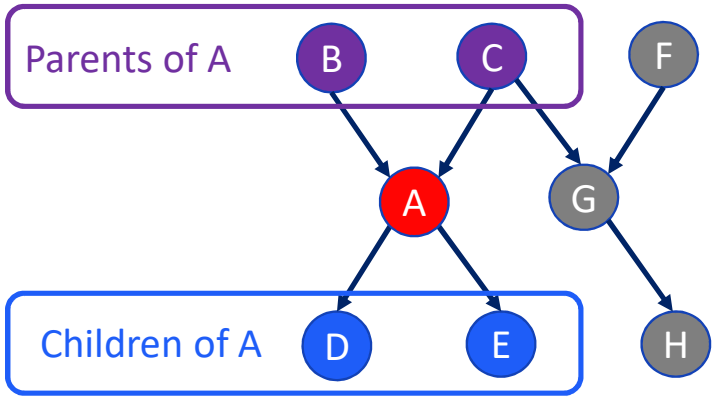
A Bayesian network has two parts: a directed acyclic graph representing the dependencies, and the conditional probability distributions for each dependency.



Bayesian networks

In a Bayesian network, each node is conditionally independent of its non-descendants (not children, etc) given its parents.

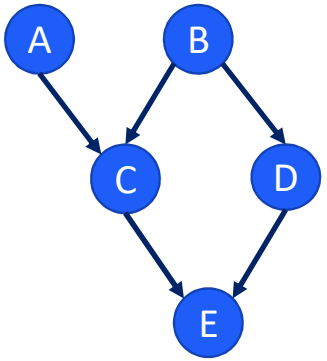
In this network, A's non-descendants are {B, C, F, G, H} so it is conditionally independent of {F, G, H}



Bayesian networks

Joint probability

$$
\begin{aligned}
&\mathrm{P}(A \cap B \cap C \cap D \cap E) \\
&= \mathrm{P}(E \mid A \cap B \cap C \cap D)\mathrm{P}(A \cap B \cap C \cap D) \\
&= \mathrm{P}(E \mid C \cap D)\mathrm{P}(A \cap B \cap C \cap D) \ (E \text{ is conditionally independent of } A \text{ and } B) \\
&= \mathrm{P}(E \mid C \cap D)\mathrm{P}(D \mid A \cap B \cap C)\mathrm{P}(A \cap B \cap C) \\
&= \mathrm{P}(E \mid C \cap D)\mathrm{P}(D \mid B)\mathrm{P}(A \cap B \cap C) \ (D \text{ is conditionally independent of } A \text{ and } C) \\
&= \mathrm{P}(E \mid C \cap D)\mathrm{P}(D \mid B)\mathrm{P}(C \mid A \cap B)\ \mathrm{P}(A \cap B) \ (D \text{ is conditionally independent of } A \text{ and } C) \\
&= \mathrm{P}(E \mid C \cap D)\mathrm{P}(D \mid B)\mathrm{P}(C \mid A \cap B)\ \mathrm{P}(B)\mathrm{P}(A) \ (A \text{ and } B \text{ are independent})
\end{aligned}
$$



Bayesian networks

Joint probability

$$\mathrm{P}(E \mid C \cap D)\mathrm{P}(D \mid B)\mathrm{P}(C \mid A \cap B)\,\mathrm{P}(B)\mathrm{P}(A)$$

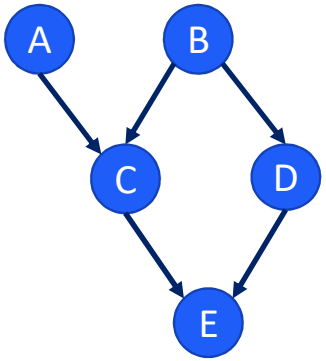
Conditional distributions

$$\mathrm{P}(E = 1 \mid C \cap D) = \begin{cases} 0.2 & C = 0, D = 0 \\ 0.3 & C = 1, D = 0 \\ 0.4 & C = 0, D = 1 \\ 0.1 & C = 1, D = 1 \end{cases}$$

$$\mathrm{P}(B = 1) = 0.6$$
$$\mathrm{P}(A = 1) = 0.9$$

$$\mathrm{P}(D = 1 \mid B) = \begin{cases} 0.6 & B = 0 \\ 0.4 & B = 1 \end{cases}$$

$$\mathrm{P}(C = 1 \mid A \cap B) = \begin{cases} 0.6 & A = 0, B = 0 \\ 0.1 & A = 1, B = 0 \\ 0.2 & A = 0, B = 1 \\ 0.1 & A = 1, B = 1 \end{cases}$$



Science and Technology Facilities Council

Hartree Centre

Kalman filters

Science and Technology Facilities Council

Hartree Centre

Kalman filter – preliminary results

Covariance transformations

If $\mathbf{x}$ is a random vector, its (auto-) covariance matrix is $\mathbf{C} = \mathrm{E}[\mathbf{x}\mathbf{x}^T] - \mathrm{E}[\mathbf{x}]\mathrm{E}[\mathbf{x}^T]$
then under a constant matrix transform $\mathbf{x}' = \mathbf{A}\mathbf{x}$, $\mathbf{C}' = \mathbf{A}\mathbf{C}\mathbf{A}^T$

If two random vectors are uncorrelated, their covariance matrix is $\mathbf{0}$.

Adding Gaussian distributions

If two multivariate normal distributions are multiplied, their product is also a multivariate normal

$$N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \cdot N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = N(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$$

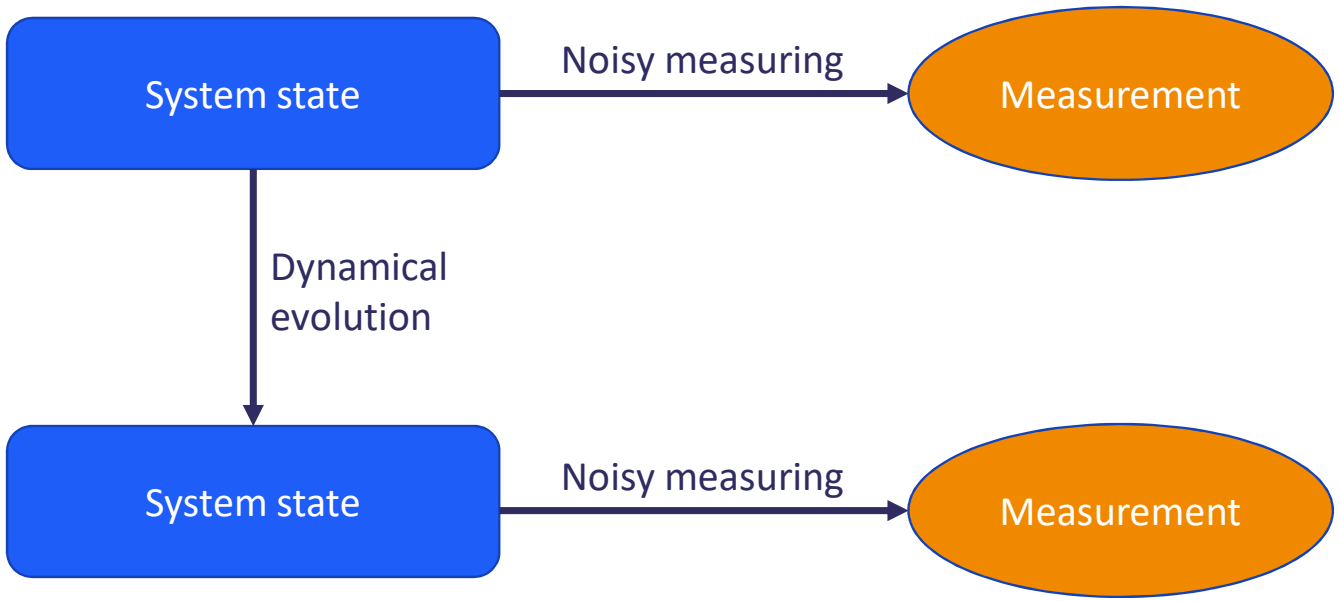
where

$$\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)^{-1}\boldsymbol{\Sigma}_0 \qquad \boldsymbol{\mu}' = \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$$

or if $\mathrm{K} = \boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)^{-1}$, $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}_0 - \mathrm{K}\boldsymbol{\Sigma}_0$, $\boldsymbol{\mu}' = \boldsymbol{\mu}_0 + \mathrm{K}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$

Kalman filter

The Kalman filter is used to model a linear dynamical system with noisy measurements.



Linear dynamical systems

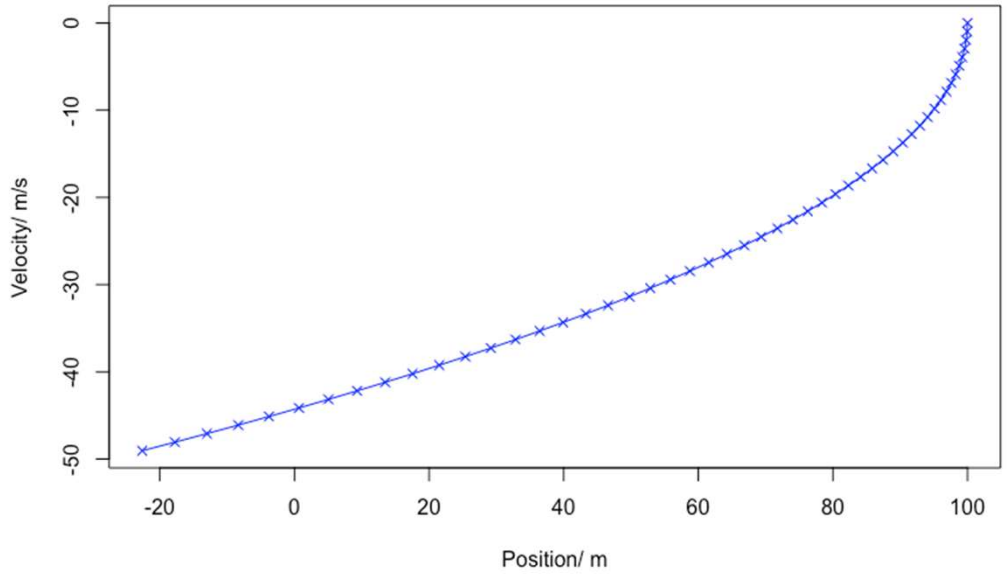
A linear dynamical system is one with a linear update equation

$$\mathbf{x}_{t+1} = \mathbf{F}\mathbf{x}_t \; (+ \mathbf{D}\mathbf{u}_t)$$

(all the systems we consider will use discrete time steps)

For example, a falling tennis ball can be modelled as a linear dynamical system with two variables, the position x and velocity v.

$$\begin{bmatrix} x_{t+1} \\ v_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ v_t \end{bmatrix} - g \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix}$$



Measurements and noise

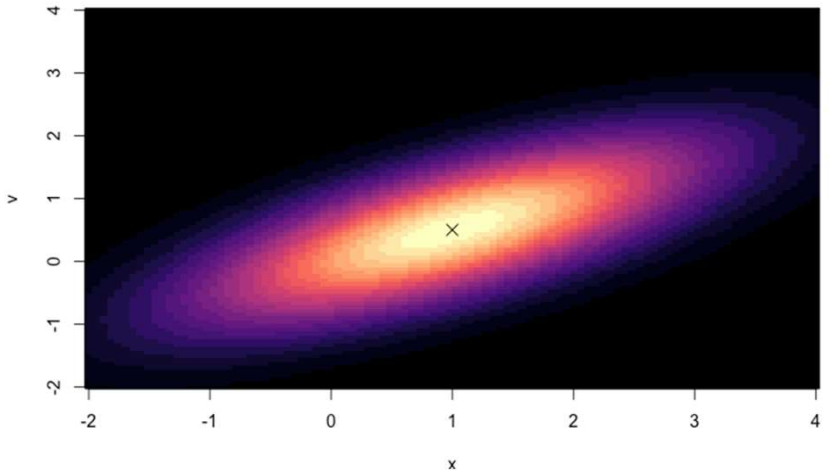
At each time step, a measurement $\mathbf{y}_t$ is taken. This is related to the system state by an observation matrix

$$\mathbf{y}_t = \mathbf{M}\mathbf{x}_t$$

Note that $\mathbf{y}_t$ and $\mathbf{x}_t$ don't have to have the same dimension. For example an object may be moving in three dimensions, but only two could be measurable - like a satellite moving against background stars.

There are two types of noise that affect this process:
Observation noise affects the measurements: $\mathbf{y}_t = \mathbf{M}\mathbf{x}_t + \mathbf{v}_t$
Process noise affects the state transitions: $\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{D}\mathbf{u}_t + \mathbf{w}_t$
(for example random forces due to air movement affecting a thrown ball)

Assume the noise is multivariate Gaussian distributed with mean 0
$\mathbf{v}_t \sim N(\mathbf{0}, \mathbf{R}_t)$, $\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q}_t)$



If the true position is at the cross, noise means that the measurement is drawn from the Gaussian distribution shown

Estimates

As there is noise, the true state of the system isn't known and has to be estimated.

$\hat{\mathbf{x}}_{t|t-1}$ estimates $\mathbf{x}_t$ given the observations made at all steps up to $t-1$
this has an error described by covariance matrix $\mathbf{P}_{t|t-1} = \mathrm{cov}\left[\mathbf{x}_t - \hat{\mathbf{x}}_{t|t-1}\right]$

When the observation $\mathbf{y}_t$ is made, this can be corrected to $\hat{\mathbf{x}}_{t|t}$, the estimate given all observations made up to t

The error in these is described by the covariance matrix $\mathbf{P}_{t|t} = \mathrm{cov}\left[\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}\right]$

Kalman filters – putting it together (1)

End of step $t-1$: have state estimate $\hat{\mathbf{x}}_{t-1|t-1}$ and error covariance estimate $\hat{\mathbf{P}}_{t-1|t-1}$

New inputs

Prediction

predict $\hat{\mathbf{x}}_{t|t-1} = \mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1} + \mathbf{D}_t\mathbf{u}_t$

$\mathbf{F}_t, \mathbf{D}_t, \mathbf{u}_t$

Estimation

estimate $\hat{\mathbf{y}}_t = \mathbf{M}_t\hat{\mathbf{x}}_{t|t-1}$(observation) and $\hat{\mathbf{P}}_{t|t-1} = \mathbf{F}_t\hat{\mathbf{P}}_{t-1|t-1}\mathbf{F}_t^T + \mathbf{Q}_t$ (error)

$\mathbf{M}_t, \mathbf{Q}_t$

Observation

observe $\mathbf{y}_t$ with some error covariance $\mathbf{R}_t$

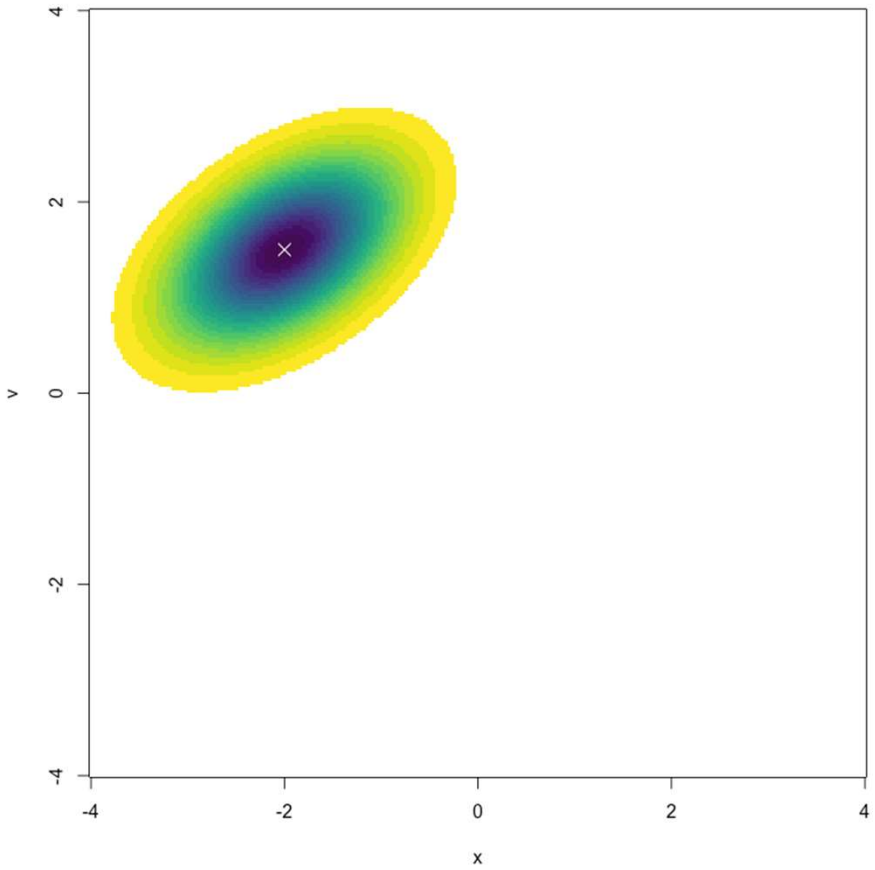
$\mathbf{R}_t$

Update

calculate $\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \hat{\mathbf{y}}_t)$ where $\mathbf{K}_t$ is the *Kalman gain*, and update $\hat{\mathbf{P}}_{t|t}$

Kalman filters

End of step $t-1$: have state estimate $\hat{\mathbf{x}}_{t-1|t-1}$ and error covariance estimate $\hat{\mathbf{P}}_{t-1|t-1}$



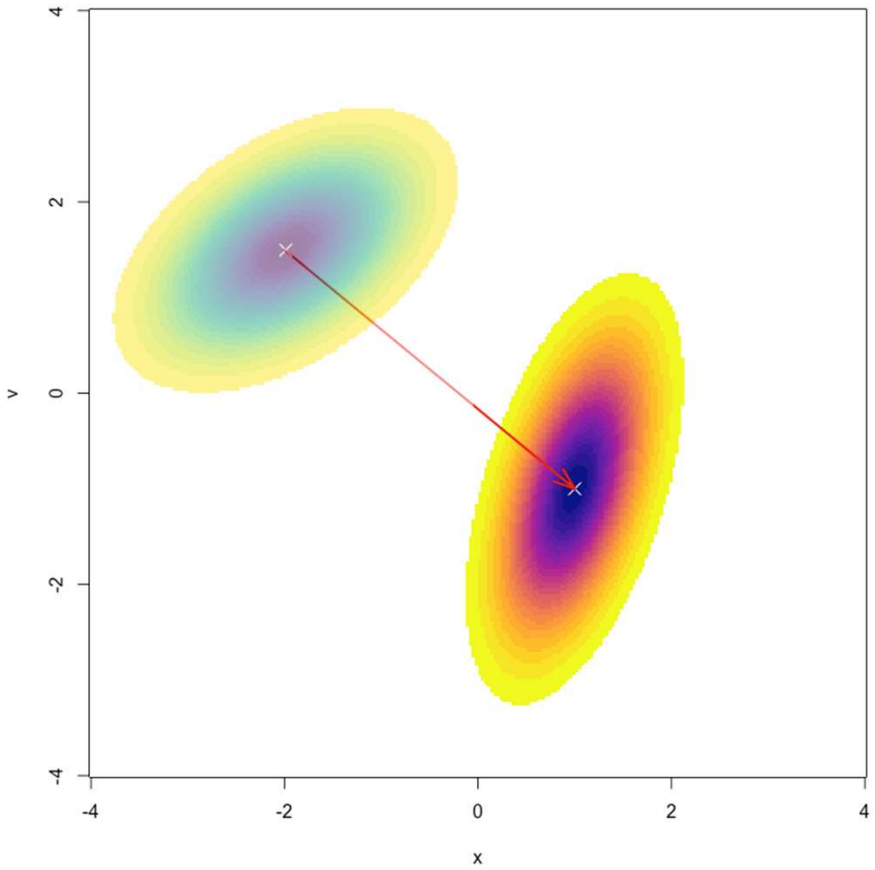
Science and Technology Facilities Council
Hartree Centre

Kalman filters

Prediction and estimation

predict $\hat{\mathbf{x}}_{t|t-1} = \mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1} + \mathbf{D}_t\mathbf{u}_t$ and

estimate $\hat{\mathbf{P}}_{t|t-1} = \mathbf{F}_t\hat{\mathbf{P}}_{t-1|t-1}\mathbf{F}_t^T$



Science and Technology Facilities Council

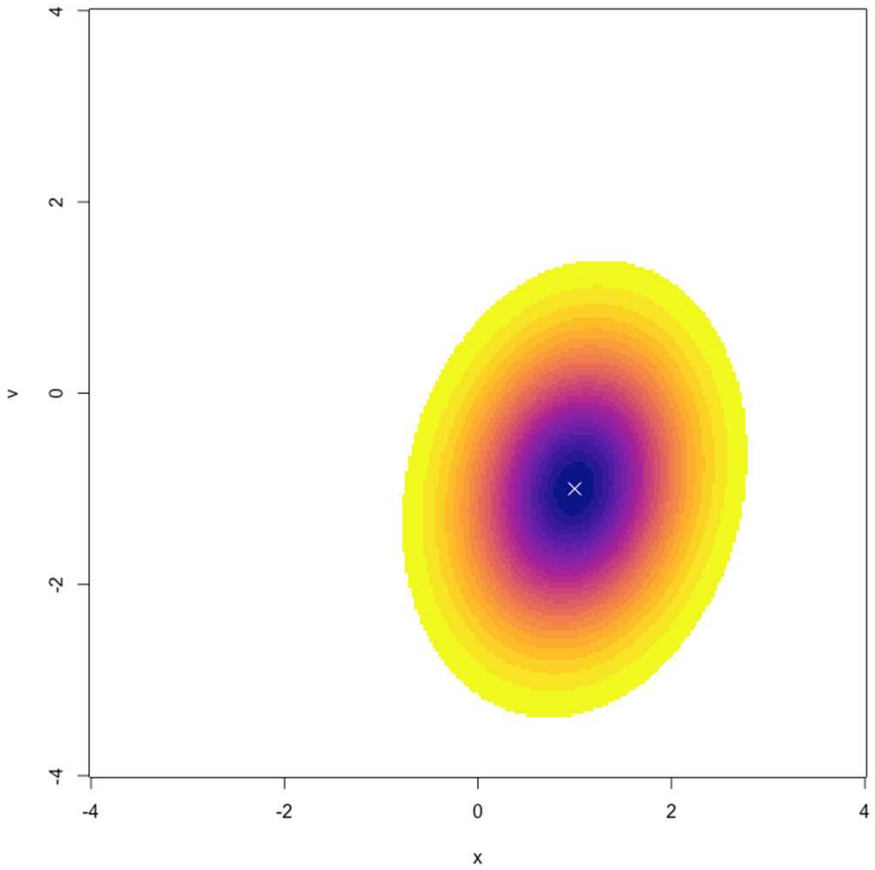
Hartree Centre

Kalman filters

Estimation

add noise

$$\hat{\mathrm{P}}_{t|t-1} = \mathrm{F}_t \hat{\mathrm{P}}_{t-1|t-1} \mathrm{F}_t^T + \mathrm{Q}_t$$

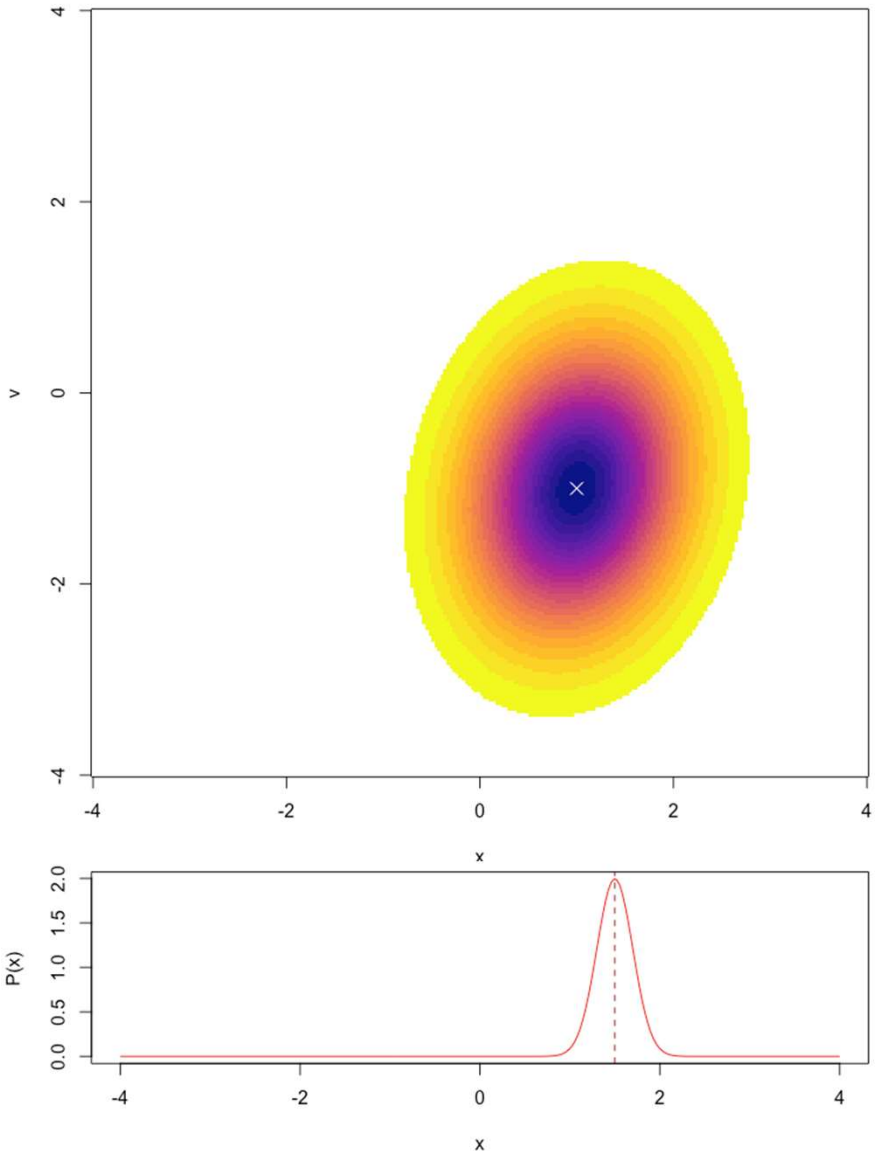


Science and Technology Facilities Council

Hartree Centre

Kalman filters

Observation
observe $\mathbf{y}_t$
with some error covariance $\mathbf{R}_t$



Science and Technology Facilities Council
Hartree Centre

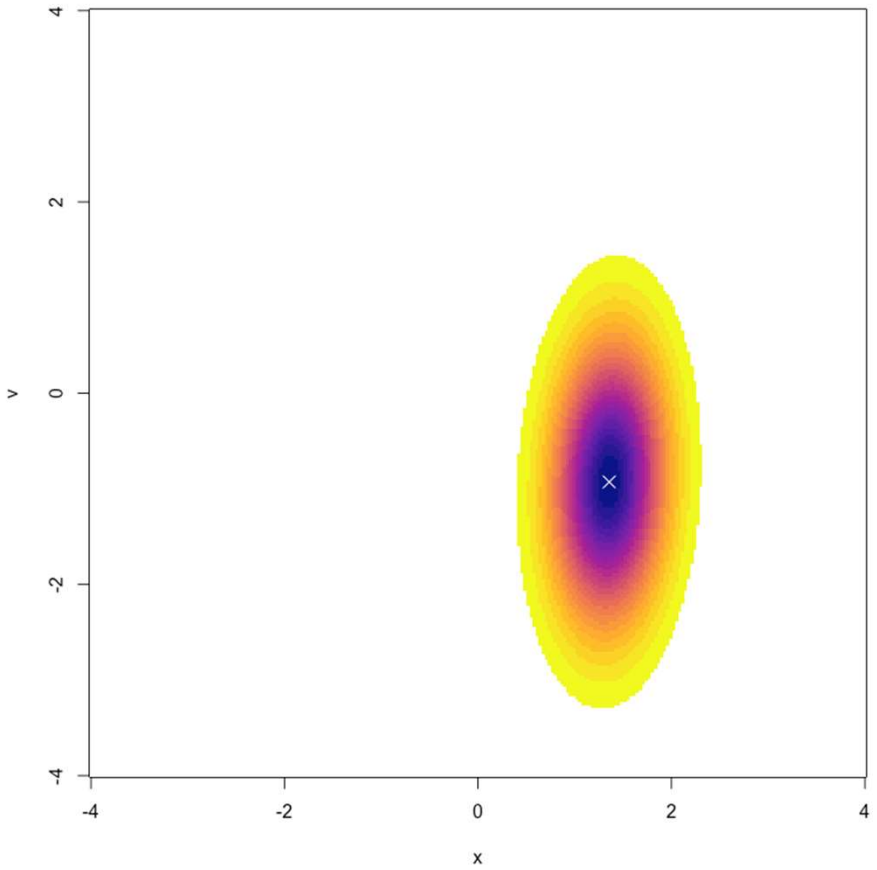
Kalman filters

Update

calculate $\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \hat{\mathbf{y}}_t)$

where $\mathbf{K}_t$ is the *Kalman gain*,

and update $\hat{\mathbf{P}}_{t|t}$



Science and Technology Facilities Council

Hartree Centre

Kalman gain

Kalman gain: $\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \hat{\mathbf{y}}_t)$

Substitute the estimate for $\hat{\mathbf{y}}_t$ and re-arrange: $\hat{\mathbf{x}}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{M}_t)\hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\mathbf{y}_t$

The error in this estimate is $\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}$.

Aim to minimise the total expected mean-squared error $\mathrm{E}\left[\left|\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}\right|^2\right]$

This gives the optimal Kalman gain

Substituting for $\mathbf{y}_t = \mathbf{M}\mathbf{x}_t + \mathbf{v}_t$ and re-arranging, $\mathbf{x}_t - \hat{\mathbf{x}}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{M}_t)(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t-1}) + \mathbf{K}_t\mathbf{v}_t$

This allows the covariance of the error to be estimated

$$\hat{\mathbf{P}}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{M}_t)\hat{\mathbf{P}}_{t|t-1}(\mathbf{I} - \mathbf{K}_t\mathbf{M}_t)^T + \mathbf{K}_t\mathbf{R}_t\mathbf{K}_t^T$$

With this covariance and calculus techniques, the mean squared error is minimised to give the optimal Kalman gain

$$\mathbf{K}_t = \hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T\left(\mathbf{R}_k + \mathbf{M}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T\right)^{-1}$$

$\hat{\mathbf{P}}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{M}_t)\hat{\mathbf{P}}_{t|t-1}$ (this expression is simplified for the optimal Kalman gain only)

Kalman gain: the Bayesian approach

Prior

estimate $\hat{\mathbf{y}}_t = \mathbf{M}_t\hat{\mathbf{x}}_{t|t-1}$(observation) and $\hat{\mathbf{P}}_{t|t-1} = \mathbf{F}_t\hat{\mathbf{P}}_{t-1|t-1}\mathbf{F}_t^T + \mathbf{Q}_t$ (error in $\hat{\mathbf{x}}_{t|t-1}$)

then $\hat{\mathbf{y}}_t \sim N(\mathbf{M}_t\hat{\mathbf{x}}_{t|t-1}, \mathbf{M}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T)$

Observation and likelihood

observe $\mathbf{y}_t$ where $(\mathbf{y}_t - \hat{\mathbf{y}}_t) \sim N(\mathbf{0}, \mathbf{R}_t)$

Posterior

calculate $\mathrm{P}(\hat{\mathbf{y}}_t|\mathbf{y}_t) \propto \mathrm{P}(\mathbf{y}_t \mid \hat{\mathbf{y}}_t)\mathrm{P}(\hat{\mathbf{y}}_t)$

and use this to calculate the probability of $\hat{\mathbf{x}}_{t|t}$

Kalman gain: the Bayesian approach

The posterior distribution is given by $\mathrm{P}(\hat{\mathbf{y}}_t|\mathbf{y}_t) \propto \mathrm{P}(\mathbf{y}_t \mid \hat{\mathbf{y}}_t)\mathrm{P}(\hat{\mathbf{y}}_t)$

As functions of $\hat{\mathbf{y}}_t$, $\mathrm{P}(\mathbf{y}_t \mid \hat{\mathbf{y}}_t) = N(\mathbf{y}_t, \mathbf{R}_t)$ and $\mathrm{P}(\hat{\mathbf{y}}_t) = N(\mathbf{M}_t\hat{\mathbf{x}}_{t|t-1}, \mathbf{M}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T)$

so using $\mathrm{K} = \boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)^{-1}$, $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}_0 - \mathrm{K}\boldsymbol{\Sigma}_0$, $\boldsymbol{\mu}' = \boldsymbol{\mu}_0 + \mathrm{K}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$

$\mathrm{P}(\hat{\mathbf{y}}_t|\mathbf{y}_t) = N(\mathbf{M}_t\hat{\mathbf{x}}_{t|t-1} + \mathrm{K}(\mathbf{y}_t - \mathbf{M}_t\hat{\mathbf{x}}_{t|t-1}), \mathbf{M}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T - \mathrm{K}\mathbf{M}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T)$

with $\mathrm{K} = \mathbf{M}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T\left(\mathbf{M}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T + \mathbf{R}_t\right)^{-1}$

Removing some of the $\mathbf{M}_t$ transformations to convert to estimates for $\hat{\mathbf{x}}_{t|t}$, this produces the same equations as before, with the new mean and covariance estimates.

Kalman filters – putting it together (2)

End of step $t-1$: have state estimate $\hat{\mathbf{x}}_{t-1|t-1}$ and error covariance estimate $\hat{\mathbf{P}}_{t-1|t-1}$

Prediction

predict $\hat{\mathbf{x}}_{t|t-1} = \mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1} + \mathbf{D}_t\mathbf{u}_t$

$\mathbf{F}_t, \mathbf{D}_t, \mathbf{u}_t$

Estimation

estimate $\hat{\mathbf{y}}_t = \mathbf{M}_t\hat{\mathbf{x}}_{t|t-1}$(observation) and $\hat{\mathbf{P}}_{t|t-1} = \mathbf{F}_t\hat{\mathbf{P}}_{t-1|t-1}\mathbf{F}_t^T + \mathbf{Q}_t$ (error)

$\mathbf{M}_t, \mathbf{Q}_t$

Observation

observe $\mathbf{y}_t$ with some error covariance $\mathbf{R}_t$

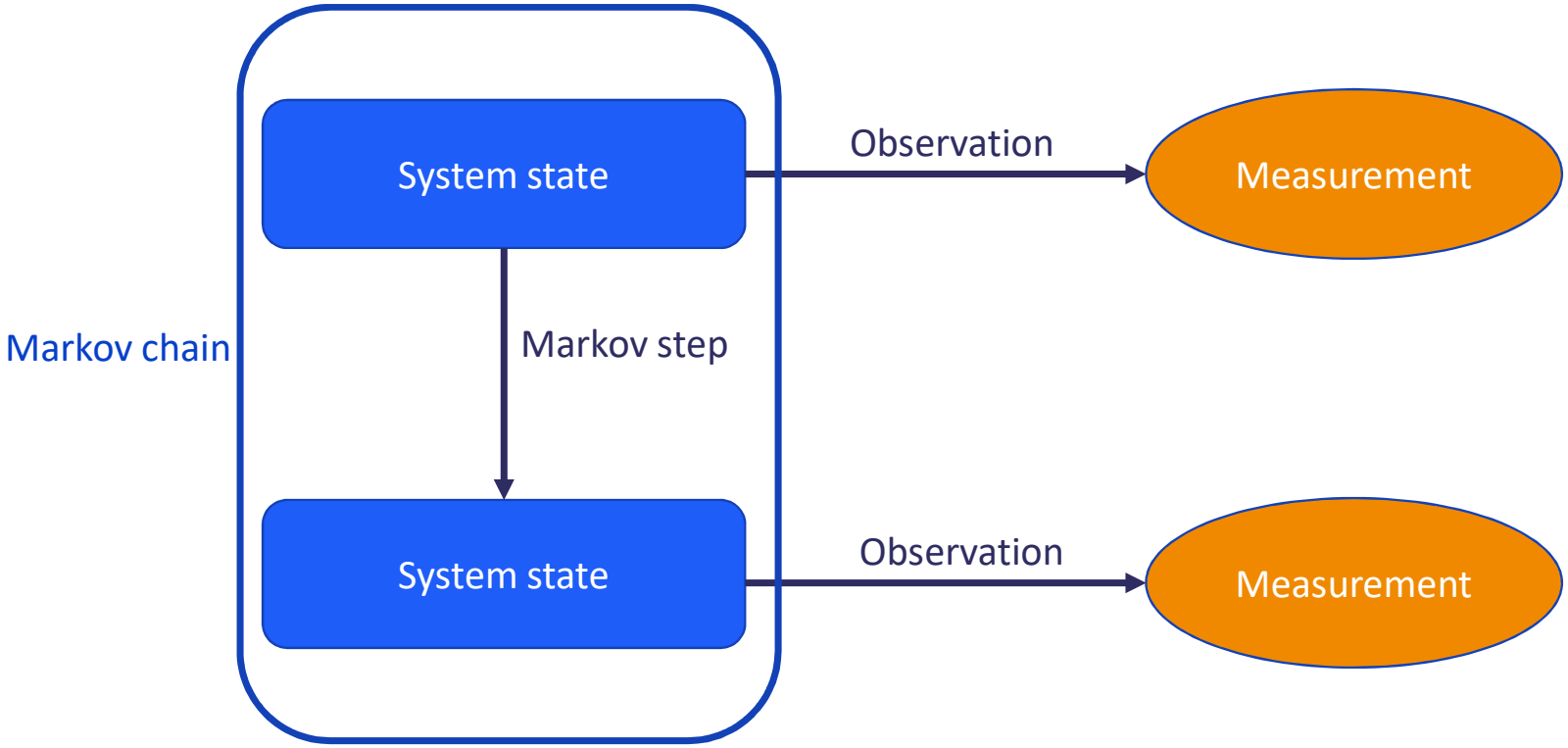
$\mathbf{R}_t$

Update

calculate $\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \hat{\mathbf{y}}_t)$ and $\hat{\mathbf{P}}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{M}_t)\hat{\mathbf{P}}_{t|t-1}$

where $\mathbf{K}_t = \hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T\left(\mathbf{R}_k + \mathbf{M}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{M}_t^T\right)^{-1}$

Hidden Markov models



Posterior distributions

In Bayes' theorem $\mathrm{P}(A|B) = \dfrac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}$ it is often easy to write down the likelihood $\mathrm{P}(B|A)\mathrm{P}(A)$ but intractable to calculate the normalisation $\mathrm{P}(B)$ as this requires summing over all possible values of A.

Markov Chain Monte Carlo is a way around this problem.

- Certain Markov chains have a stationary distribution – a probability distribution which doesn't change with the update step.
- Algorithms have been developed (e.g. the Metropolis-Hastings algorithm) to construct a Markov chain with a stationary distribution $\mathrm{P}(A|B)$ using only $\mathrm{P}(B|A)\mathrm{P}(A)$ – only proportionality is needed.
- Run the Markov chain for many steps until it should have reached its stationary distribution (burn-in)
- Now the results of continuing to run the Markov chain are samples from the desired distribution.

This can be computationally intensive.

Science and Technology Facilities Council

Hartree Centre

Questions?

Science and Technology Facilities Council

Hartree Centre

Thank You

Dr Simon Goodchild
simon.goodchild@stfc.ac.uk
sggoodc@liv.ac.uk

Science and Technology Facilities Council

Hartree Centre