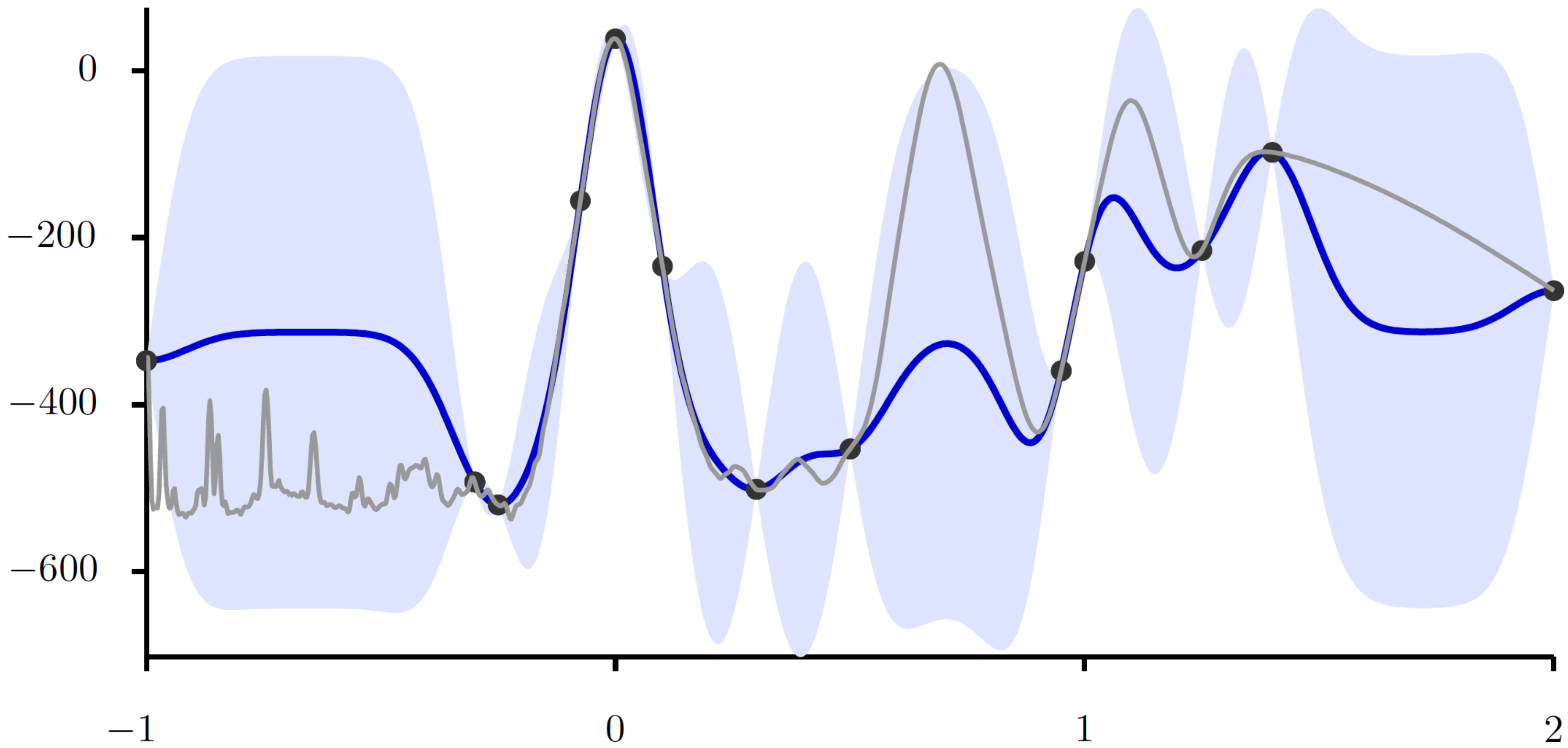
Big model configuration with Bayesian quadrature

David Duvenaud,
Roman Garnett,
Tom Gunter,
Philipp Hennig,
Michael A Osborne
and
Stephen Roberts.

UNIVERSITY OF OXFORD

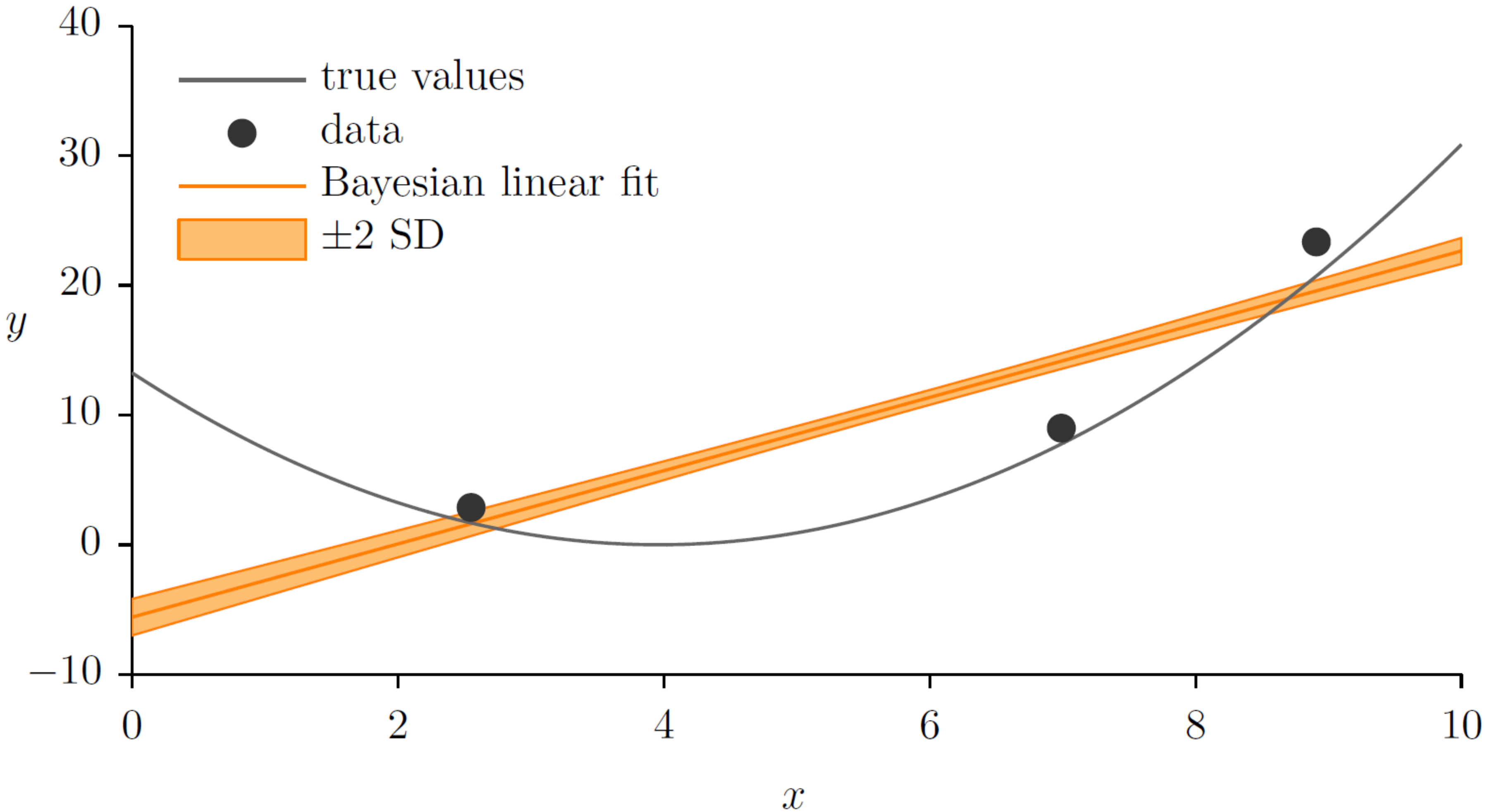
This talk will develop Bayesian quadrature approaches to building ensembles of models for big and complex data.



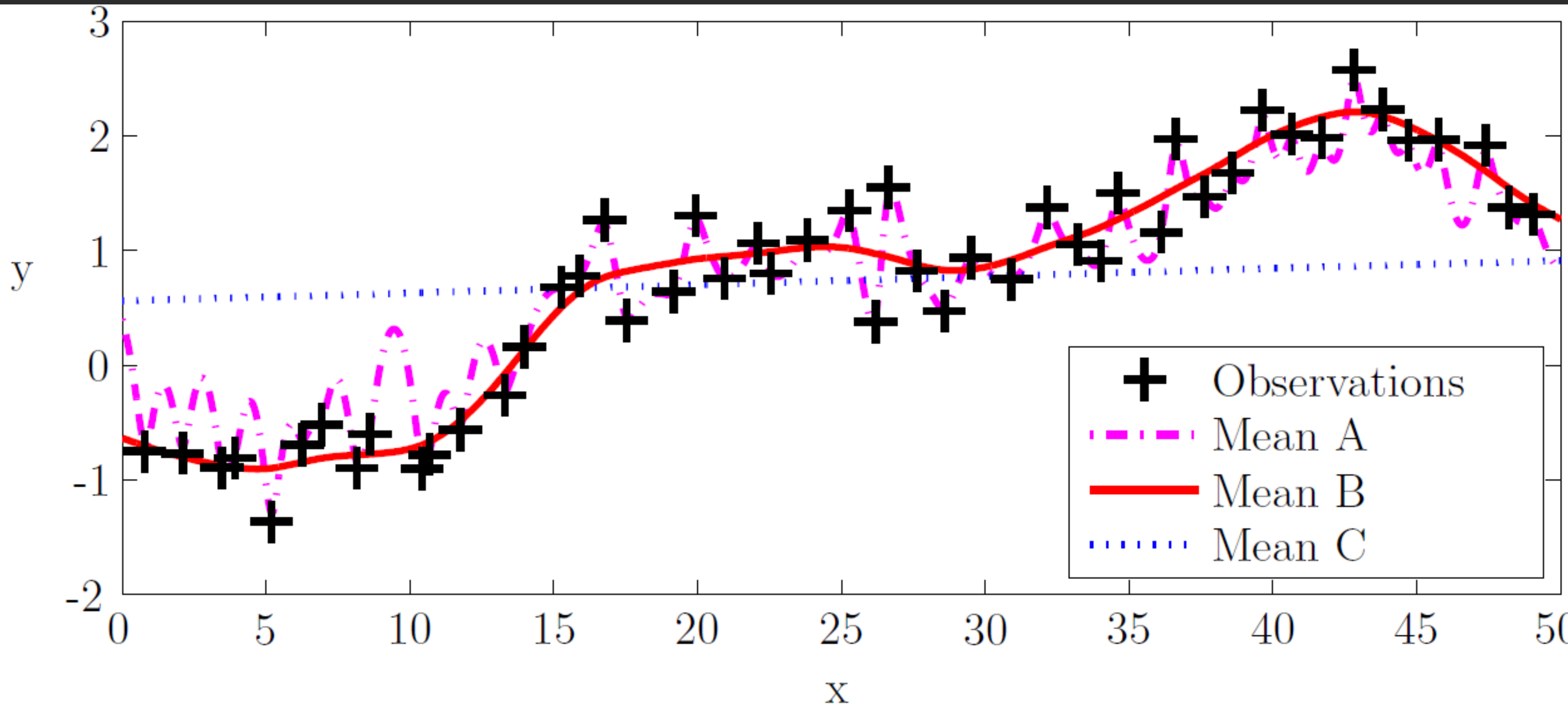
UNIVERSITY OF OXFORD

For big and complex data, it is difficult to pick the right model parameters.

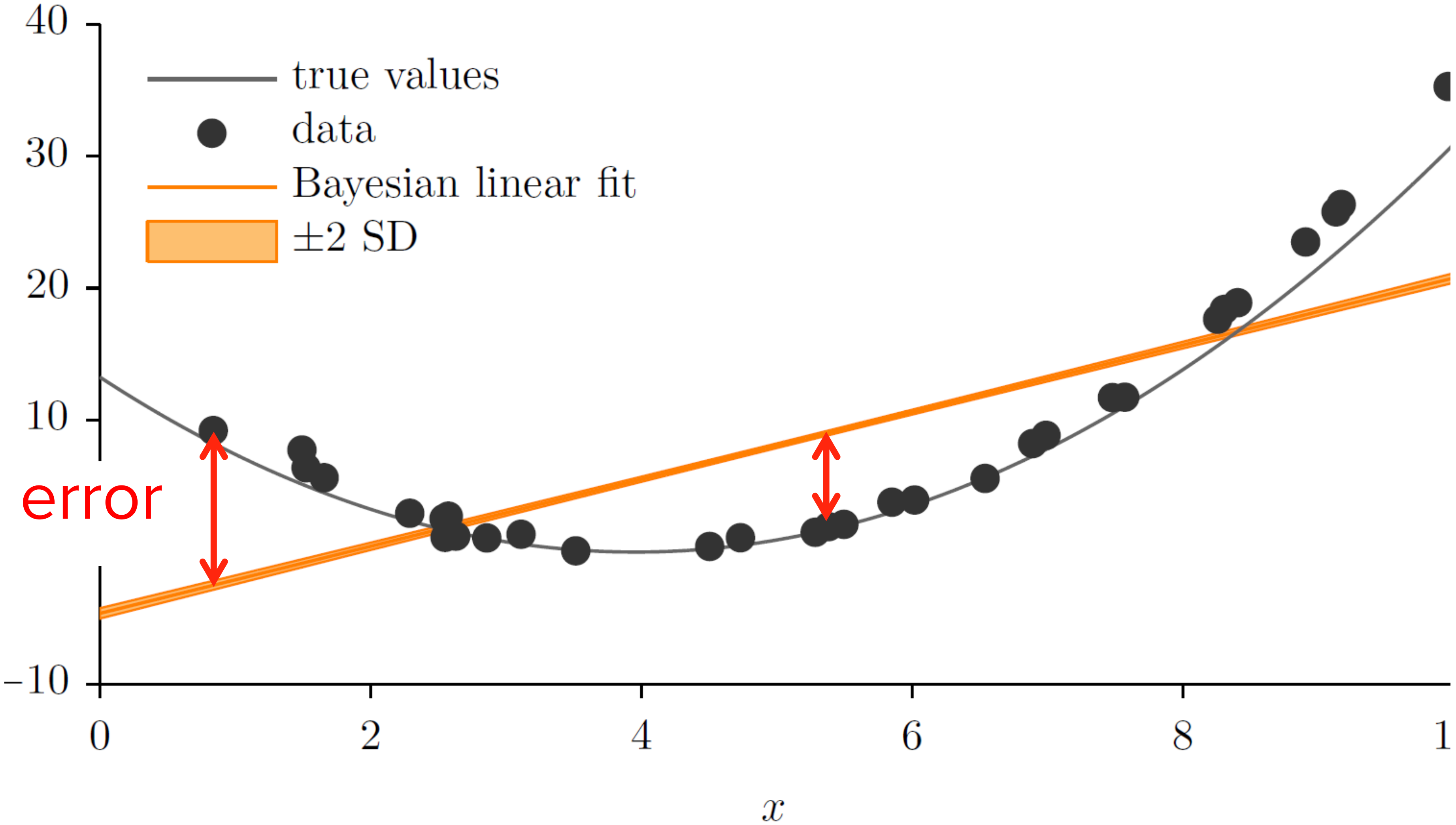
Modelling data requires fitting parameters, such as the *a* and *b* of $y = a\,x + b$.



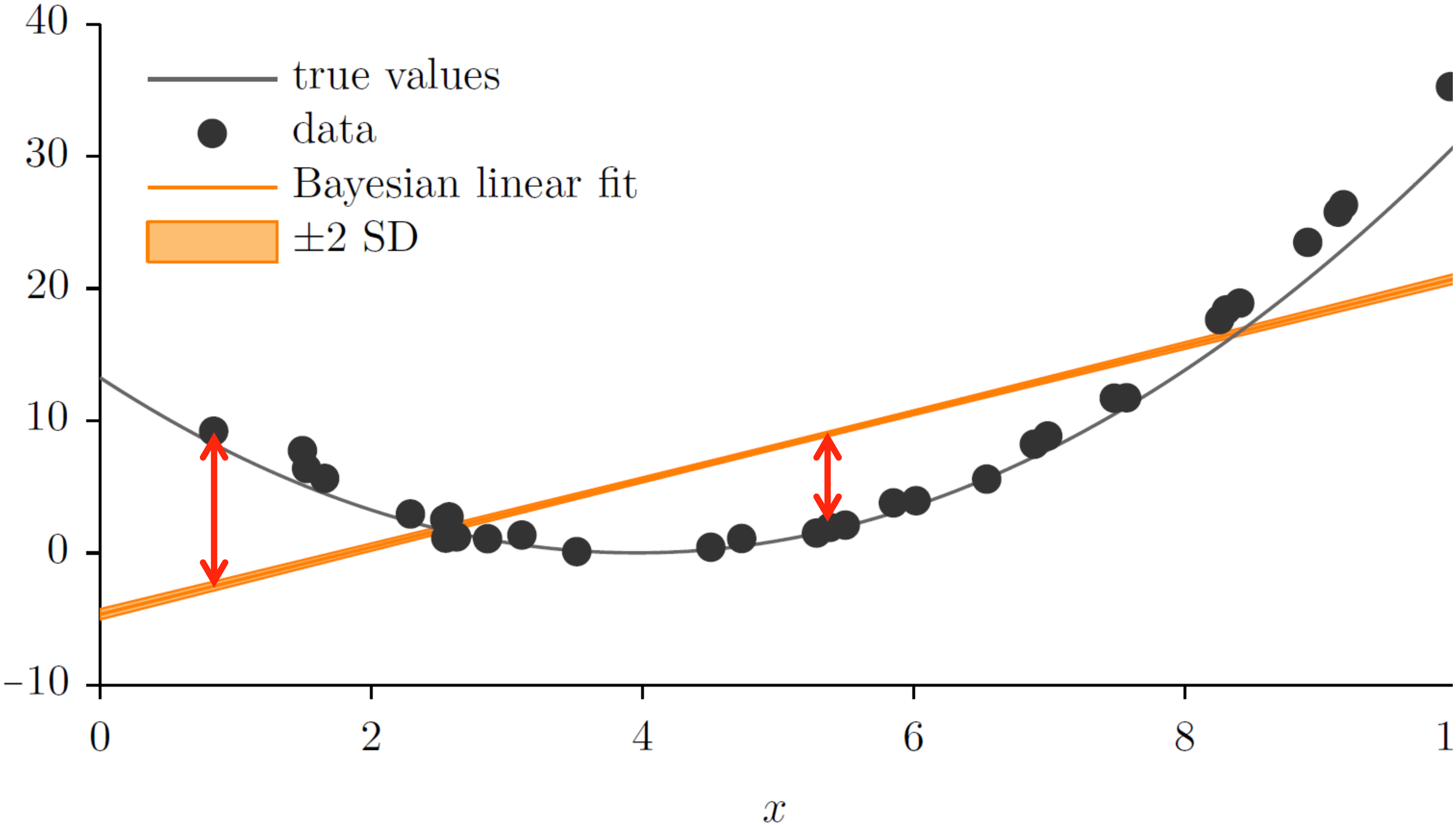
The performance of 'non-parametric' models is sensitive to the selection of hyperparameters.



Evaluating the quality of model fit (or the likelihood) is expensive for big data.

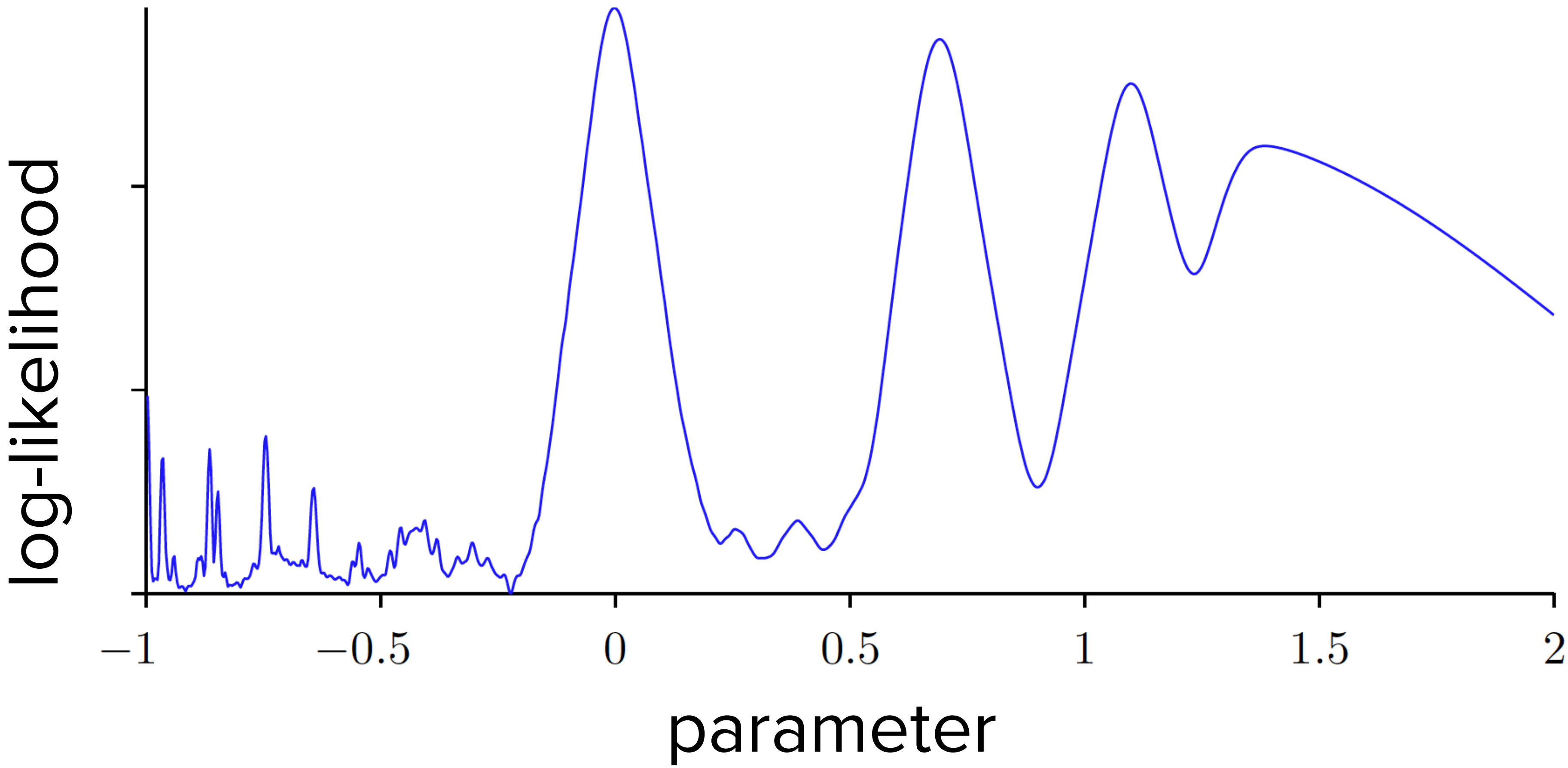


Assuming correlated errors, computation cost will be supralinear in the number of data.



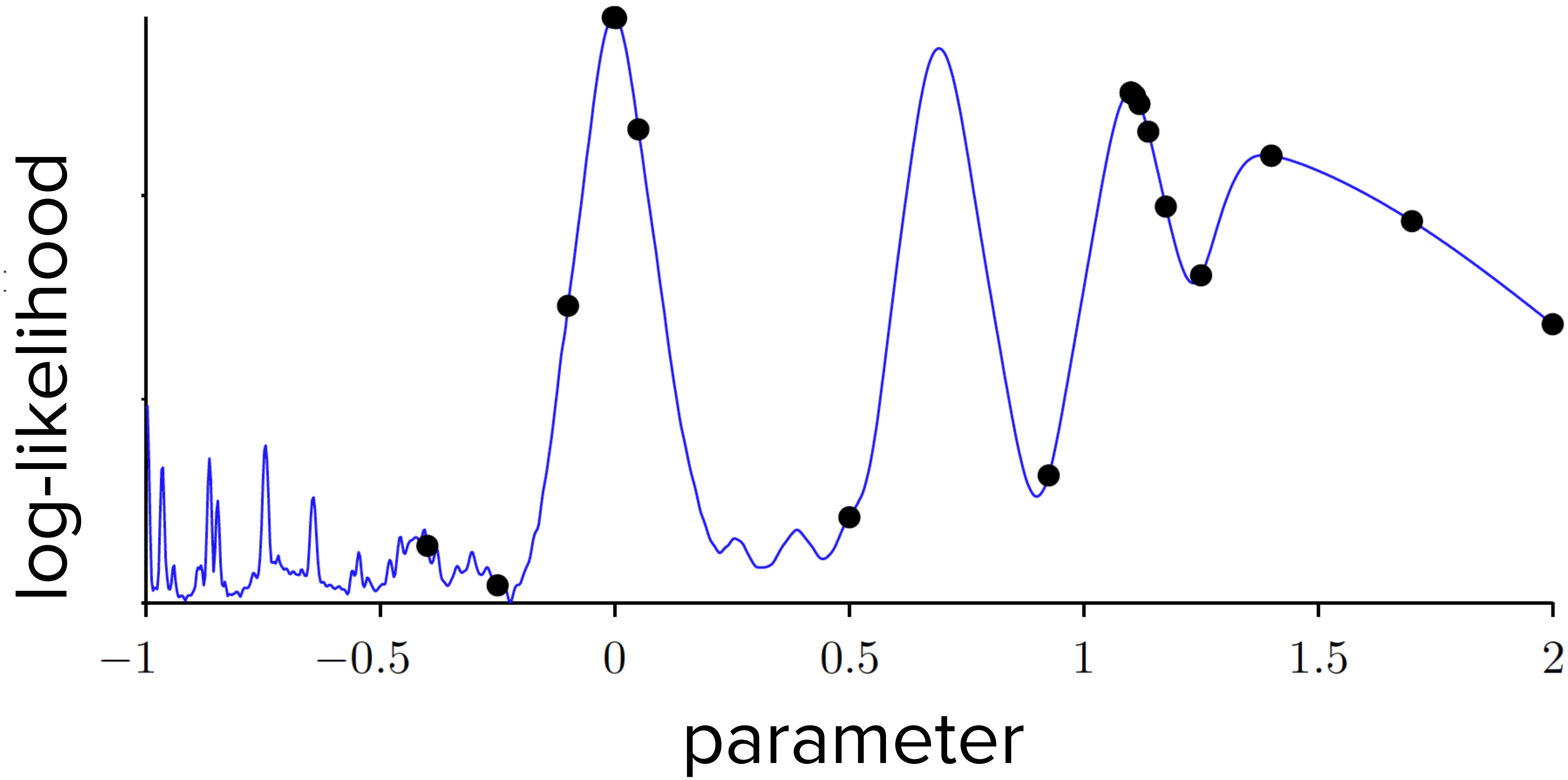
UNIVERSITY OF OXFORD

Complex models and real data often lead to multi-modal likelihood functions.

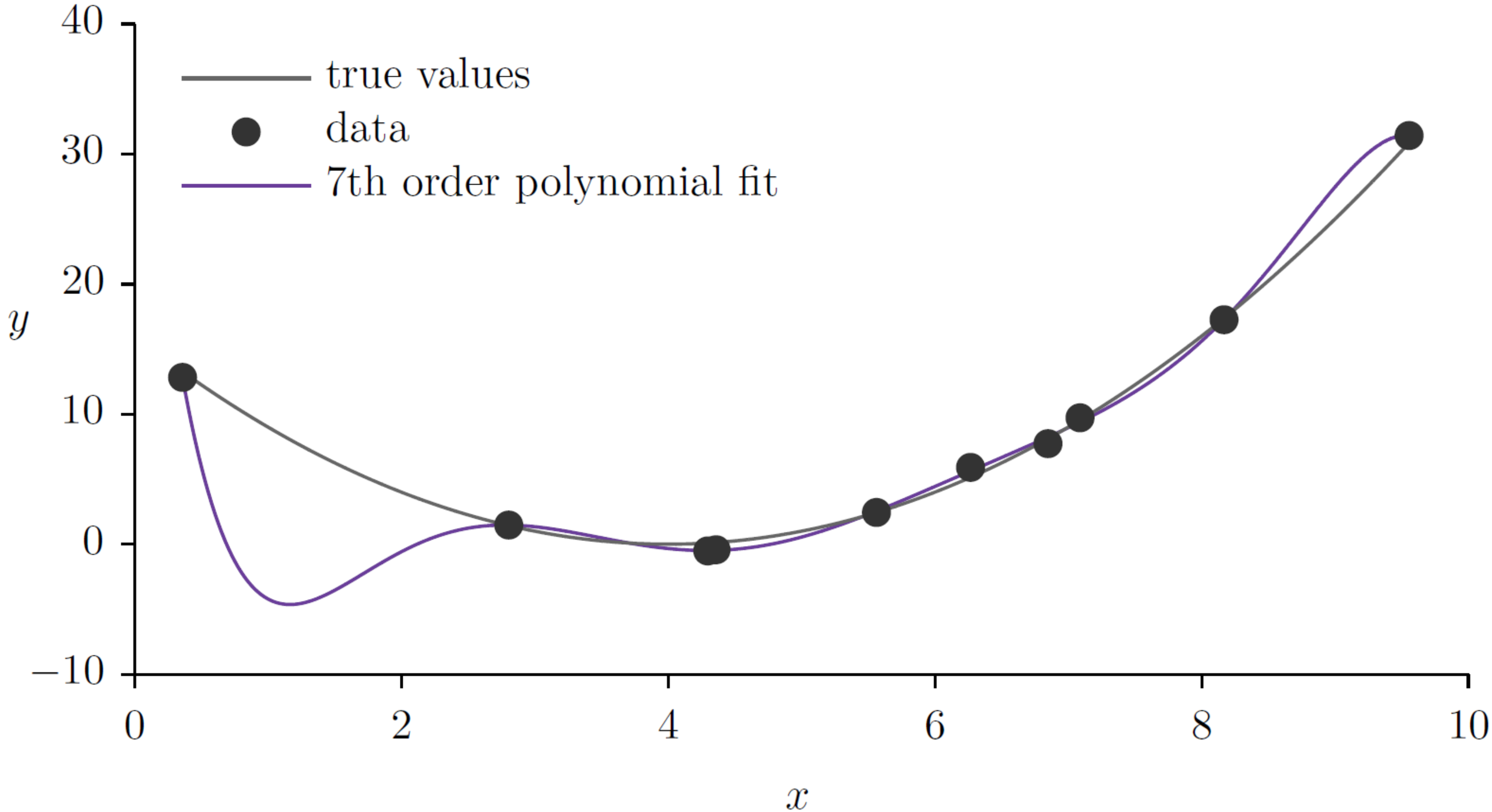


UNIVERSITY OF OXFORD

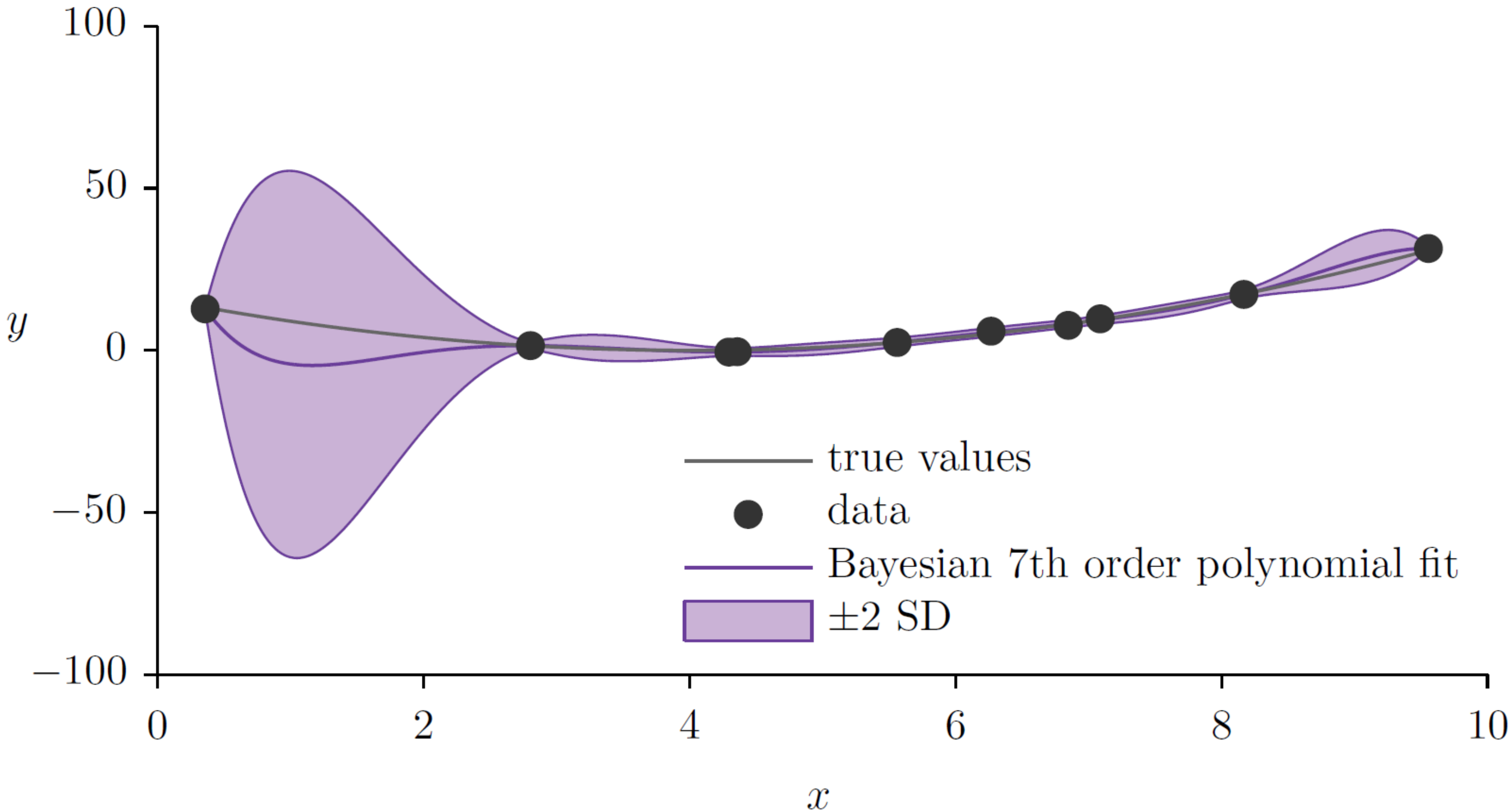
Optimisation (as in maximum likelihood or least squares), gives a reasonable heuristic for exploring the likelihood.



The naïve fitting of models to data performed by maximum likelihood can lead to overfitting.

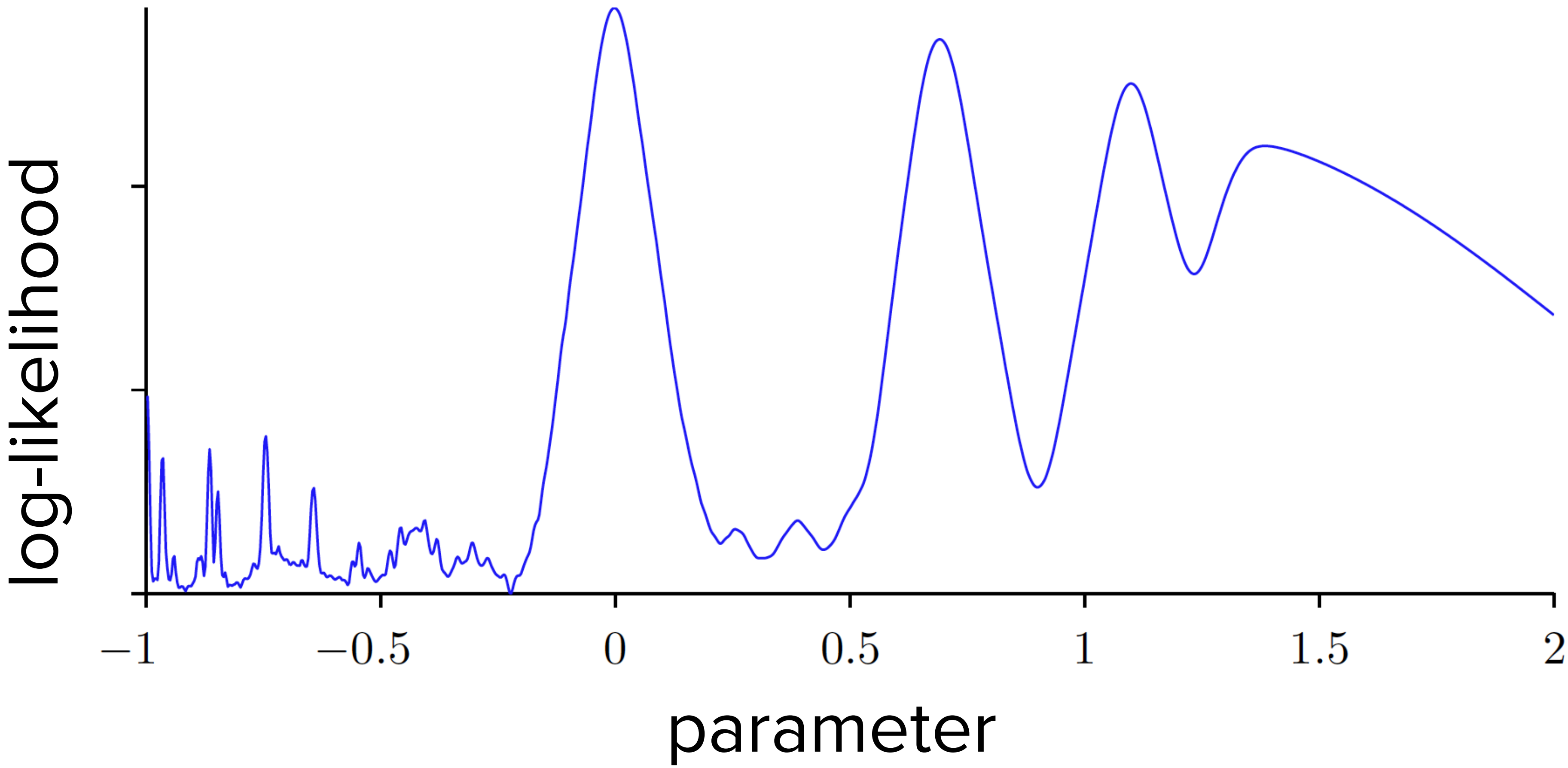


Bayesian averaging over ensembles of models reduces overfitting, and provides more honest estimates of uncertainty.



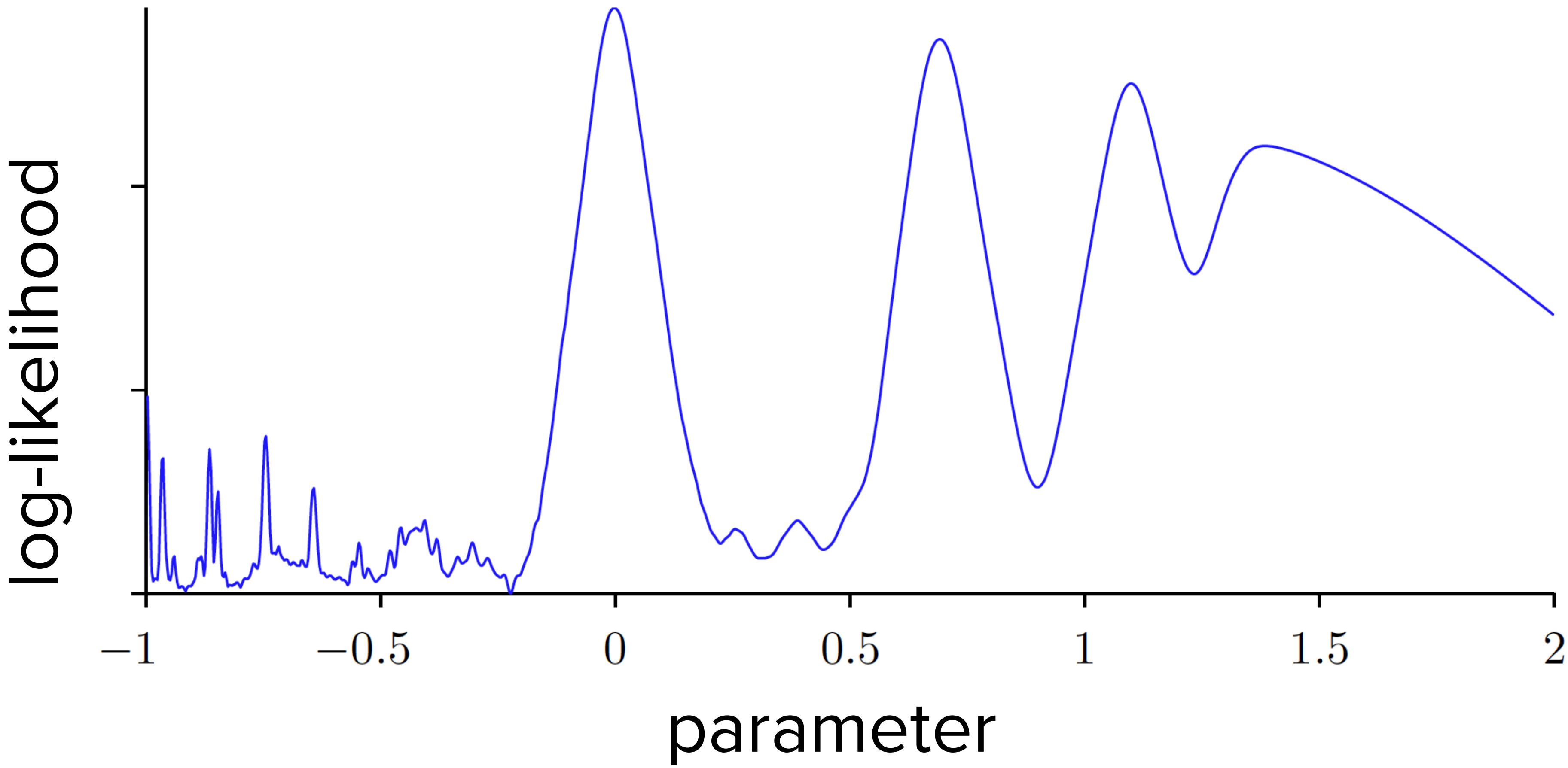
UNIVERSITY OF OXFORD

Averaging requires integrating over the many possible states of the world consistent with data: this is often non-analytic.

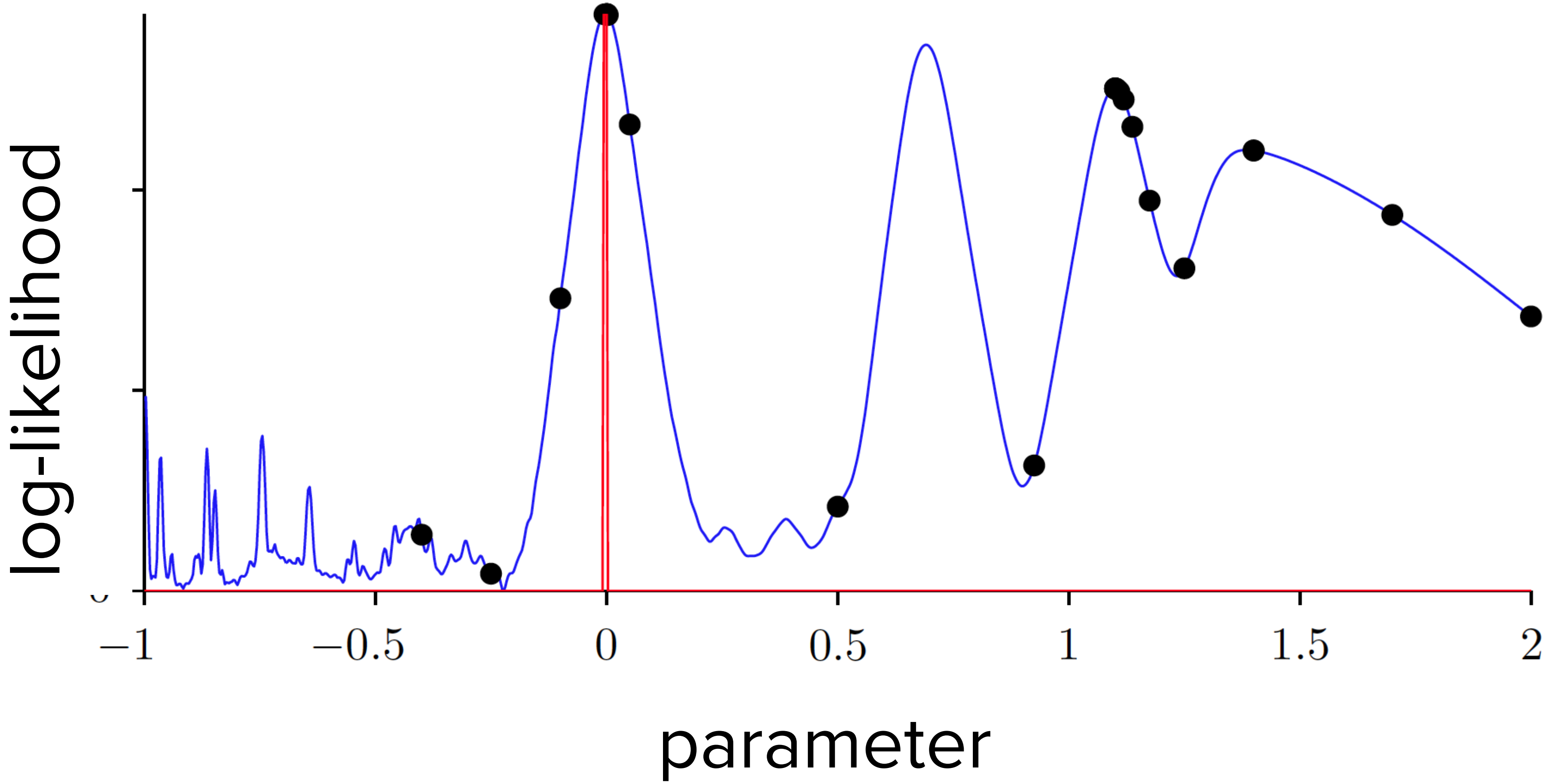


UNIVERSITY OF OXFORD

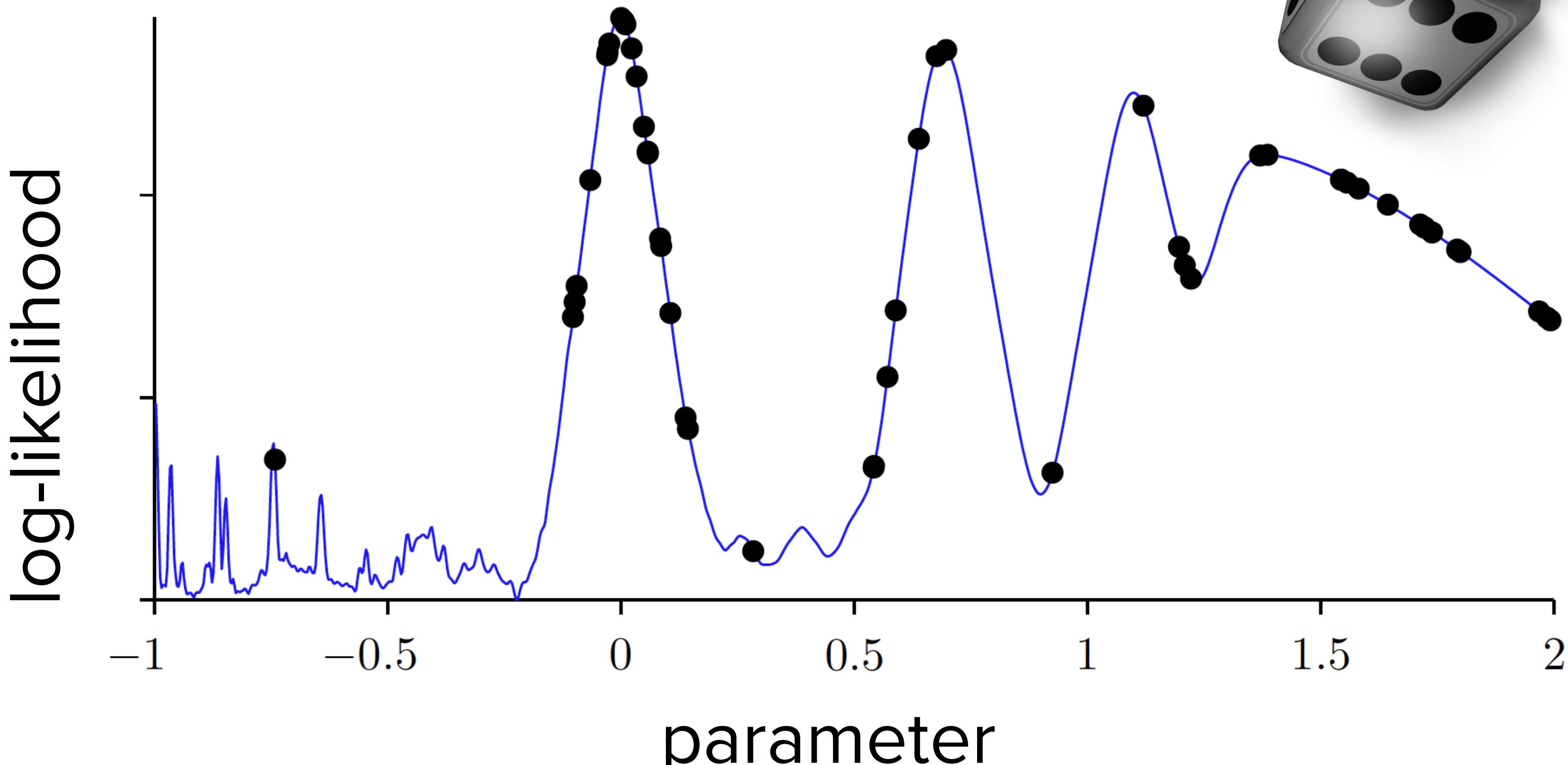
There are many different approaches to quadrature (numerical integration); integrand estimation is undervalued by most.



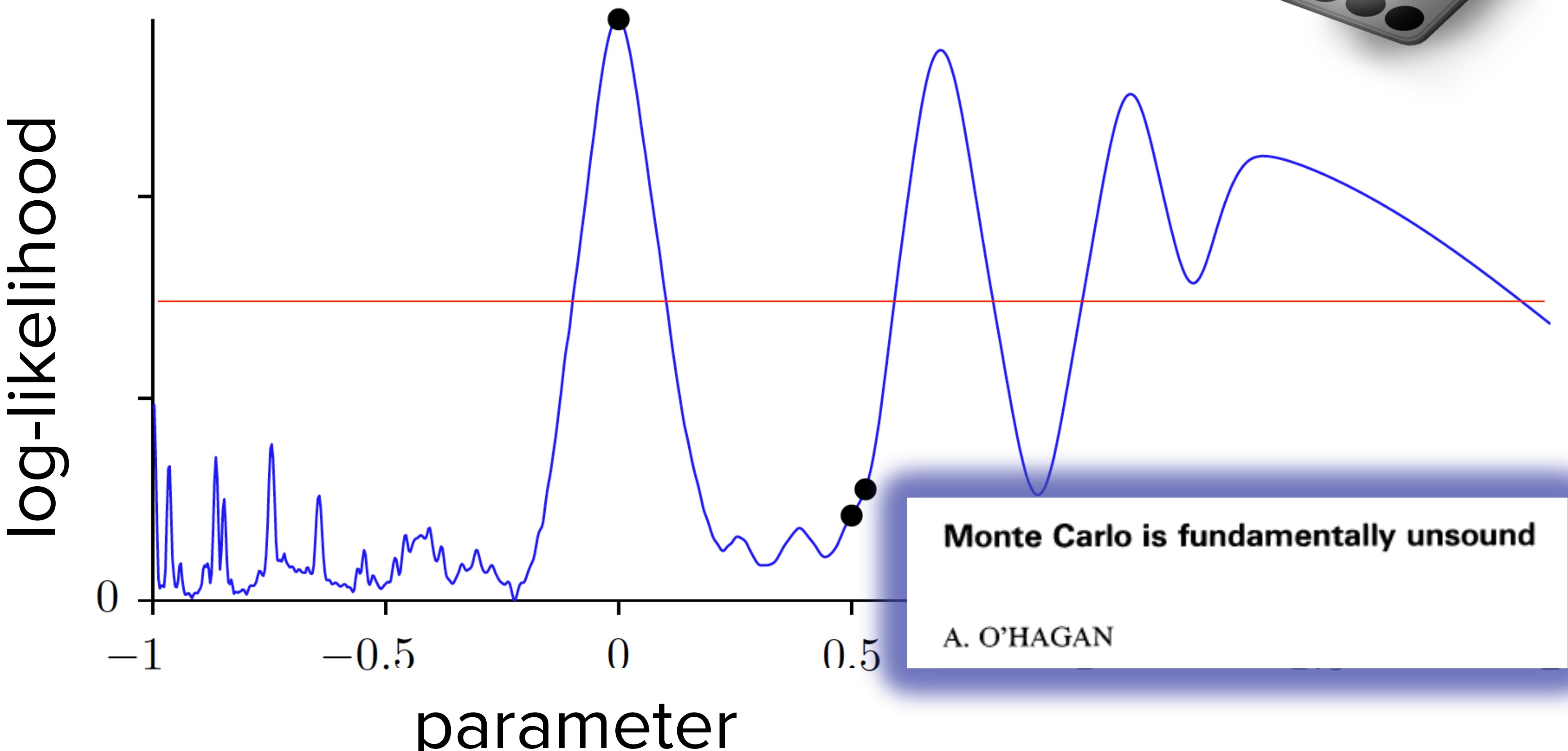
Maximum likelihood is an unreasonable way of estimating a multi-modal or broad likelihood integrand.



Monte Carlo schemes give powerful methods of exploration that have revolutionised Bayesian inference.



Monte Carlo schemes give a reasonable method of exploration, but an unsound means of integrand estimation.



Monte Carlo schemes have some other potential issues:

some parameters must be hand-tuned;

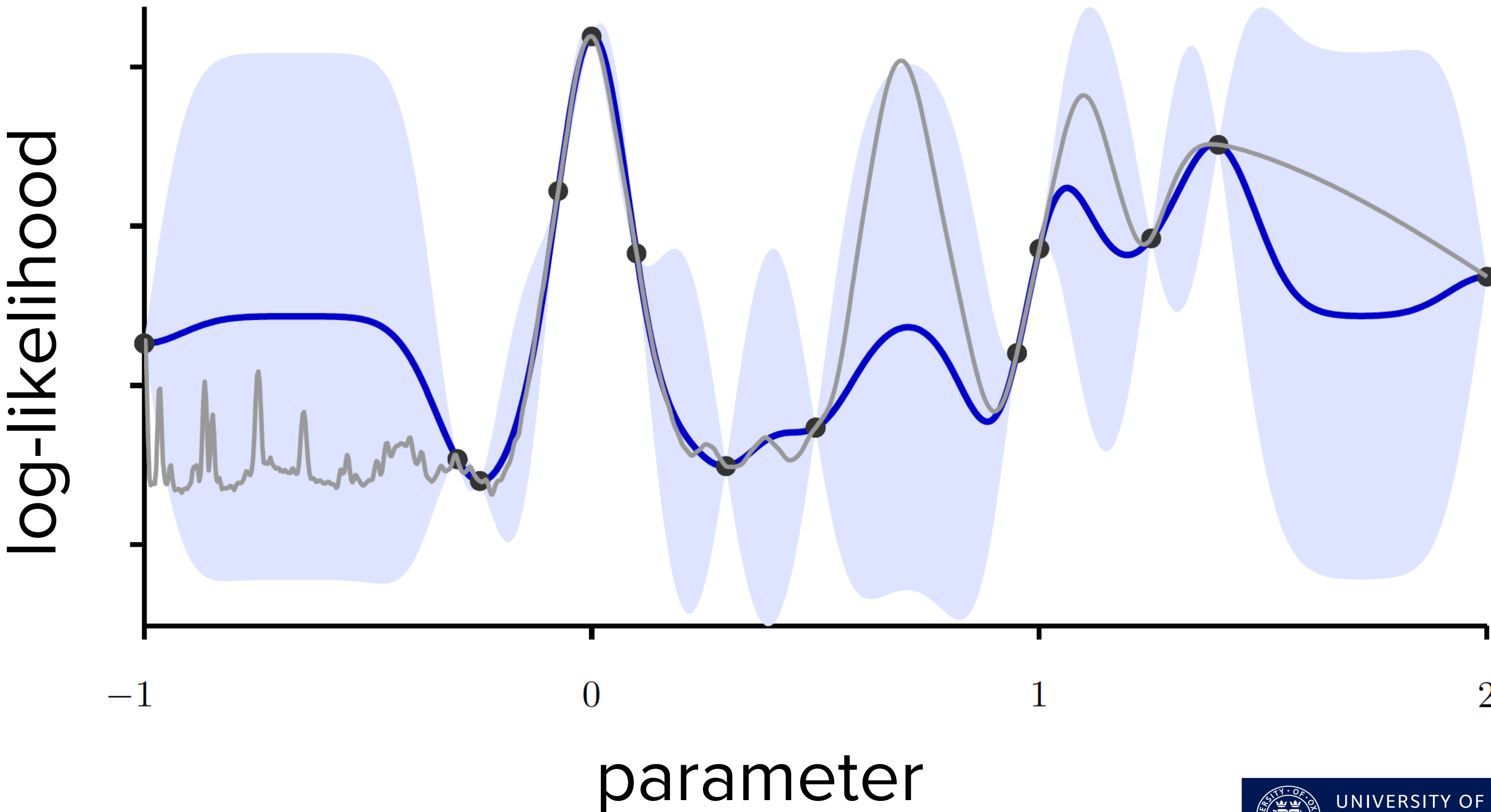
convergence diagnostics are often unreliable.

UNIVERSITY OF OXFORD

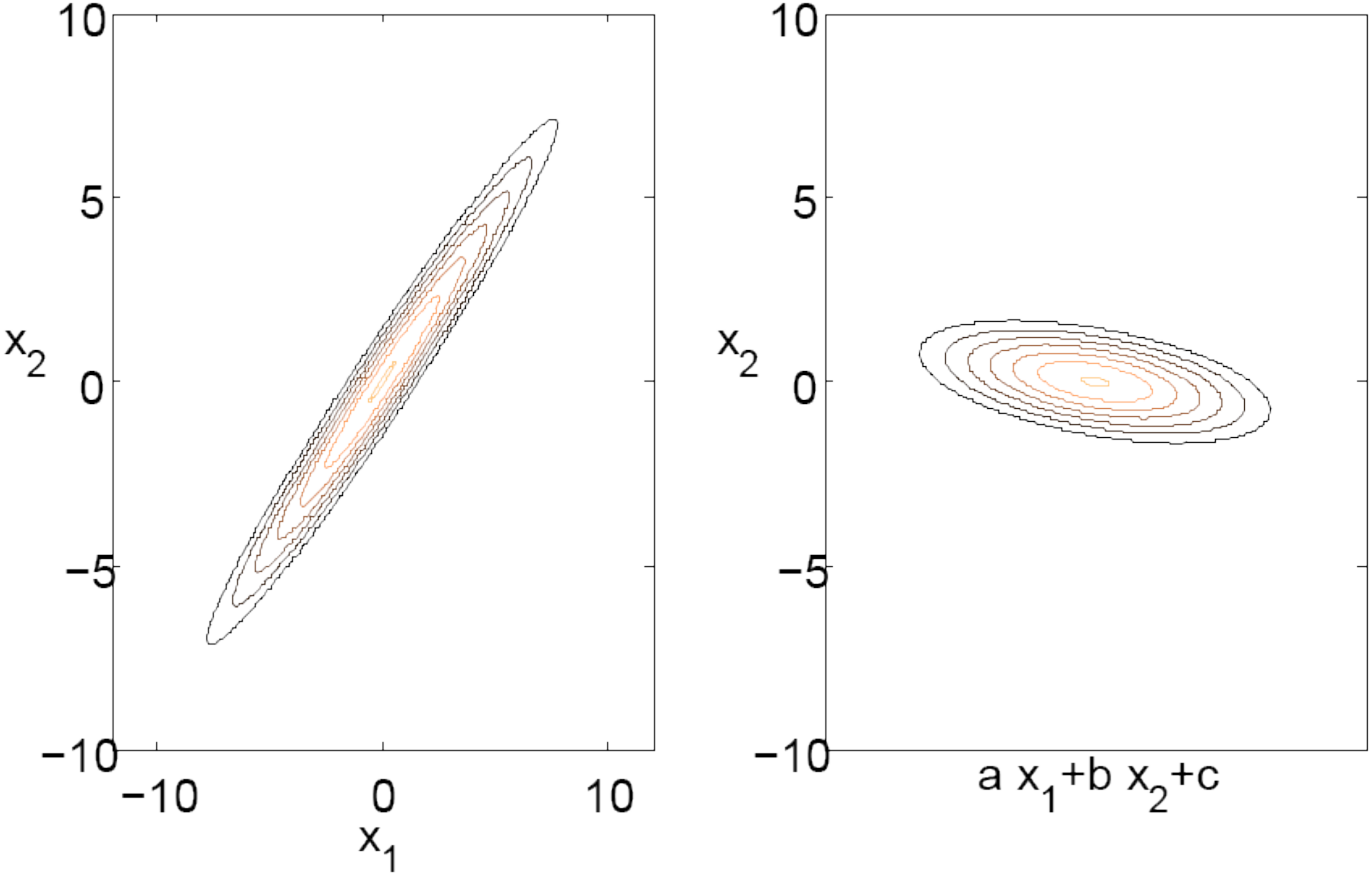
Bayesian quadrature provides optimal ensembles of models for big data.

UNIVERSITY OF OXFORD

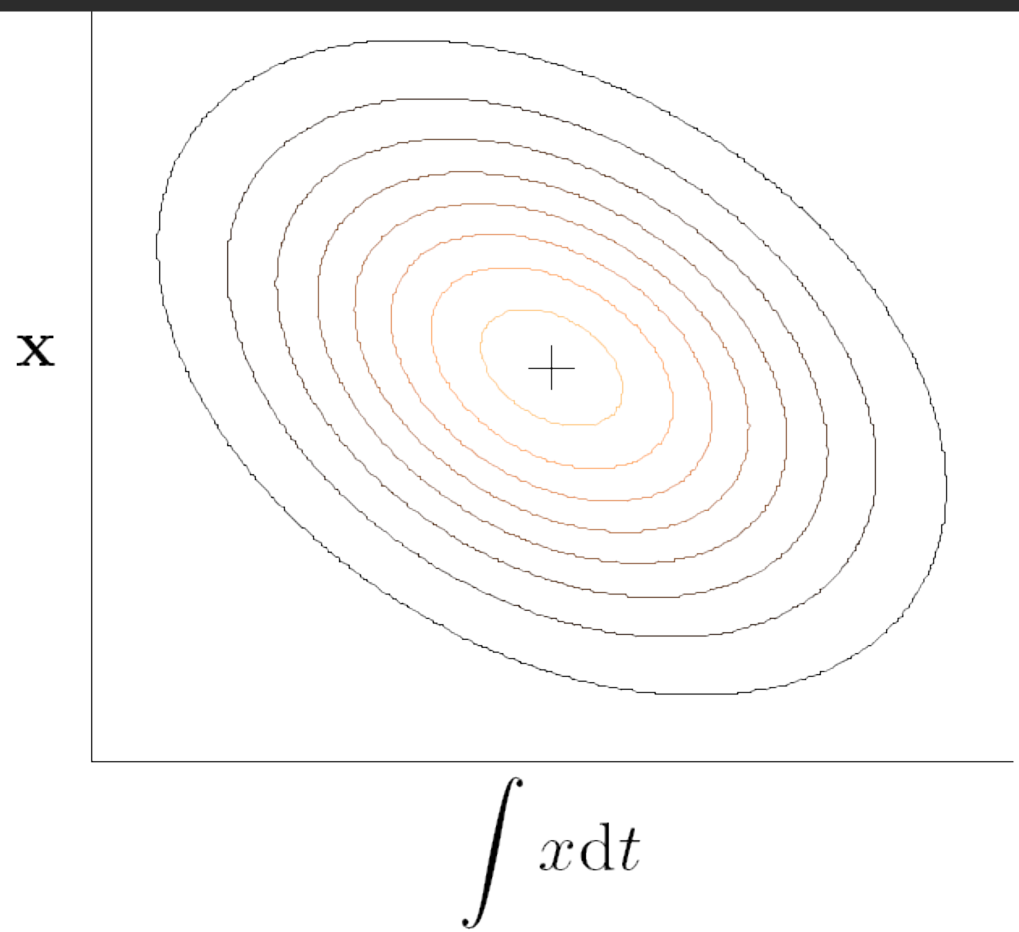
Bayesian quadrature gives a powerful method for estimating the integrand: a Gaussian process.



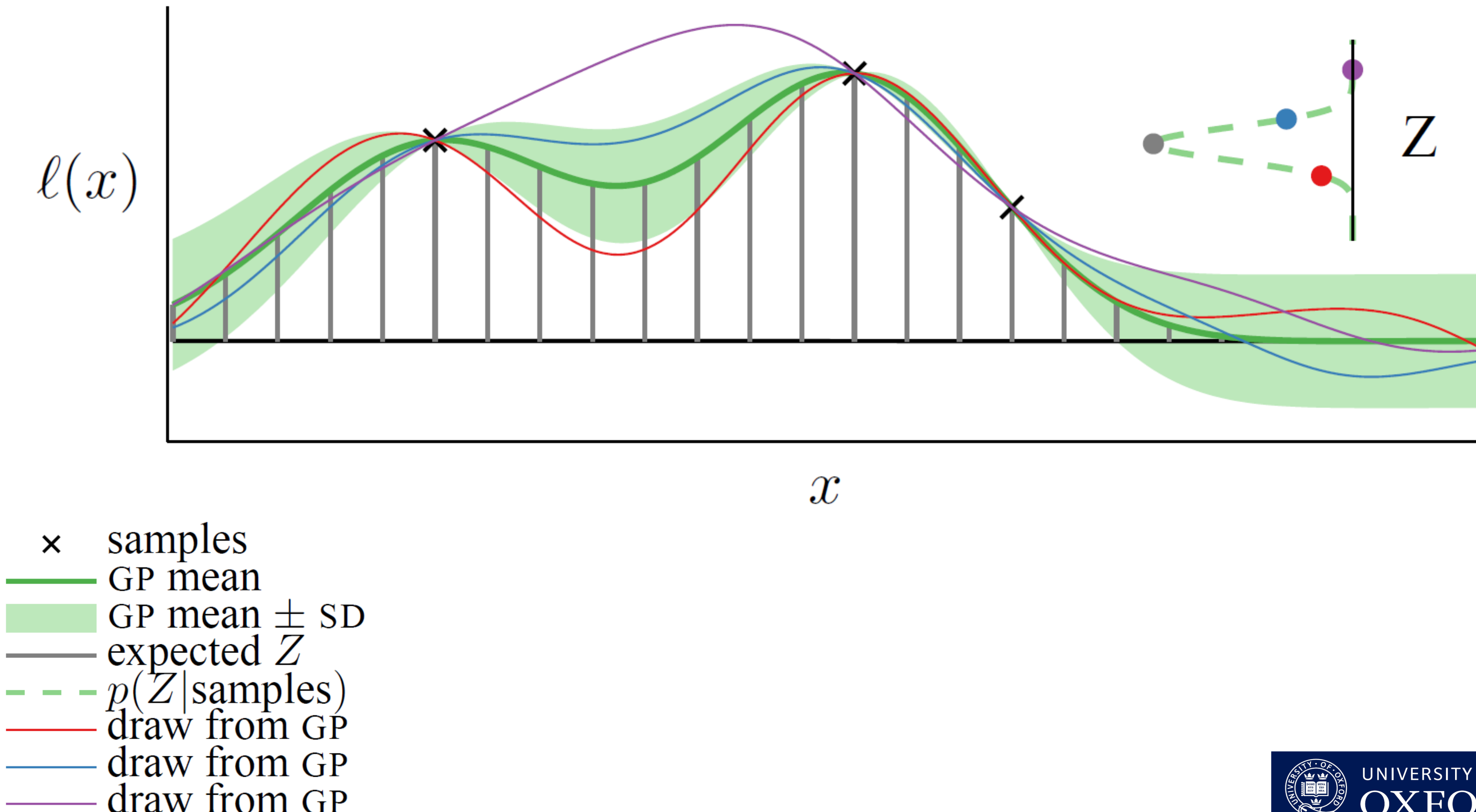
Gaussian distributed variables are joint Gaussian with any affine transform of them.



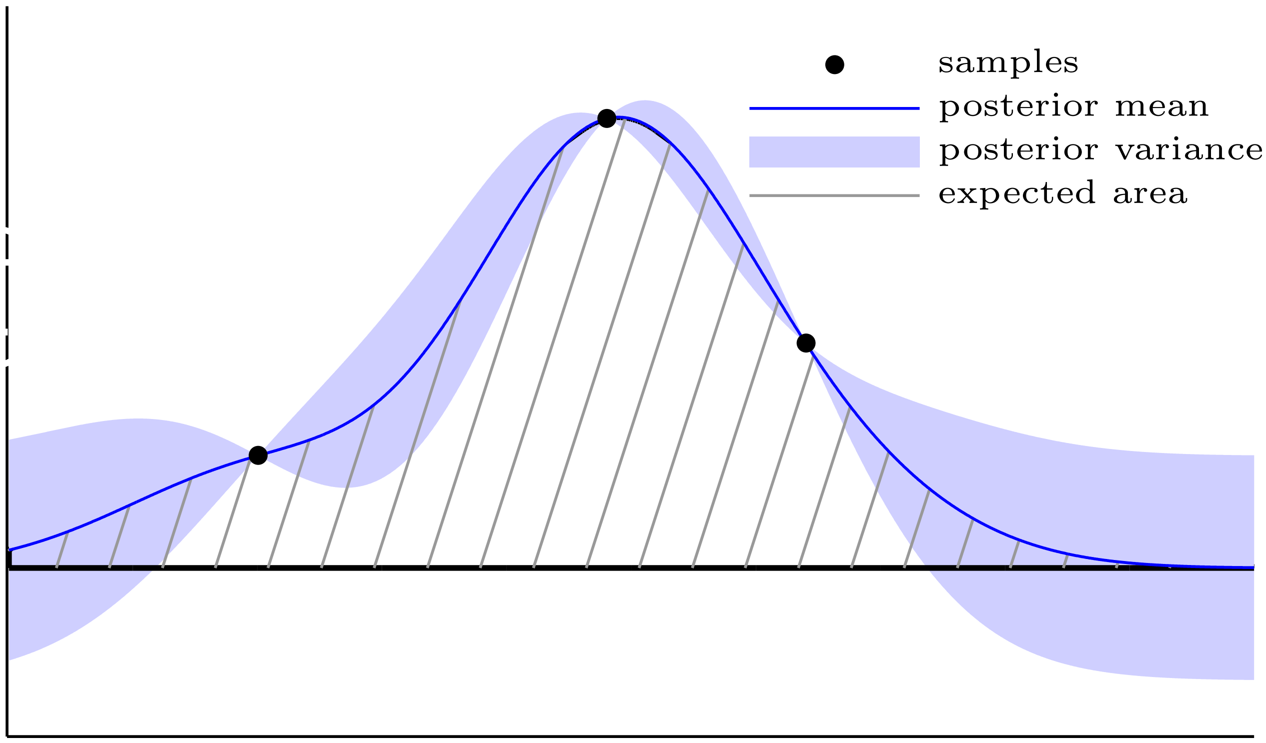
A function over which we have a Gaussian process is joint Gaussian with any integral or derivative of it, as integration and differentiation are linear.



We can use observations of an integrand ℓ in order to perform inference for its integral, Z: this is known as Bayesian Quadrature.



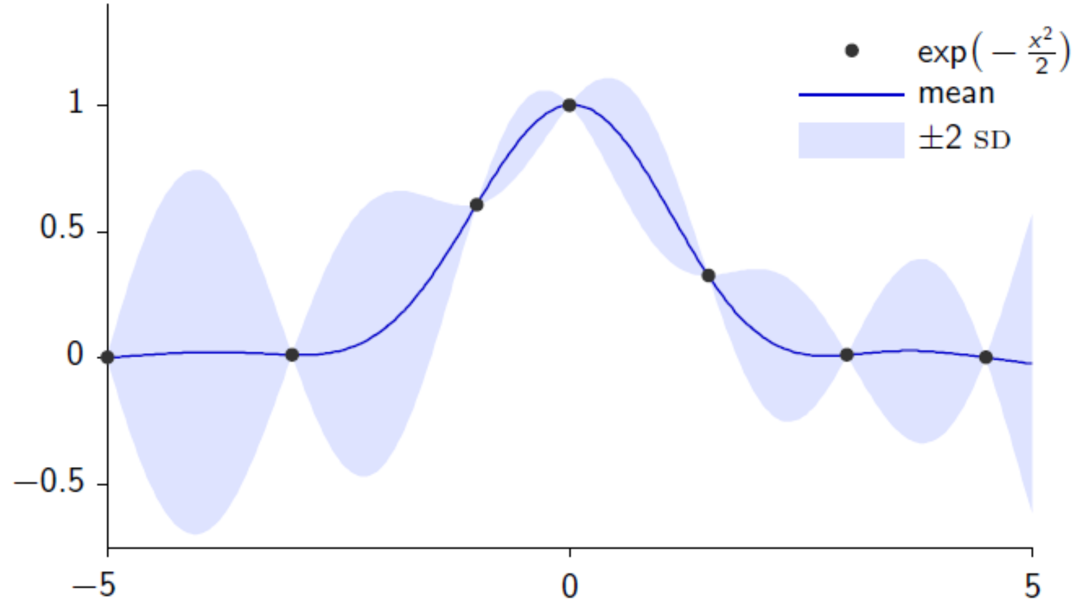
Bayesian quadrature makes best use of our evaluations of model fit, important for big data, where such evaluations are expensive.

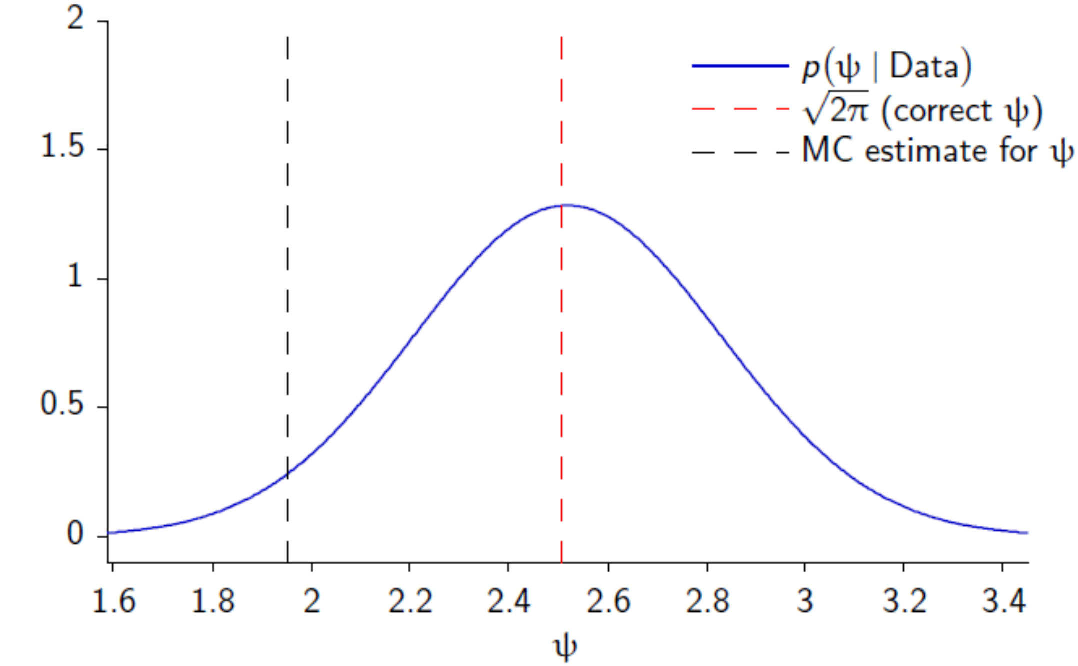


Consider the integral

$$\psi = \int_{-5}^{5} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \, .$$

Bayesian quadrature achieves more accurate results than Monte Carlo, and provides an estimate of our uncertainty.





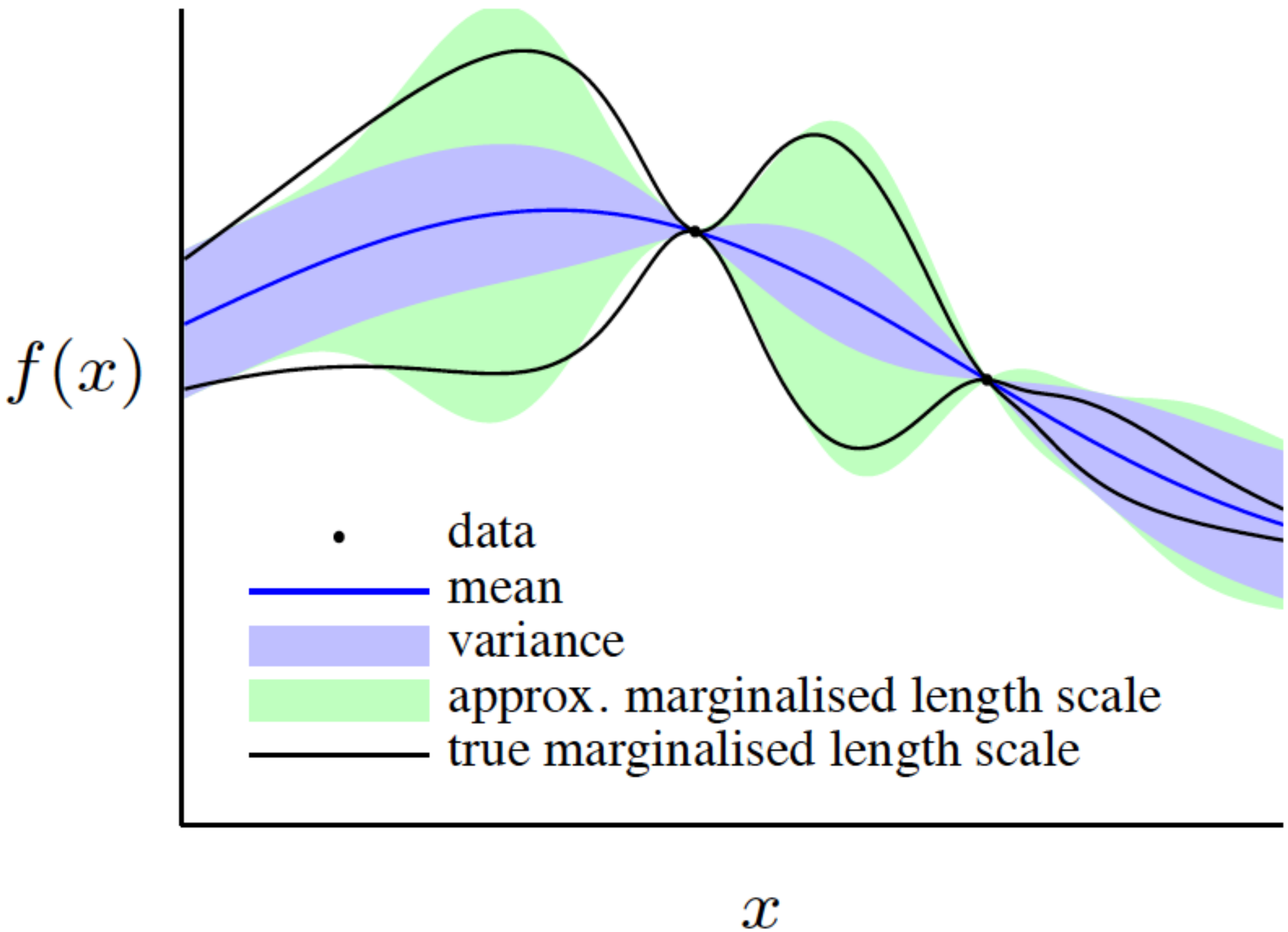
UNIVERSITY OF OXFORD

Bayesian quadrature promises solutions to some of the issues of Monte Carlo:

hyper-parameters can be automatically set by maximum likelihood;

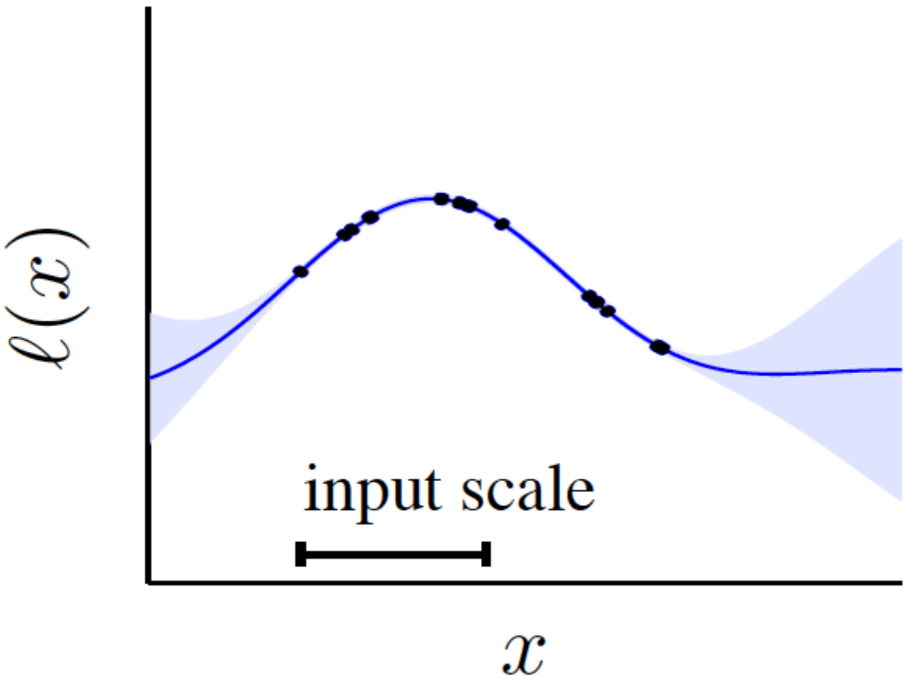
the variance in the integral is a natural convergence diagnostic.

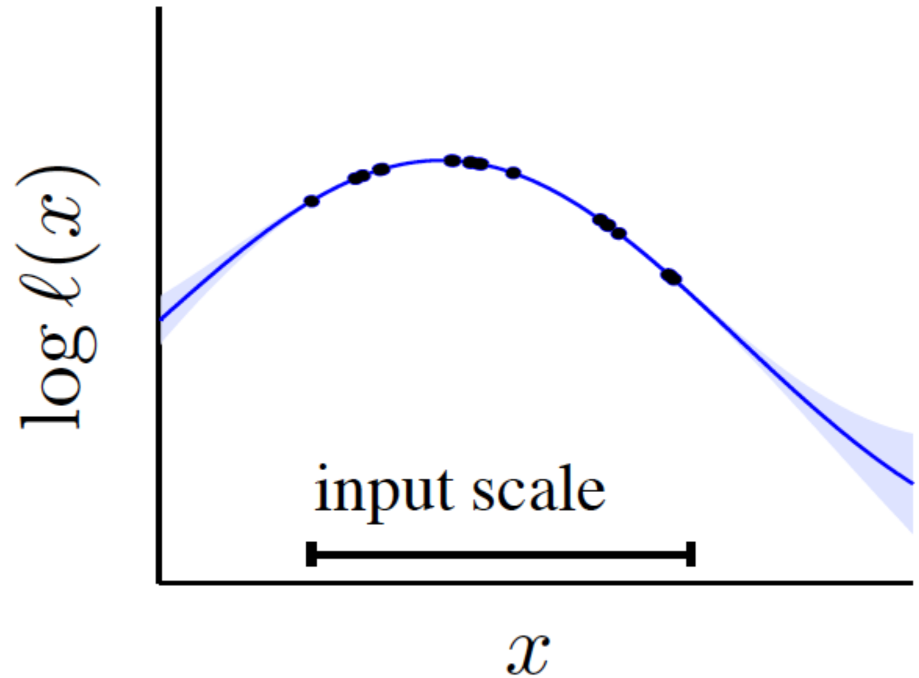
We use a Laplace approximation to marginalise the hyperparameters of the Gaussian process model.



UNIVERSITY OF OXFORD

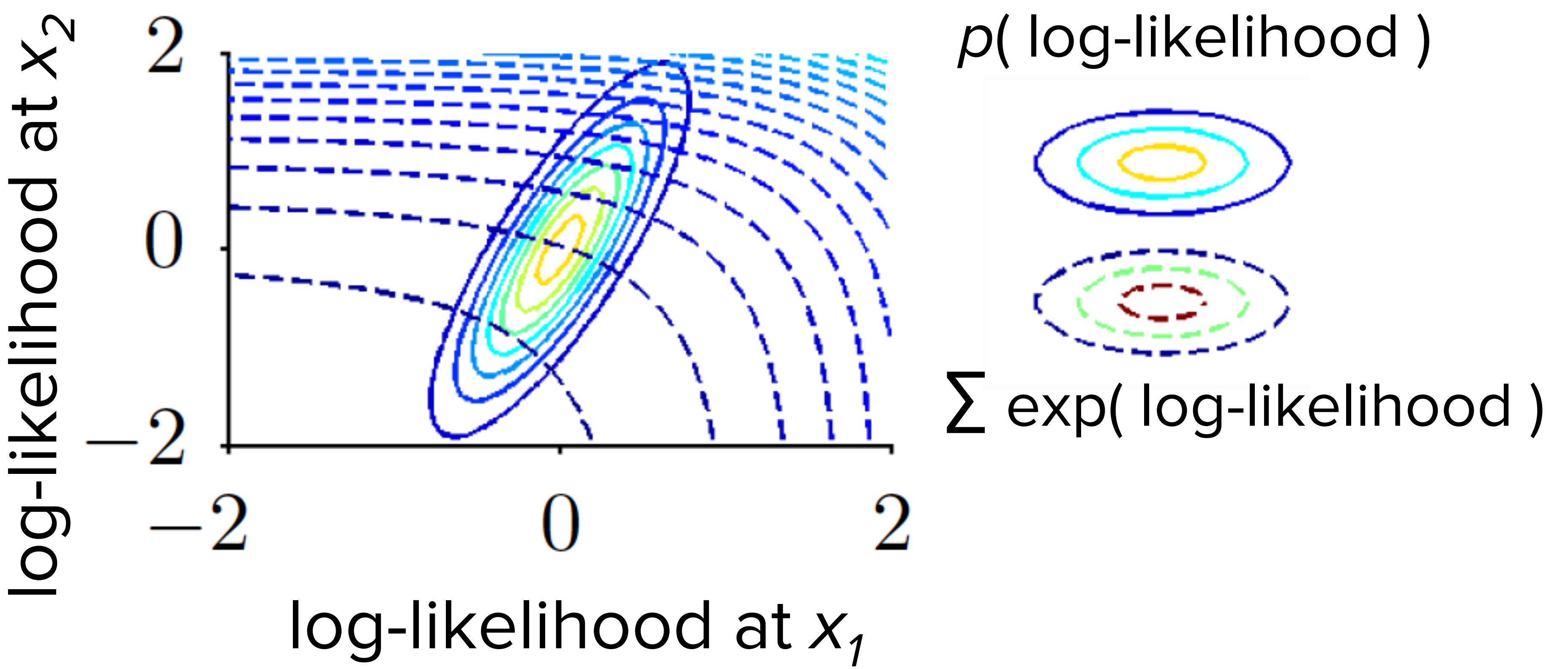
We really want to use a Gaussian process to model the log-likelihood, rather than the likelihood.

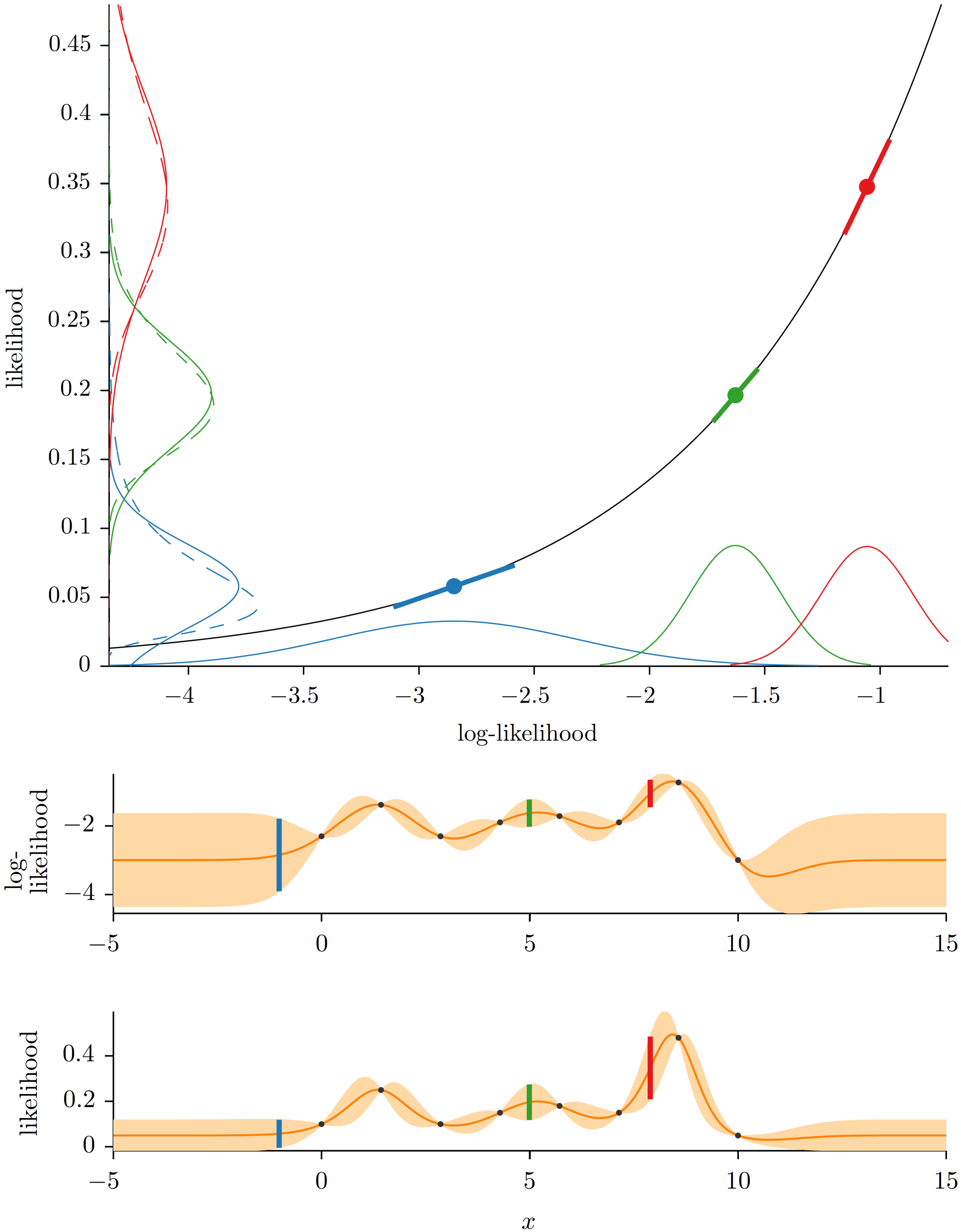




Doing so better captures the dynamic range of likelihoods, and extends the correlation range.

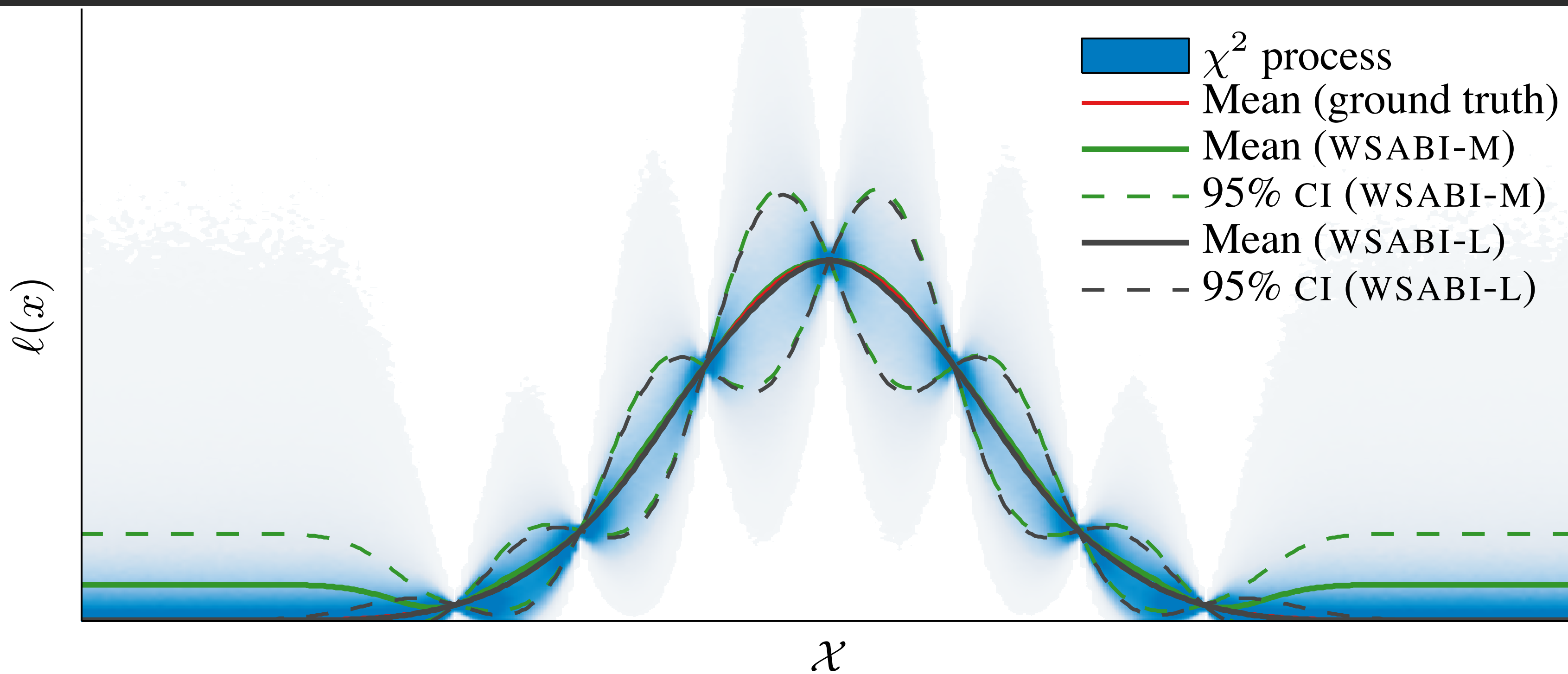
Using a Gaussian process for the log-likelihood means that the distribution for the integral of the likelihood is no longer analytic.





We could linearise the likelihood as a function of the log-likelihood. This renders the likelihood and log-likelihood jointly part of one Gaussian process (along with integrals).

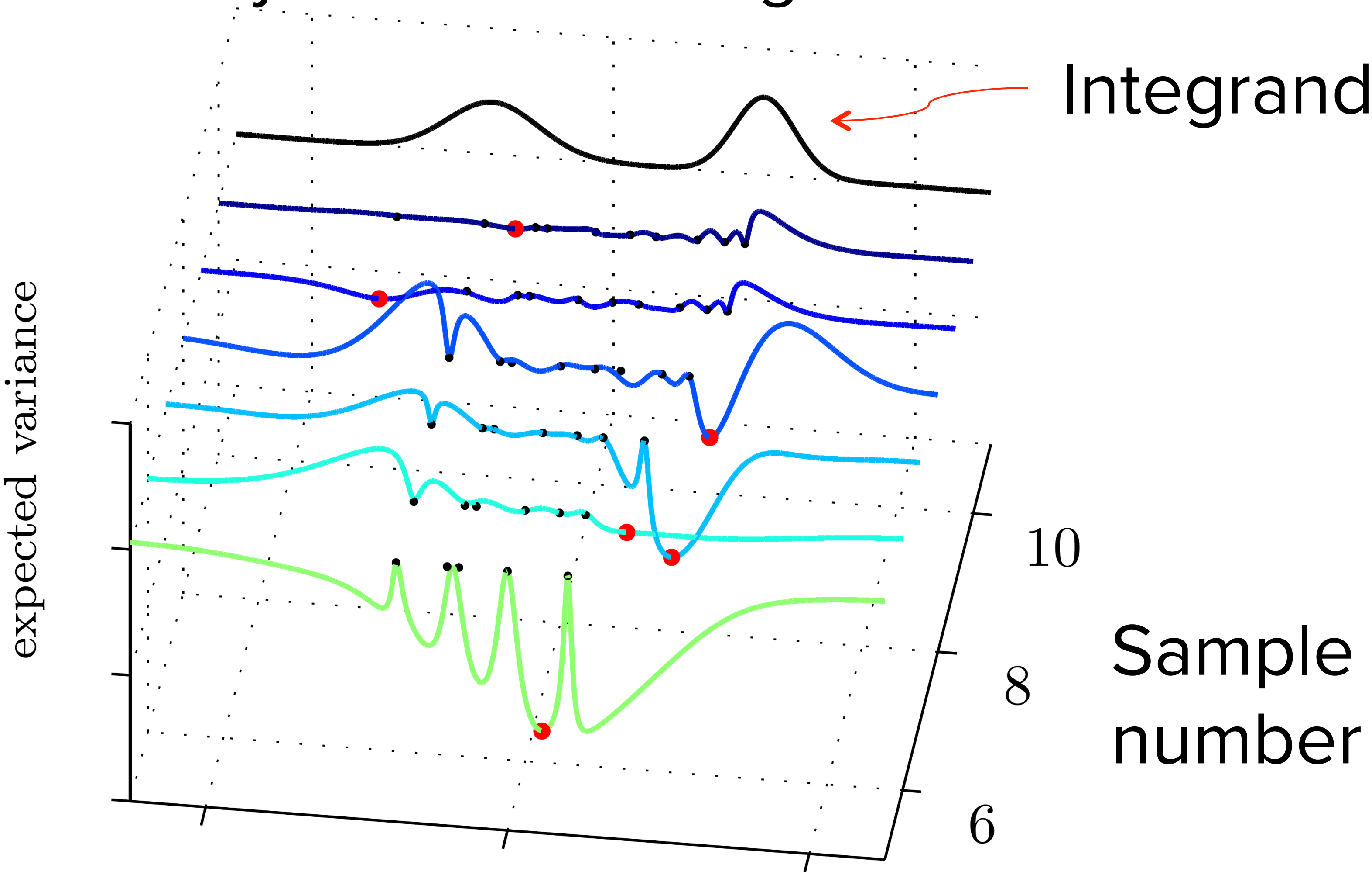
However, this linearisation is typically poor for the extreme log transform.



Rather than the log, we model the square-root (WSABI), which is more amenable to linearisation.

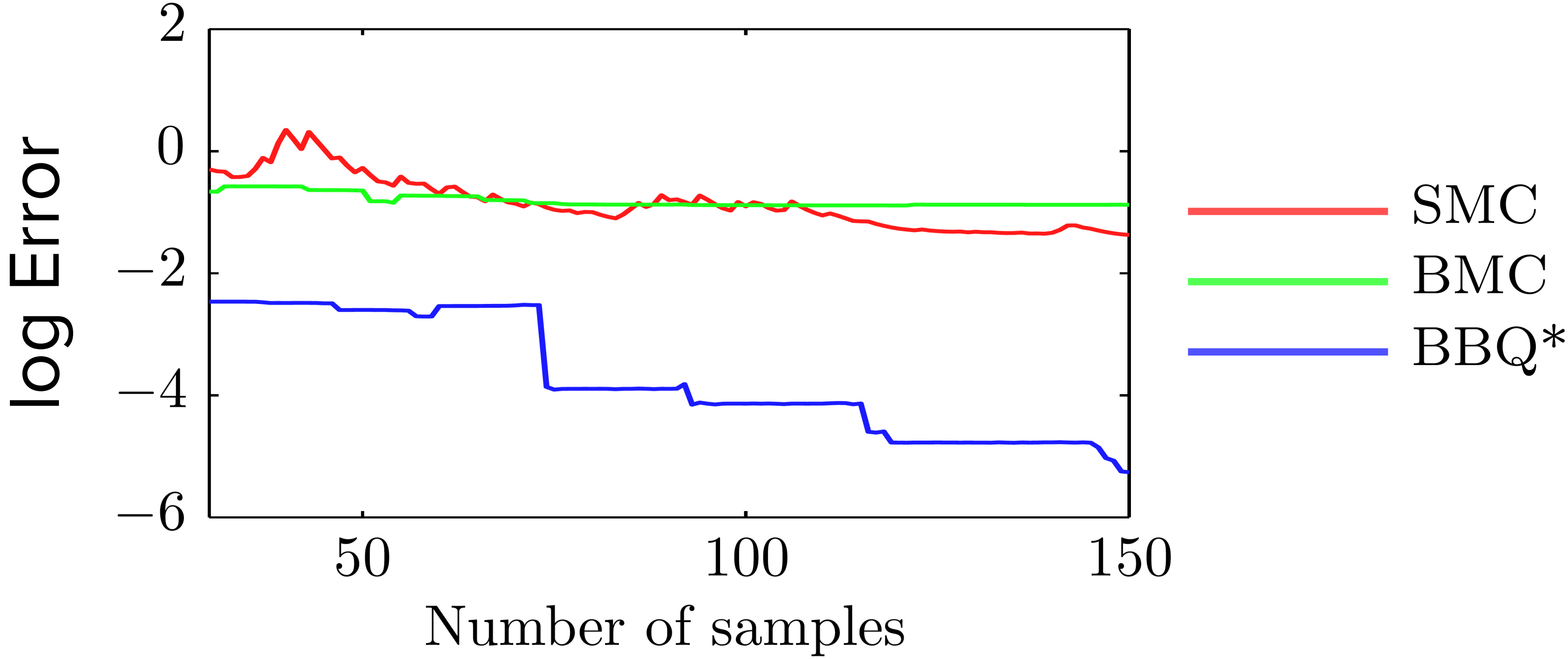
UNIVERSITY OF OXFORD

Doubly-Bayesian quadrature (BBQ) additionally explores the integrand so as to minimise the uncertainty about the integral.

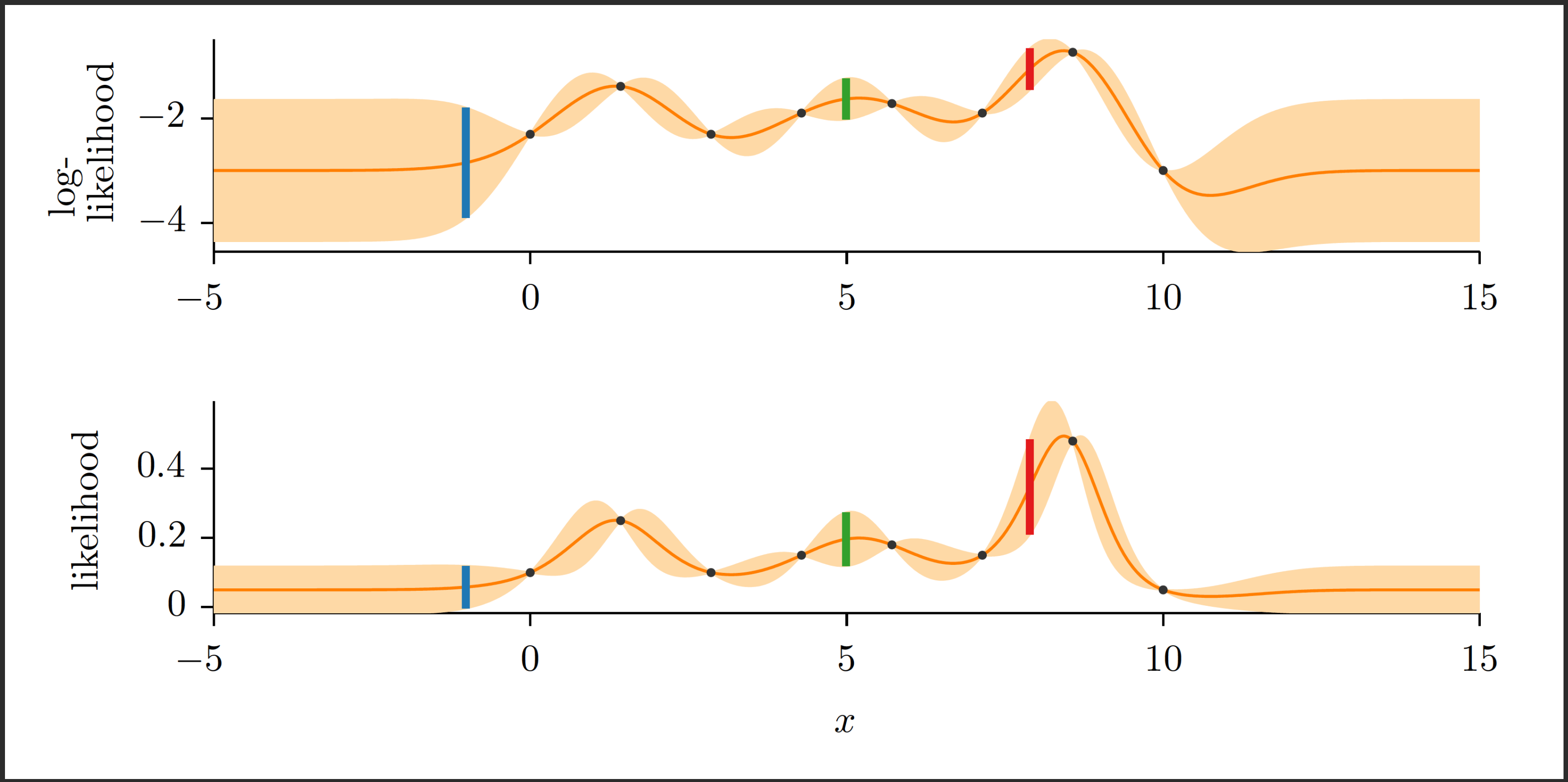


UNIVERSITY OF OXFORD

Doubly-Bayesian Quadrature (BBQ) selects efficient samples, but the computation of the expected reduction in integral variance is extremely costly.

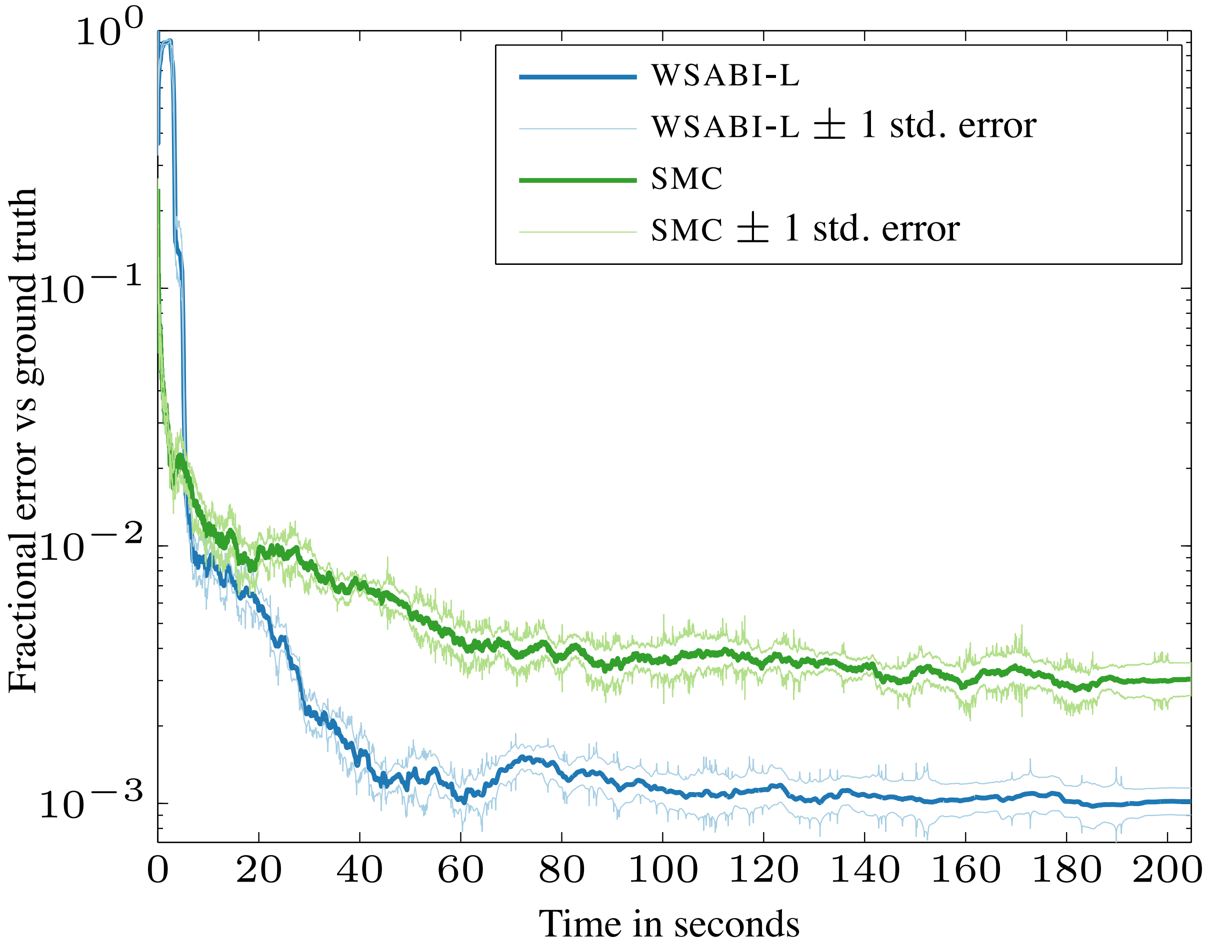


Our linearisation implies we are more uncertain about large likelihoods than small likelihoods.



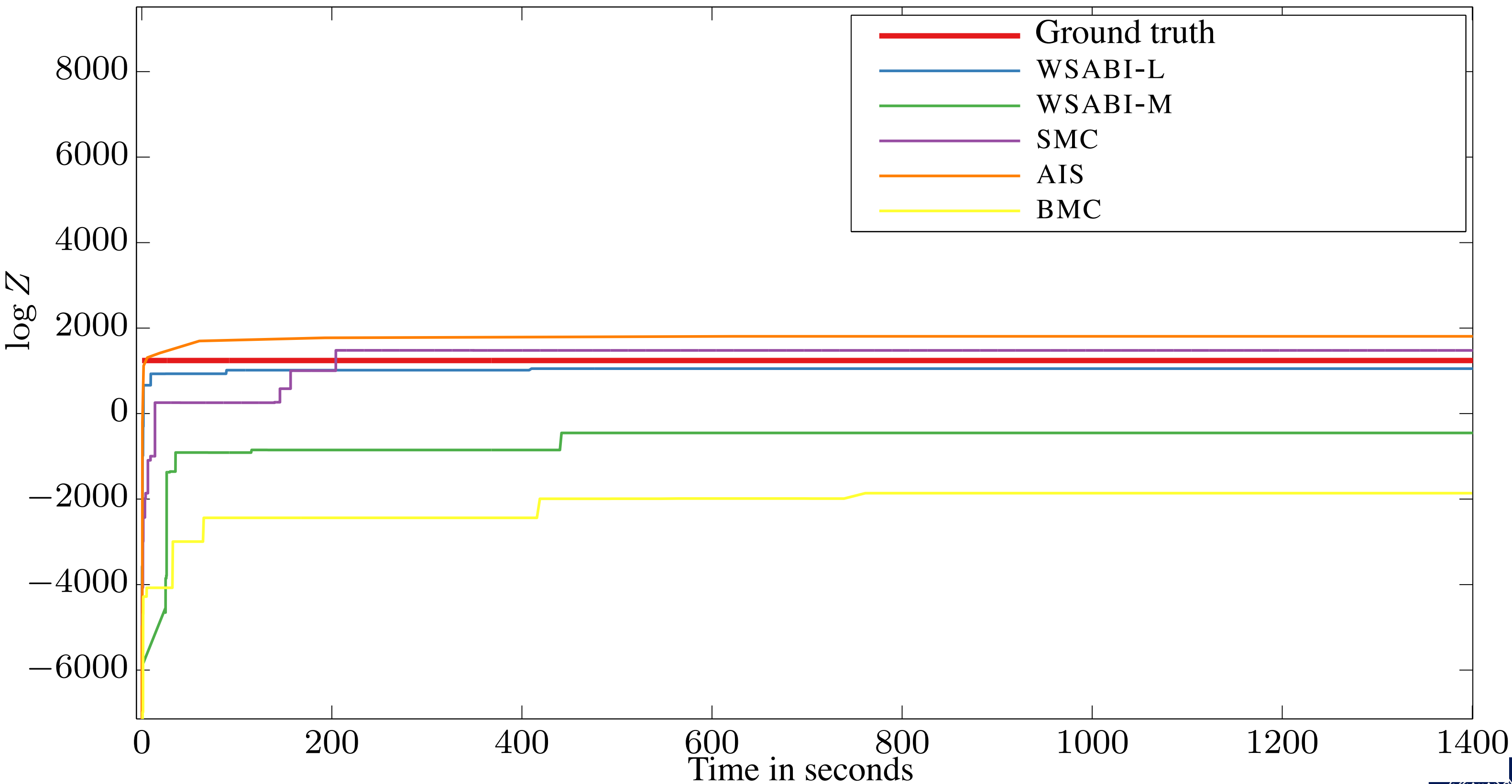
Hence selecting samples with large variance promotes both exploration and exploitation.

Our method (Warped Sequential Active Bayesian Integration) converges quickly in wall-clock time for a synthetic integrand.



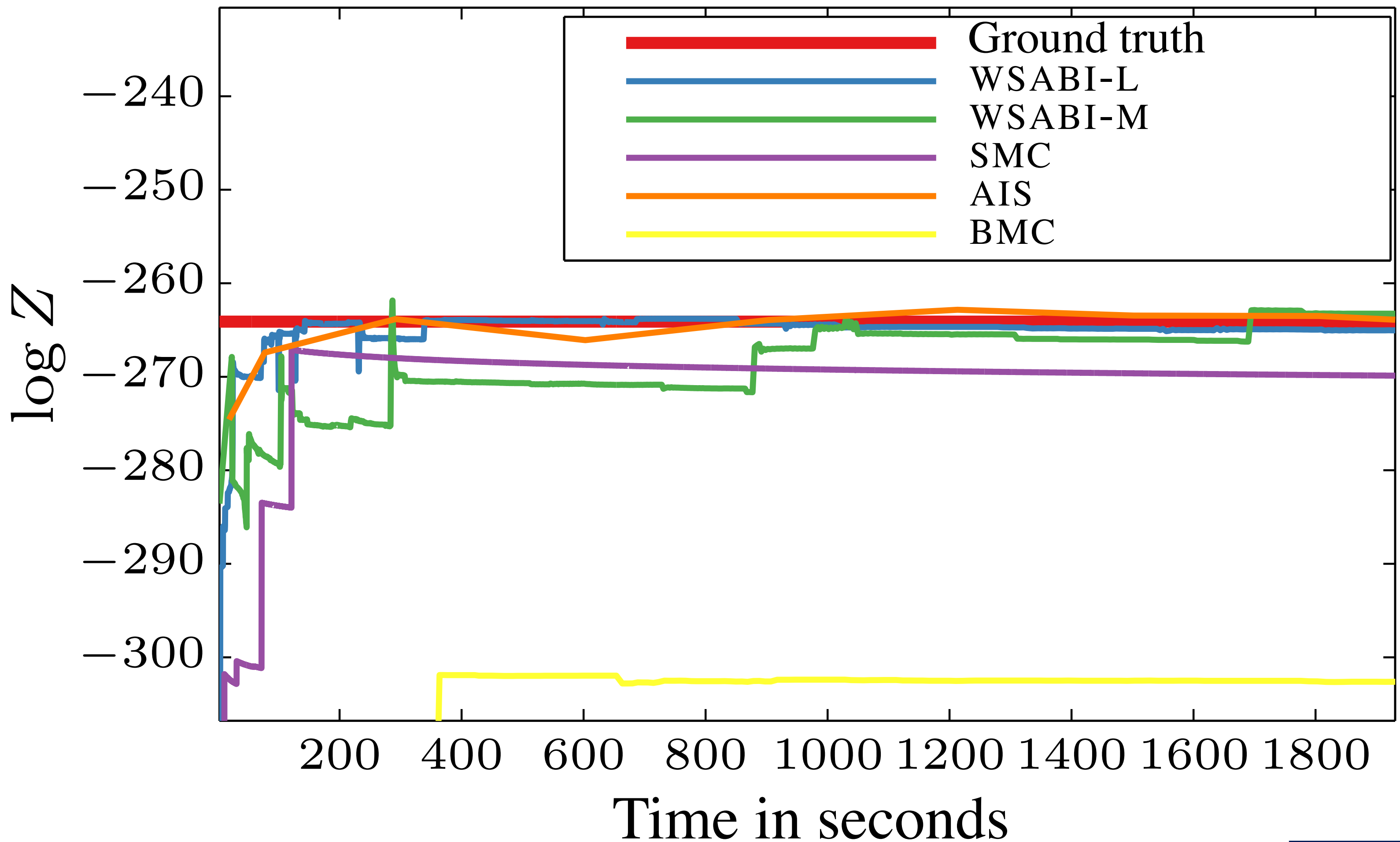
UNIVERSITY OF OXFORD

WSABI-L converges more quickly than Annealed Importance Sampling in integrating out eight hyperparameters in a Gaussian process regression problem (yacht).



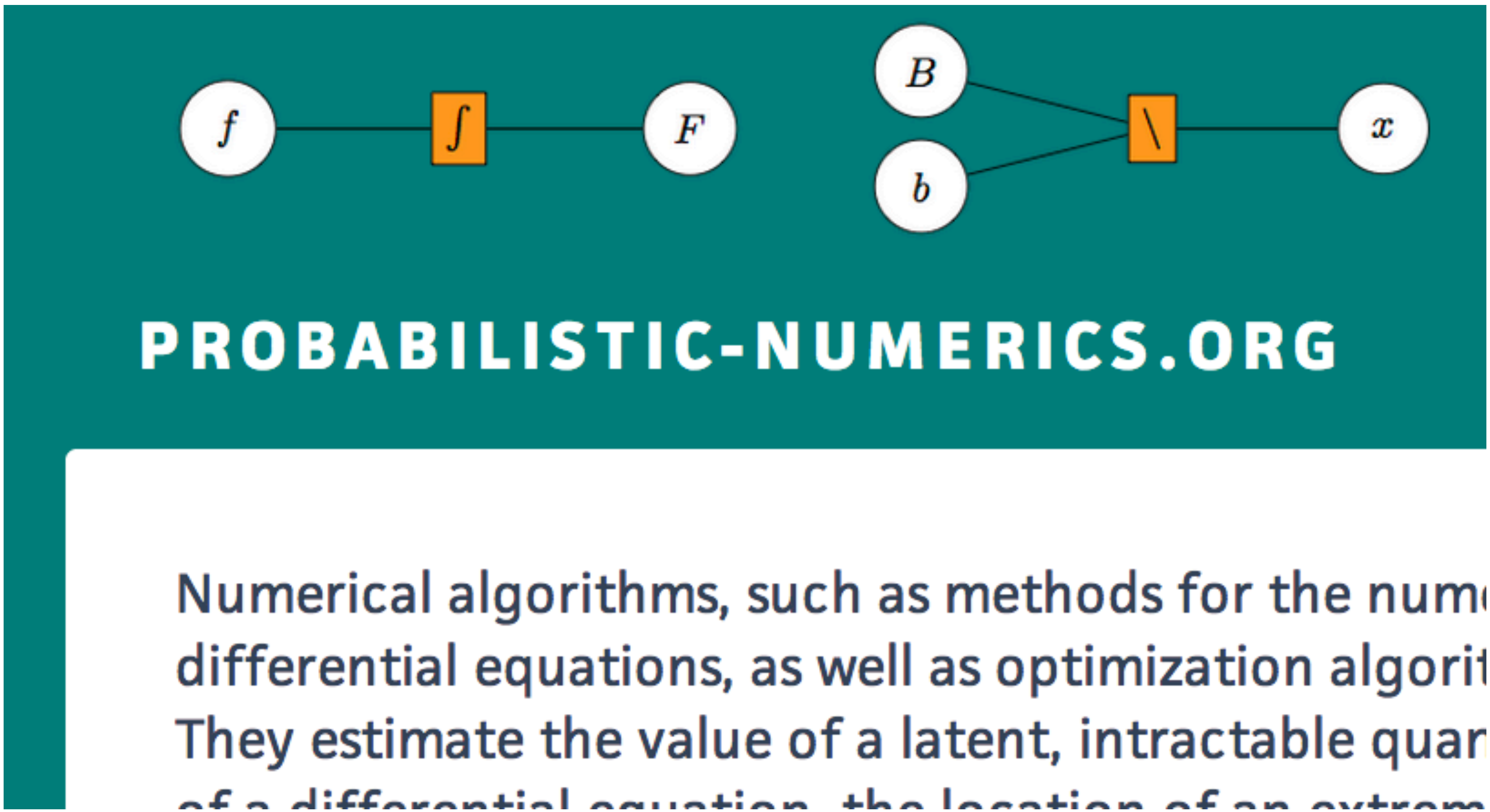
University of Oxford

WSABI-L converges quickly in integrating out hyperparameters in a Gaussian process classification problem (CiteSeer^x data).



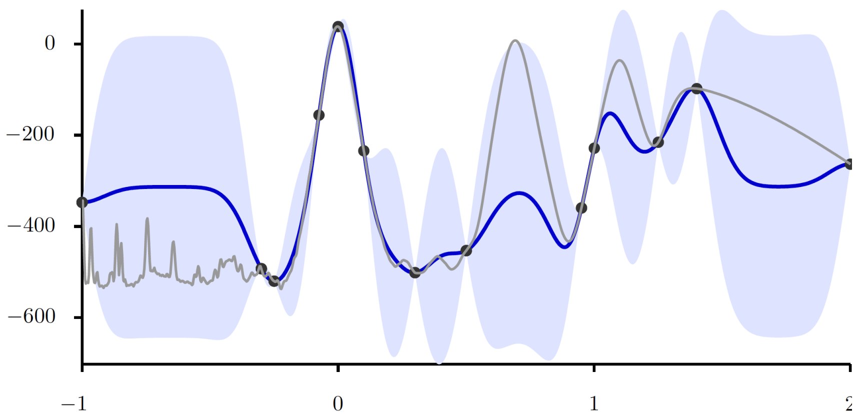
Active Bayesian quadrature gives optimal averaging over models for big and complex data.

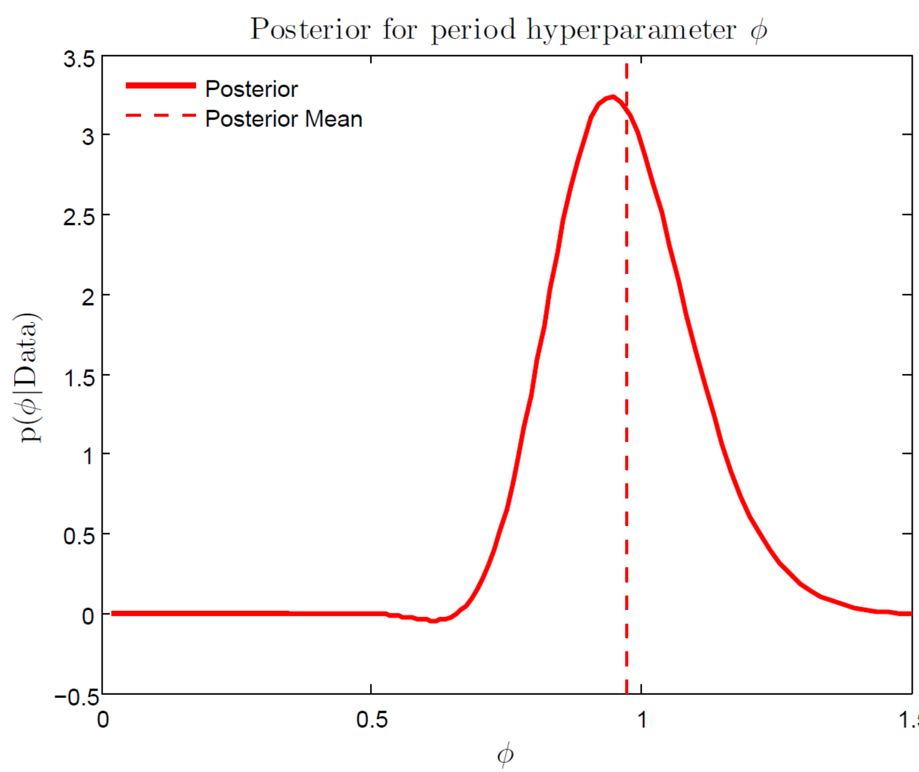
Bayesian quadrature is an example of probabilistic numerics: the study of numeric methods as learning algorithms.



UNIVERSITY OF OXFORD

With Bayesian quadrature, we can also estimate integrals to compute posterior distributions for any hyperparameters.



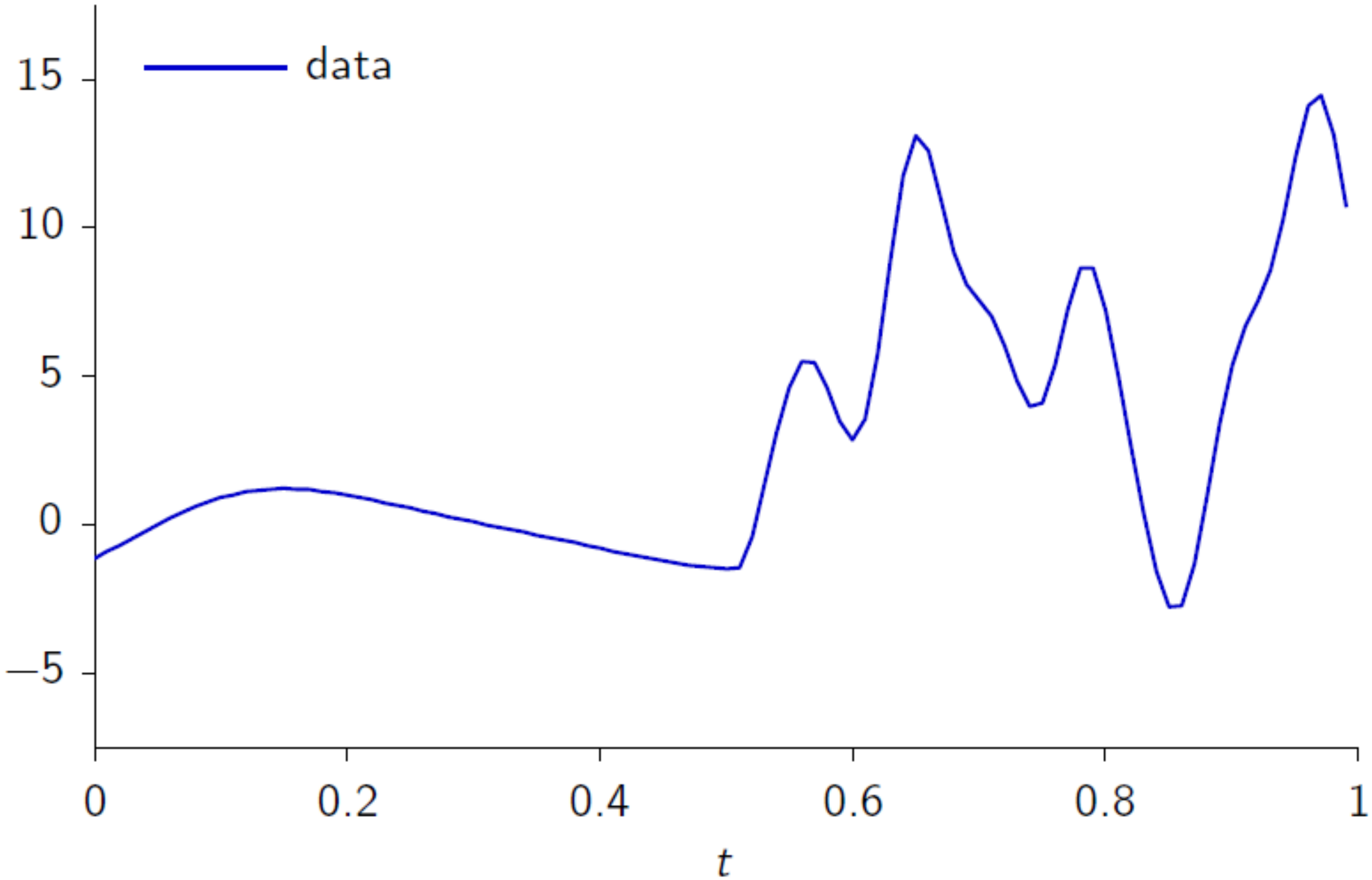


UNIVERSITY OF OXFORD

Complex data with anomalies, changepoints and faults demands model averaging.

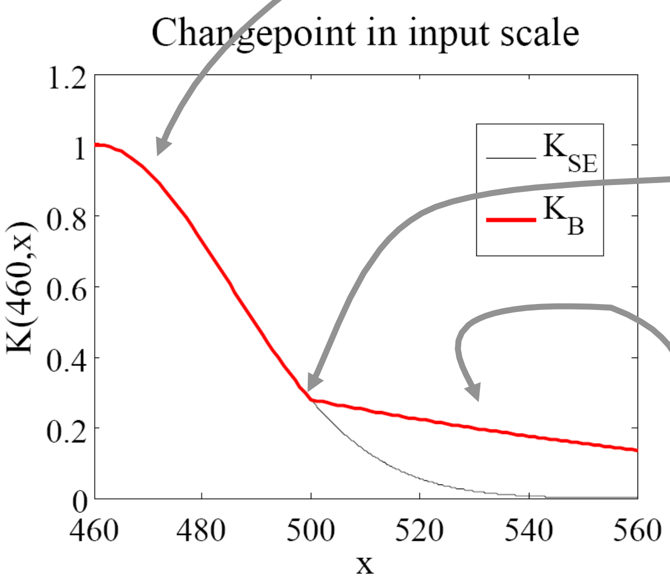
UNIVERSITY OF OXFORD

In considering data with changepoints and faults, we must entertain multiple hypotheses using Bayesian quadrature.



UNIVERSITY OF OXFORD

Changepoint covariances feature hyperparameters, for which we can produce posterior distributions using quadrature.



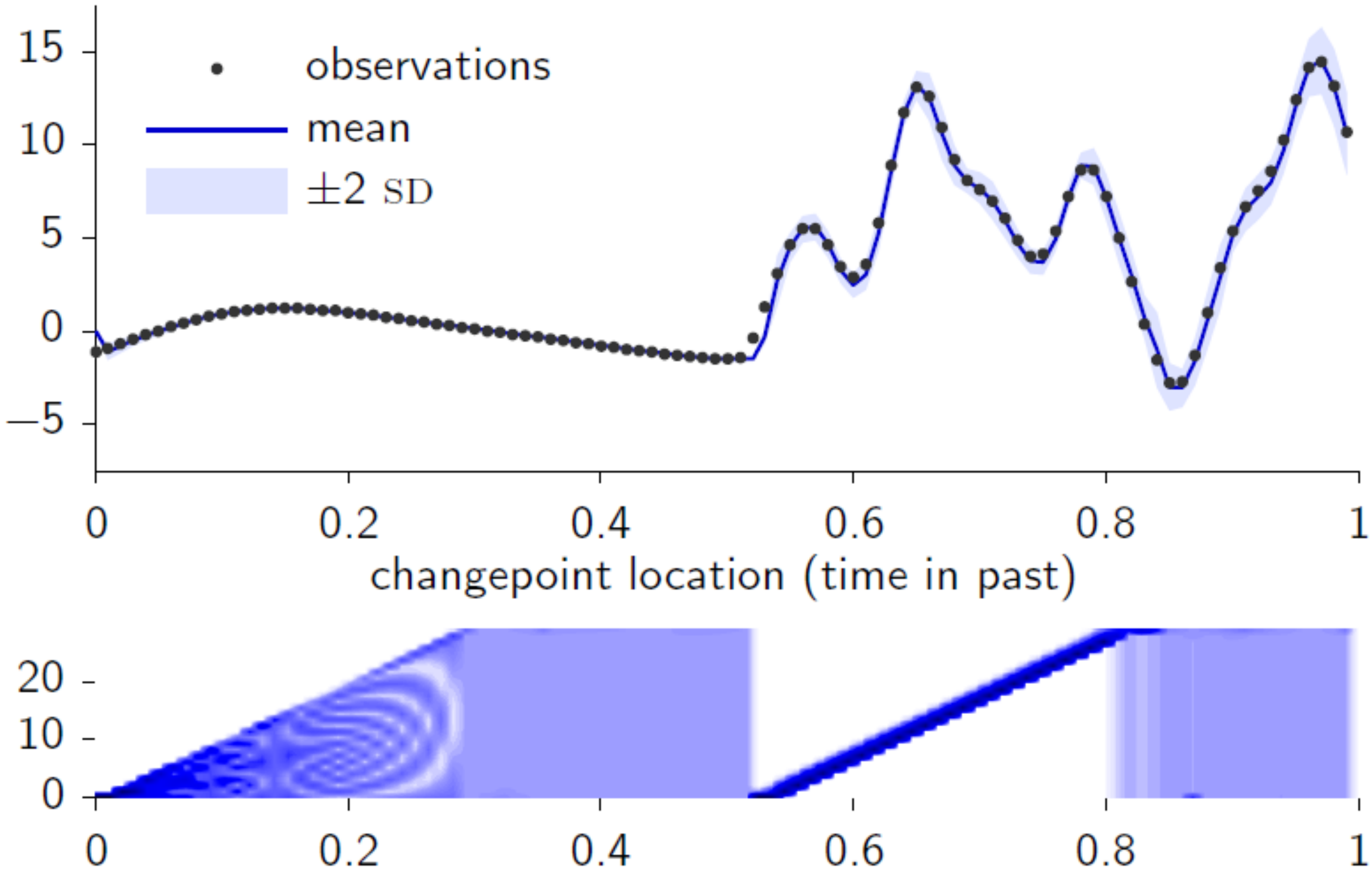
input scale pre-changepoint

changepoint location

input scale post-changepoint

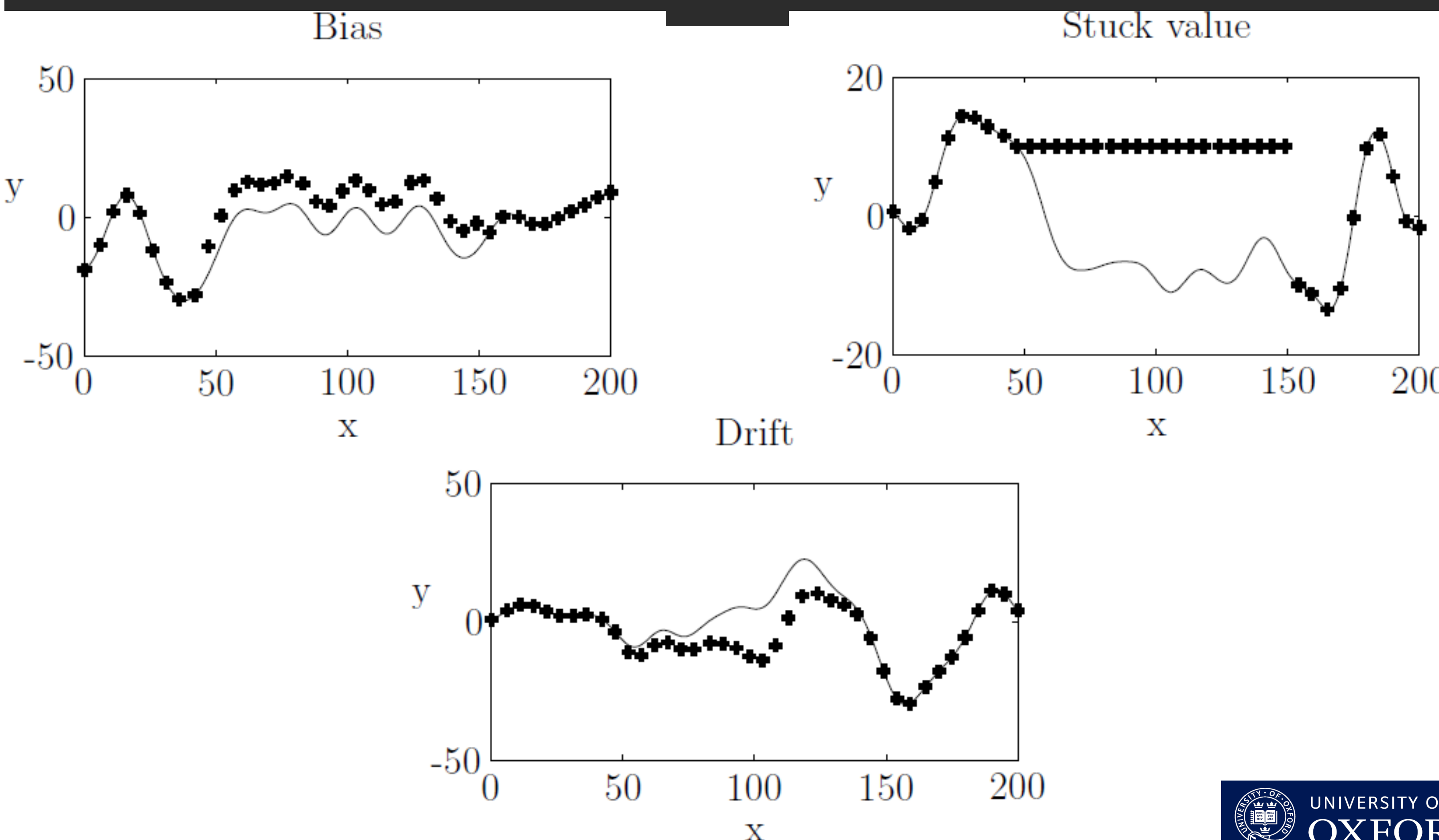
Changepoint detection requires the posterior for the changepoint location hyperparameter.

We can perform both prediction and changepoint detection using Bayesian quadrature.



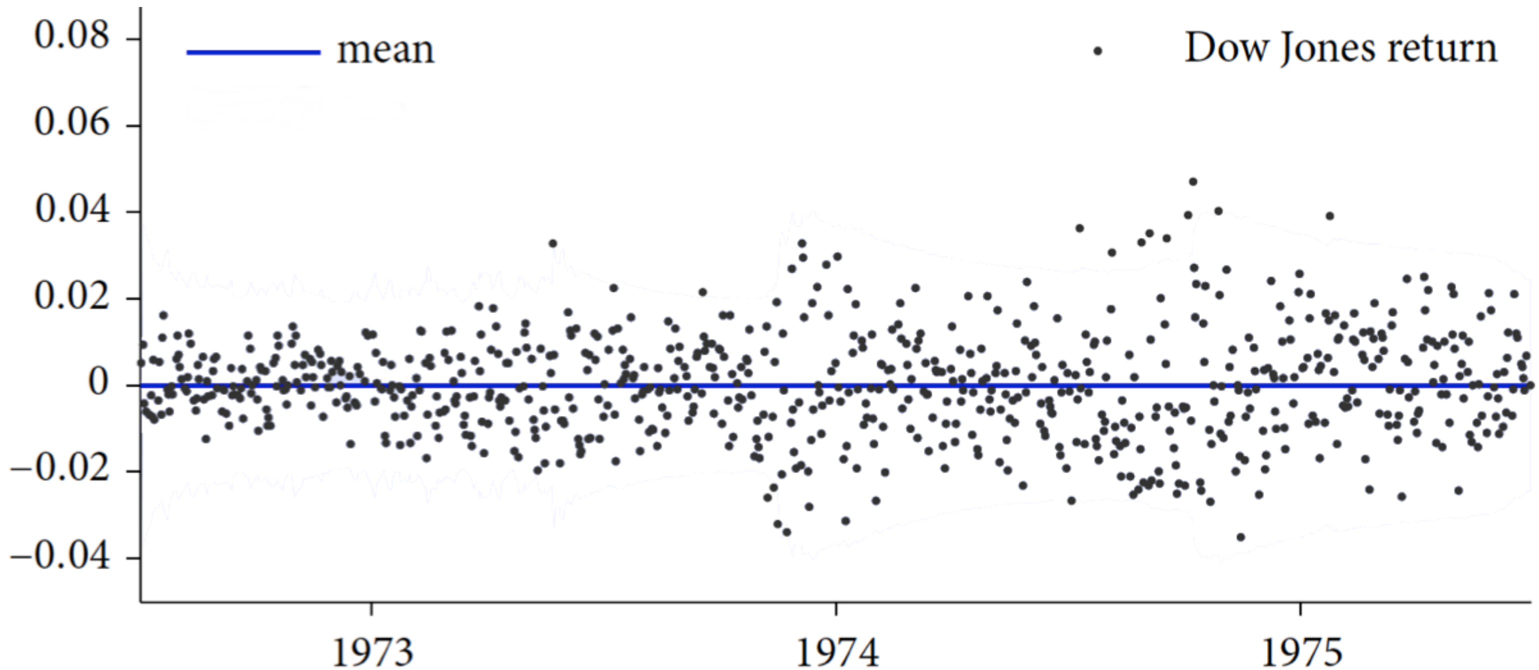
UNIVERSITY OF OXFORD

We can build covariances to accommodate faults, a common challenge in sensor networks.



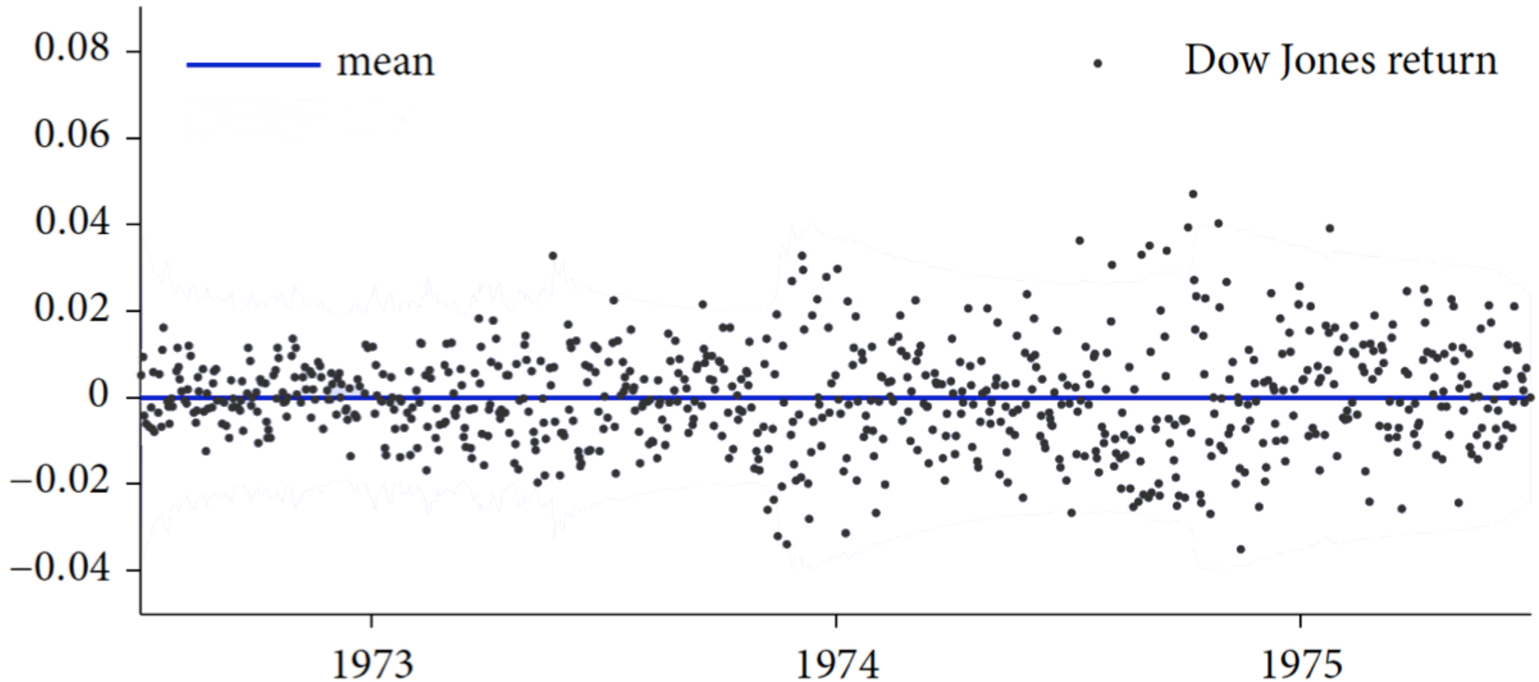
UNIVERSITY OF OXFORD

We use algorithms capable of spotting hidden patterns and anomalies in on-line data.

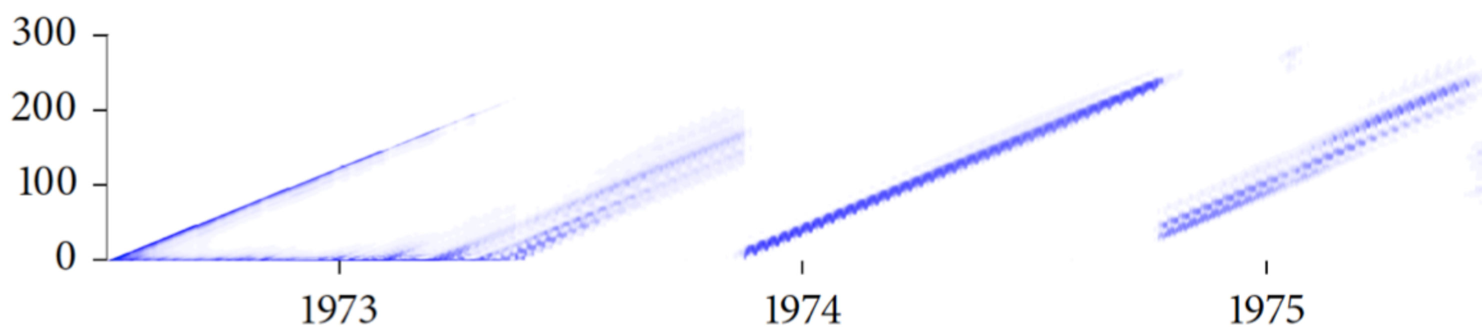


UNIVERSITY OF OXFORD

We identify the OPEC embargo in Oct 1973 and the resignation of Nixon in Aug 1974.

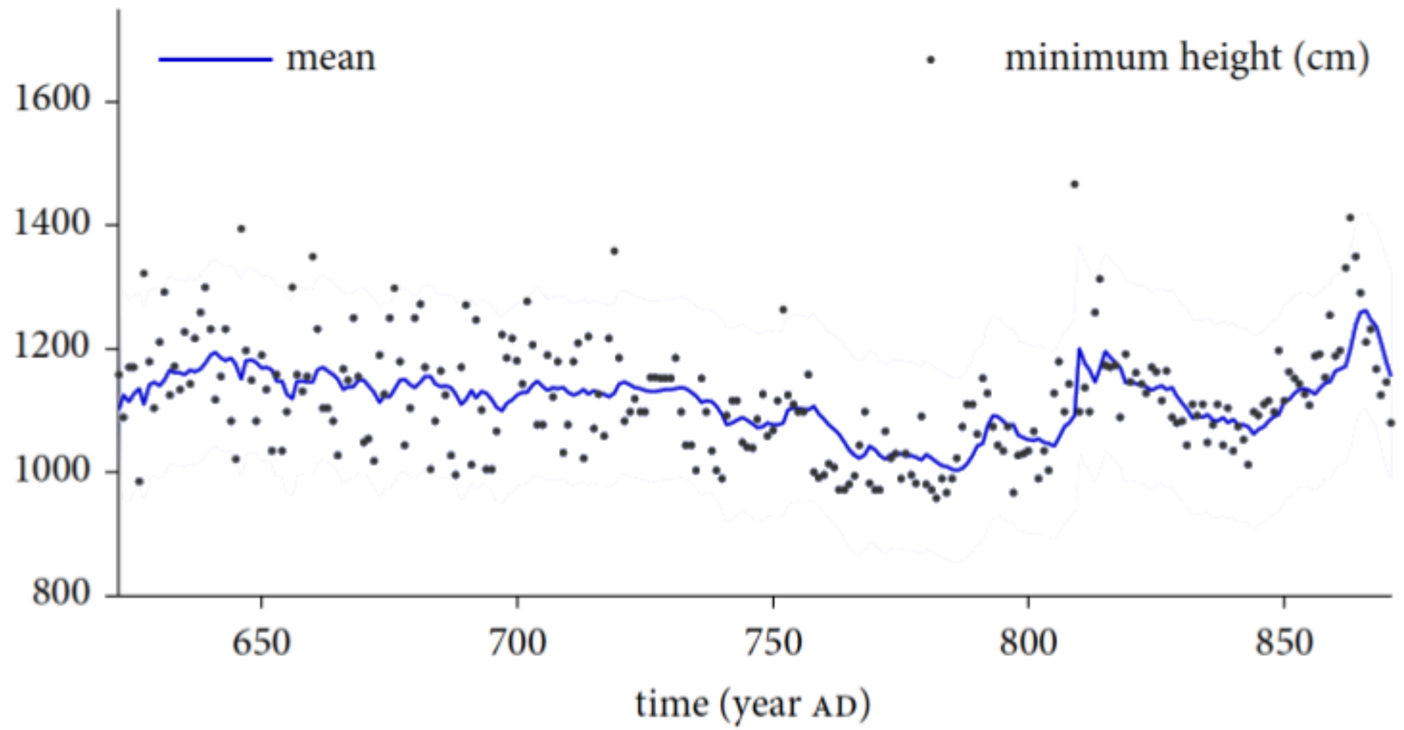


time since last change



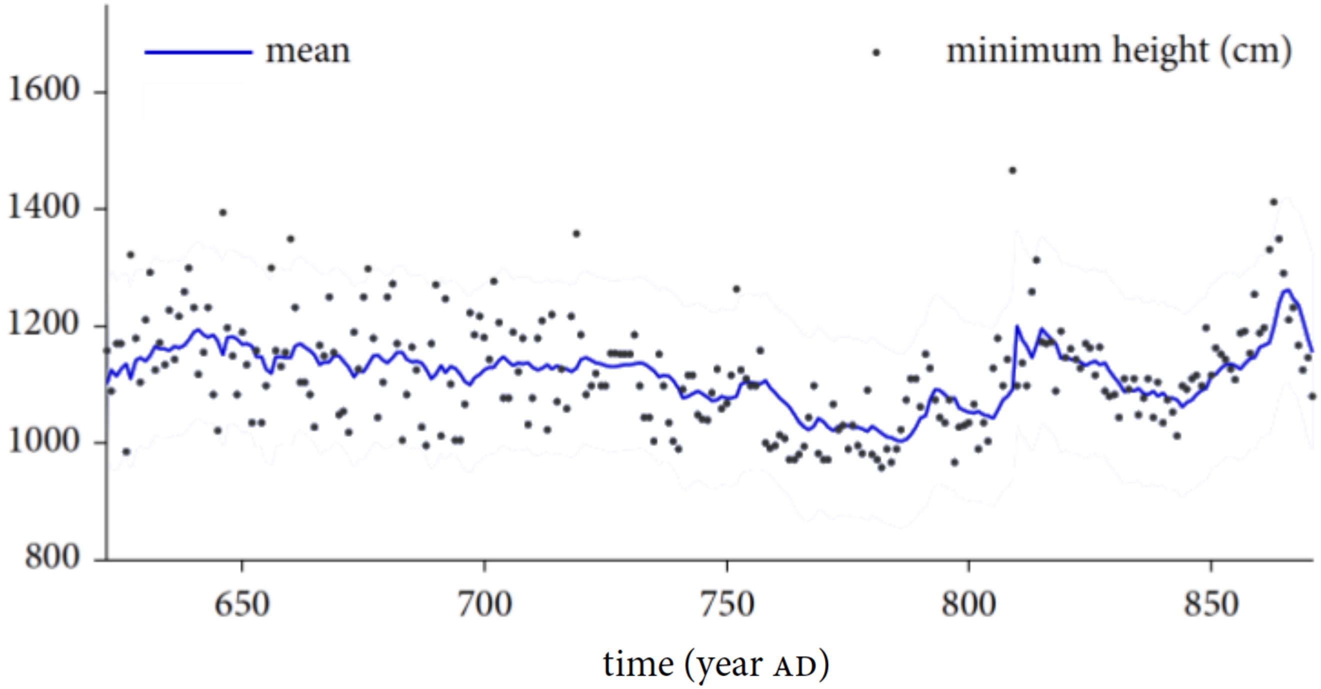
We use algorithms capable of spotting hidden patterns and anomalies in on-line data.

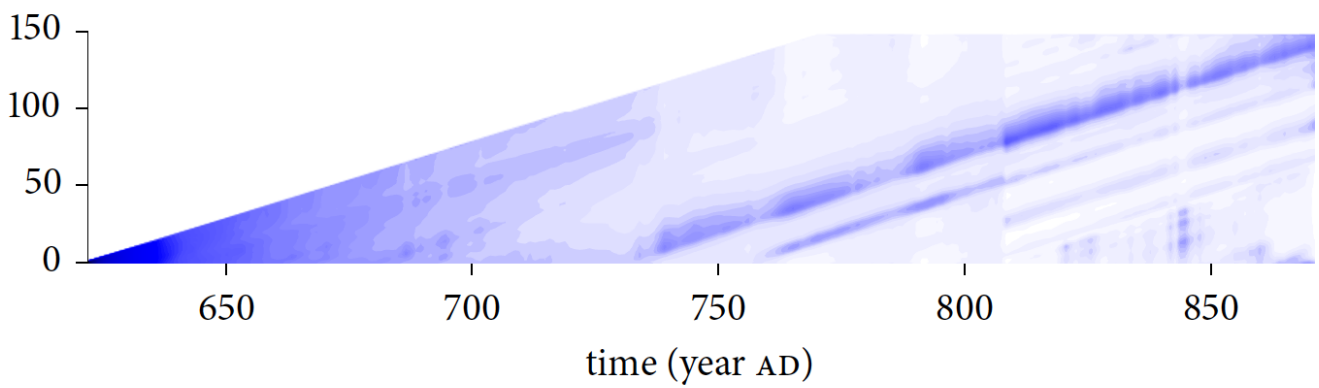
Nile flood levels



UNIVERSITY OF OXFORD

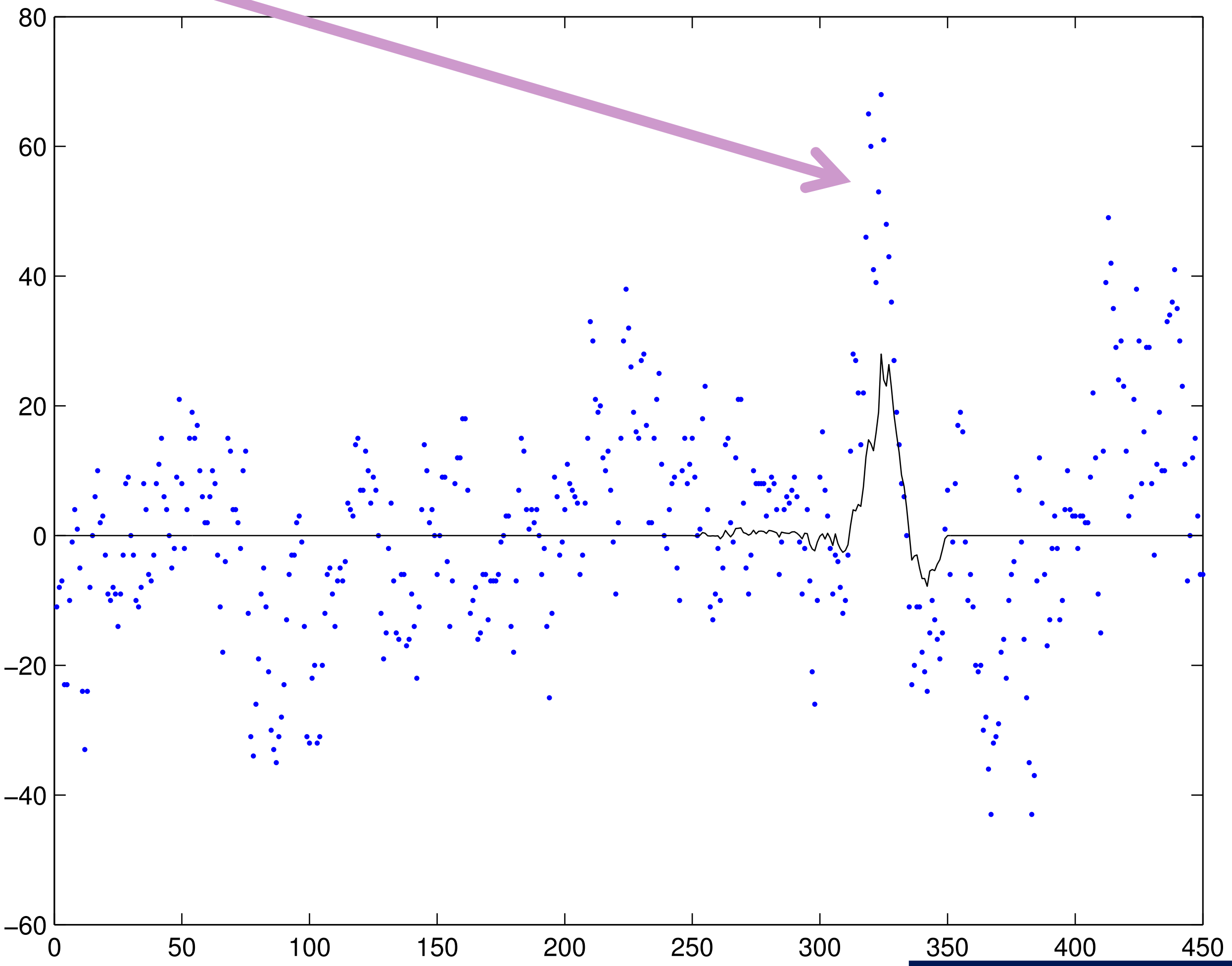
Our algorithm detects a possible change in measurement noise in AD715.





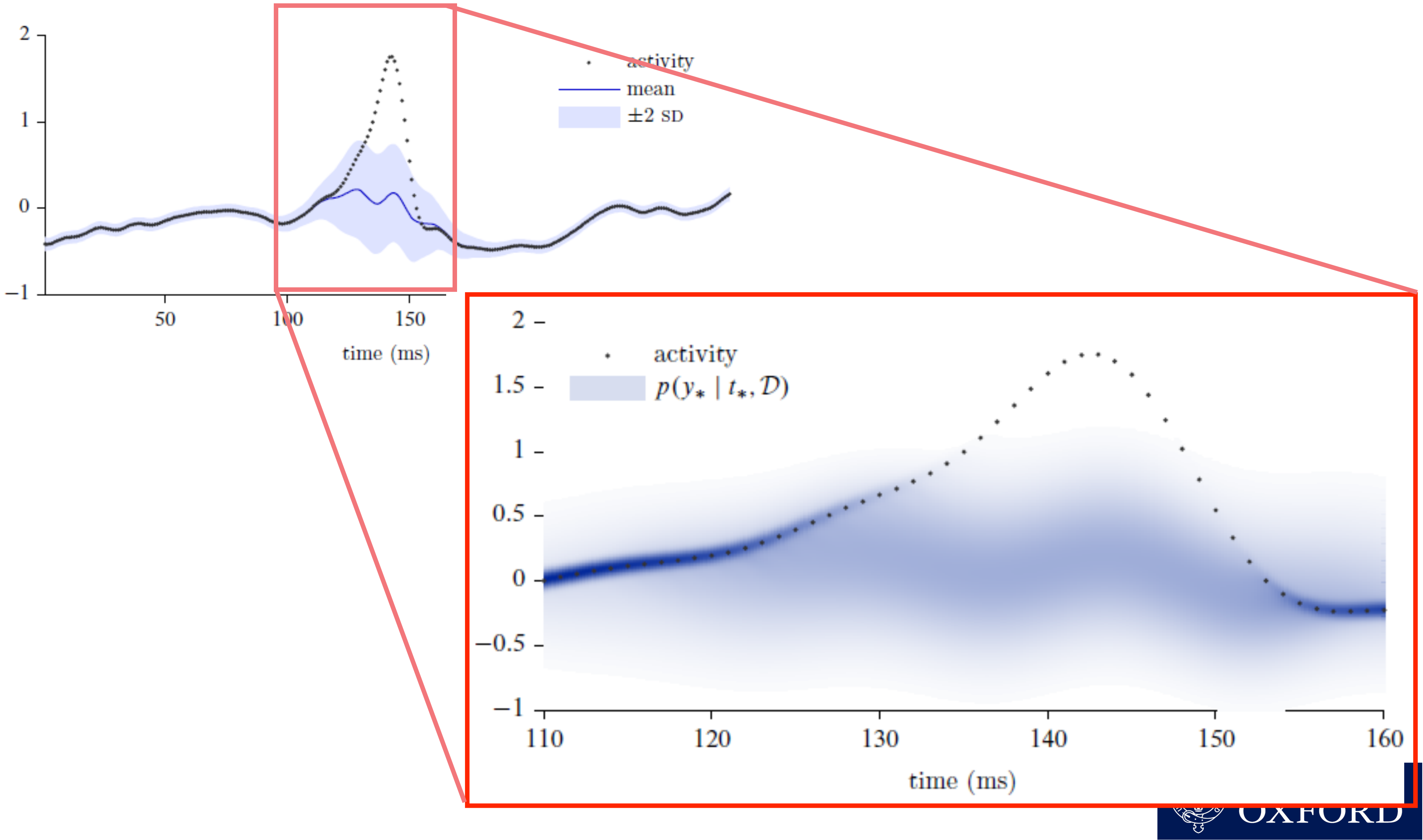
We detect the Nilometer built in AD 715.

Saccades (sudden eye movements) introduce spurious peaks into EEG data.



University of Oxford

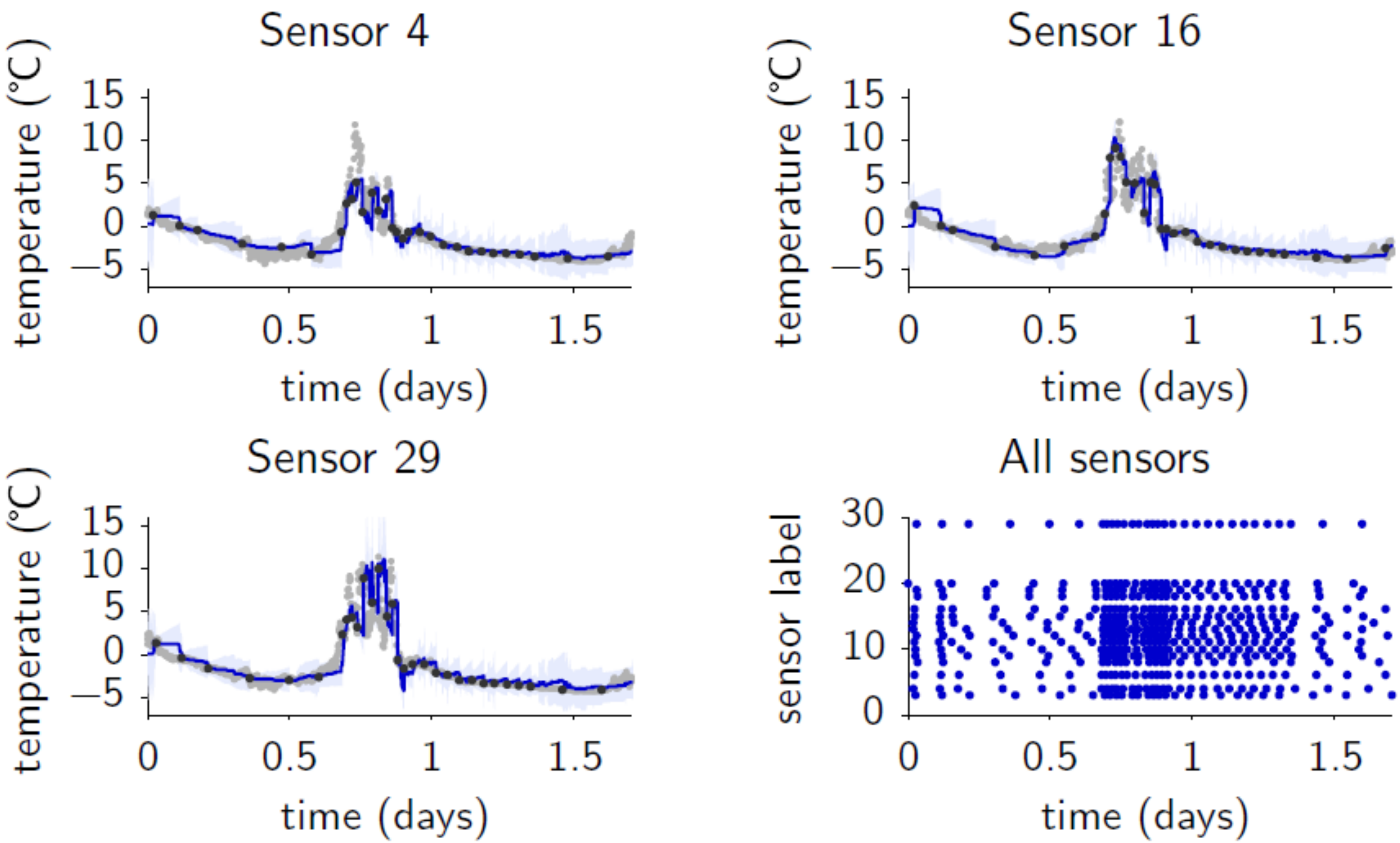
We can perform honest prediction for this complex signal during saccade anomalies.



Wannengrat hosts a remote weather sensor network used for climate change science, for which observations are costly.

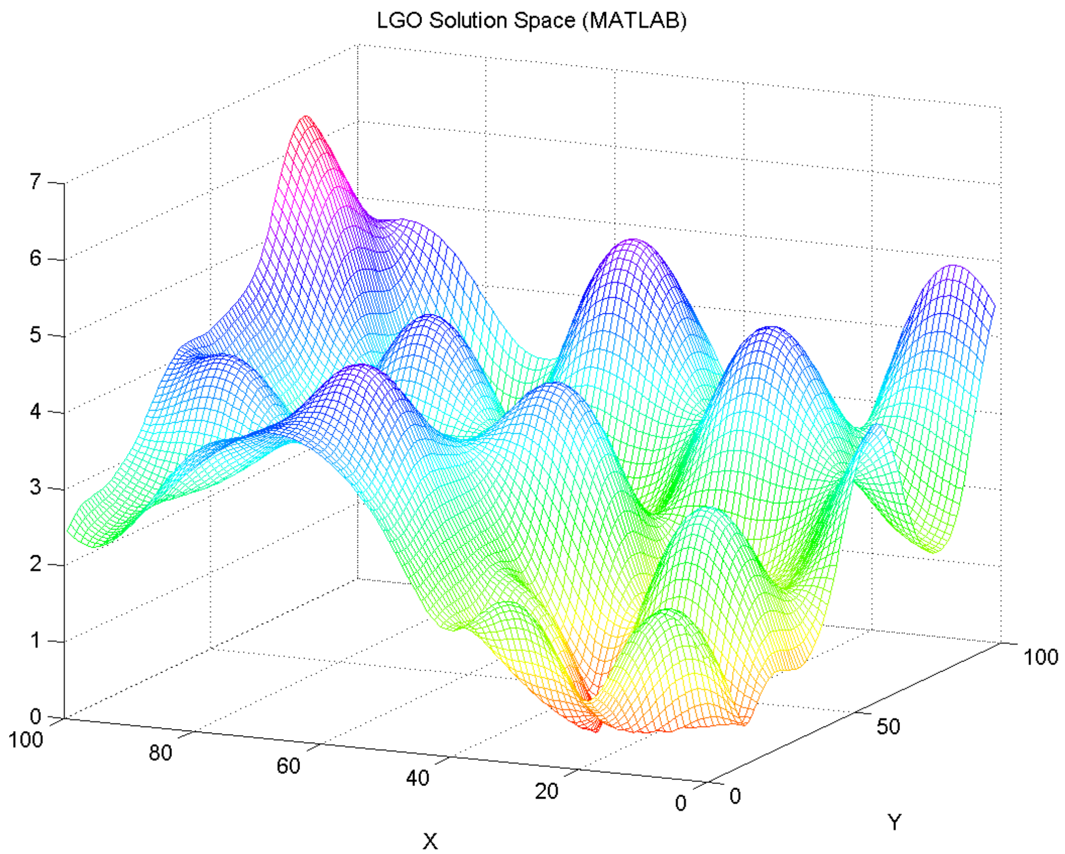
UNIVERSITY OF
OXFORD

Our algorithm acquires more data during interesting volatile periods.

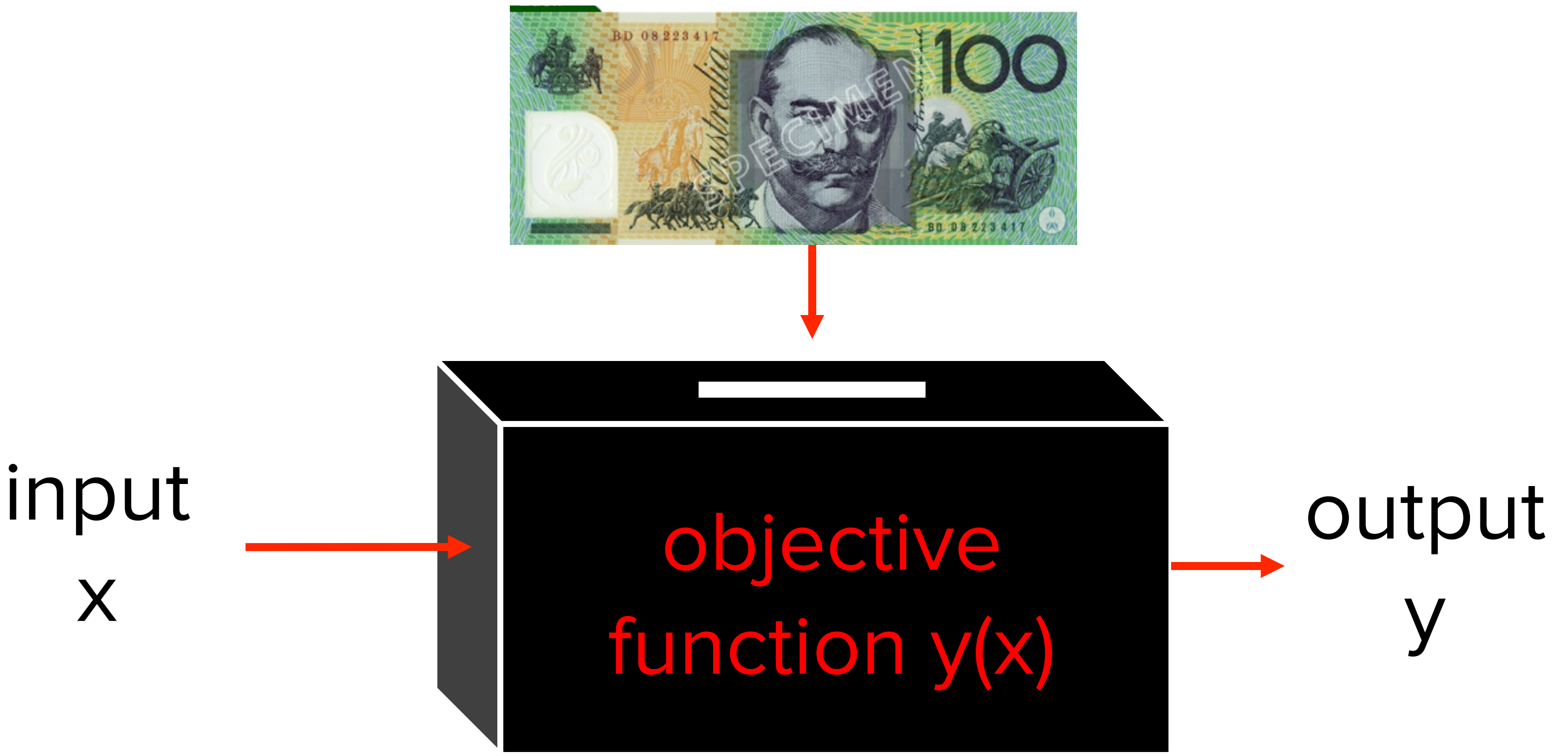


Bayesian quadrature has enabled changepoint detection through efficient model averaging.

Global optimisation considers objective functions that are multi-modal and expensive to evaluate.

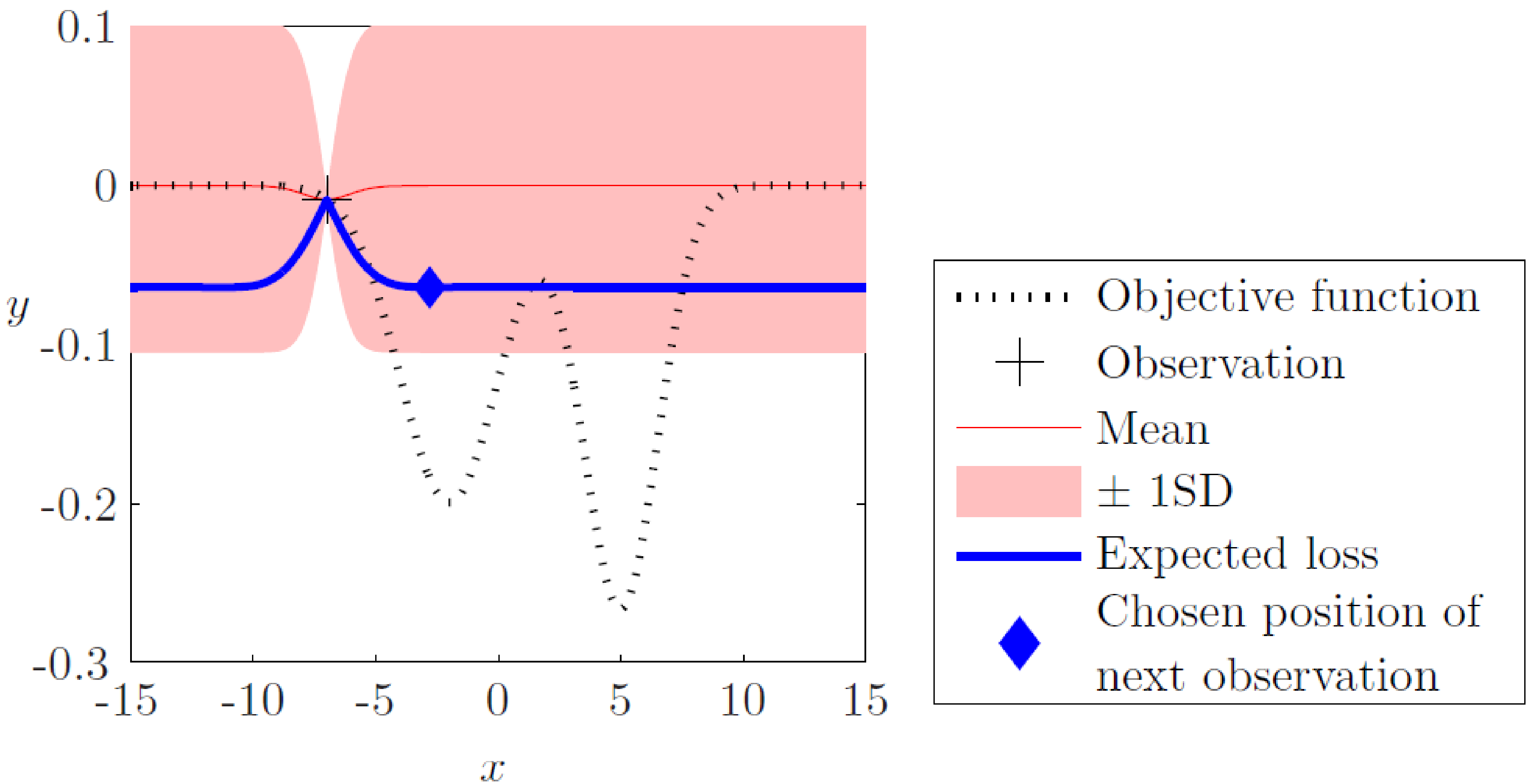


By defining the costs of observation and uncertainty, we can select evaluations optimally by **minimising the expected loss** with respect to a probability distribution.

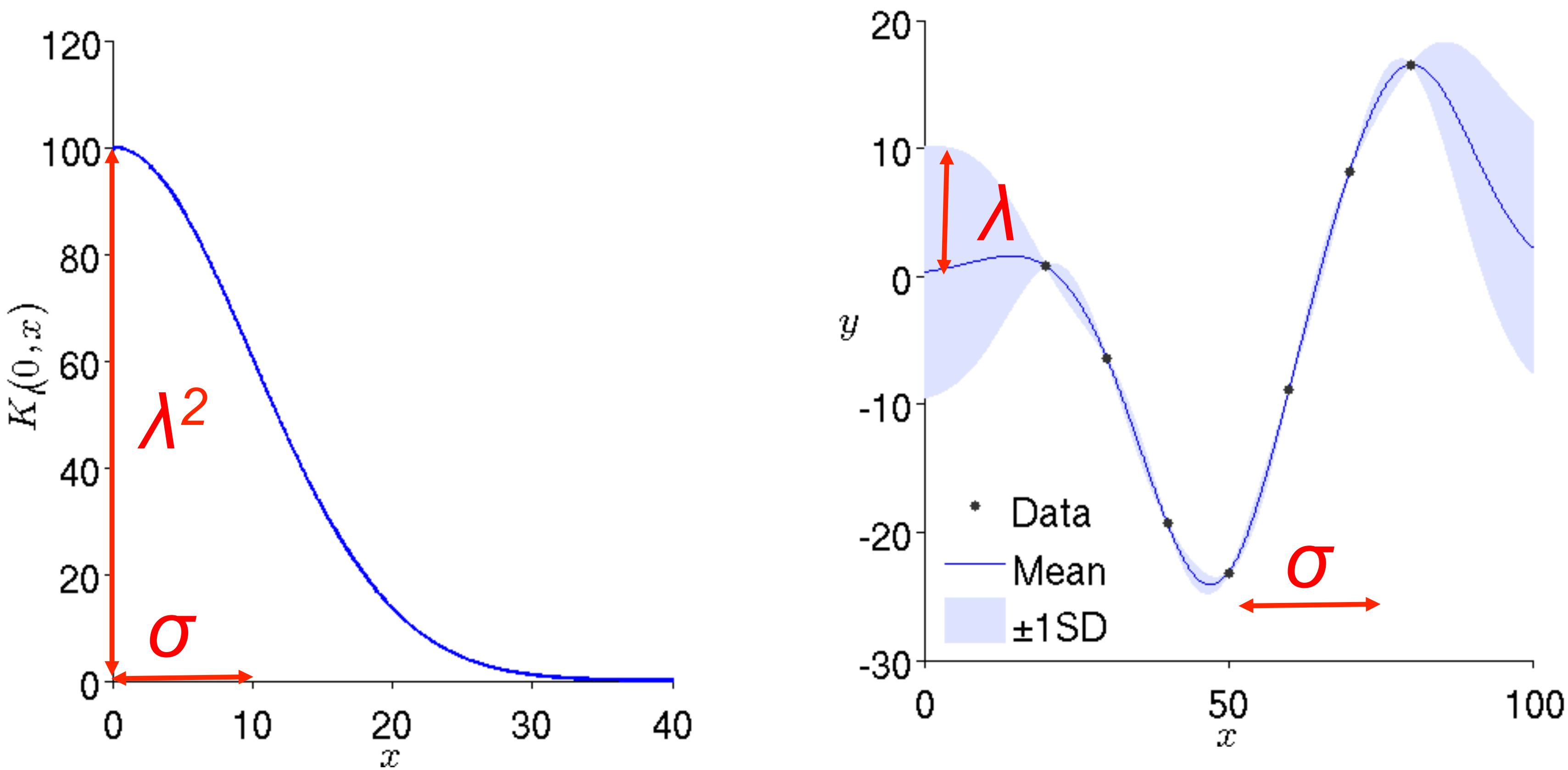


UNIVERSITY OF OXFORD

We choose a Gaussian process as the probability distribution for the objective function, giving a tractable expected loss.

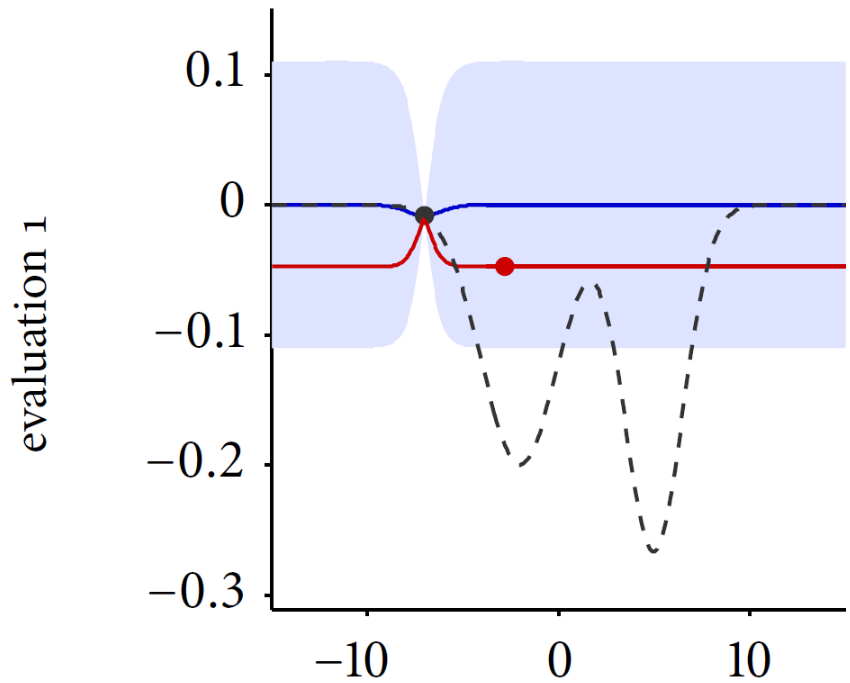


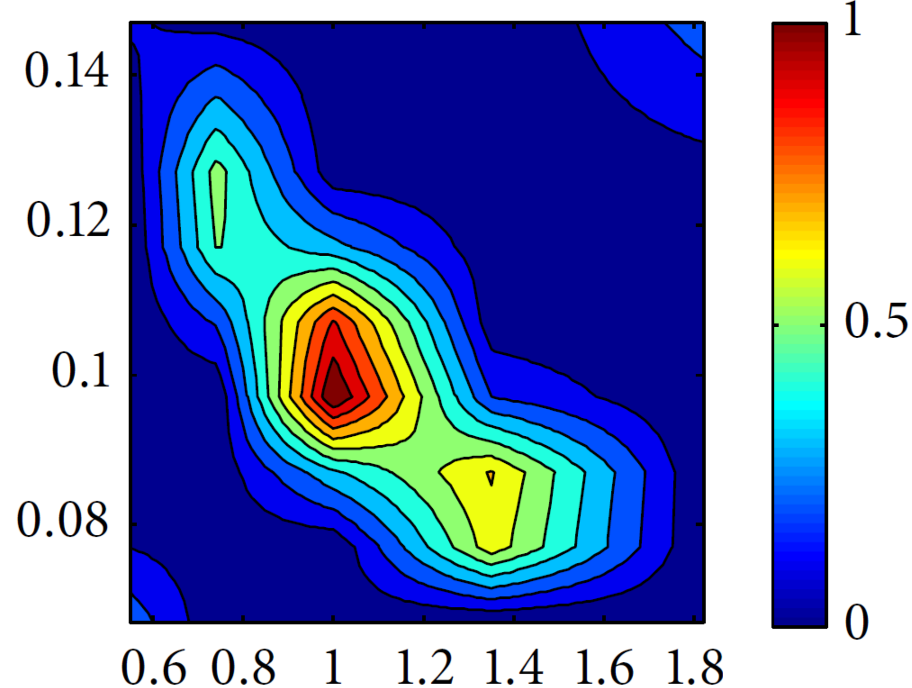
Our Gaussian process is specified by hyper-parameters λ and σ, giving expected length scales of the function in output and input spaces respectively.



UNIVERSITY OF OXFORD

Management of hyperparameters is important for optimisation: we start with no data!



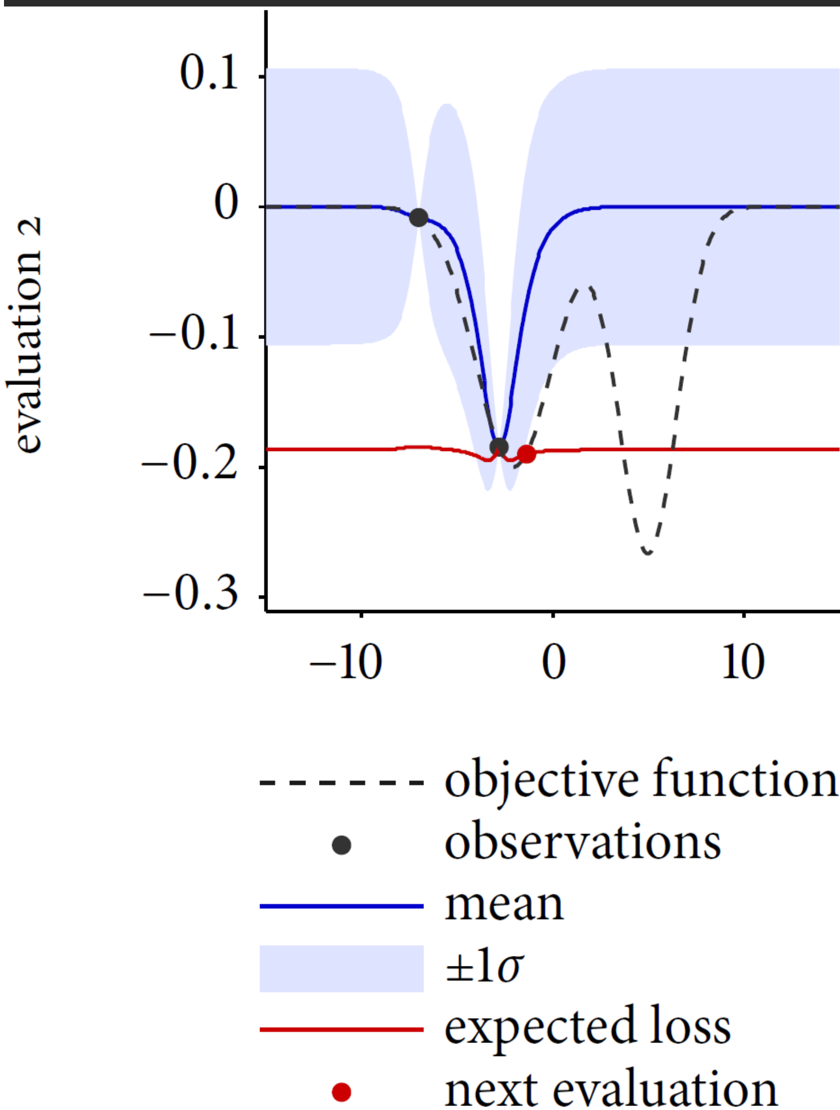


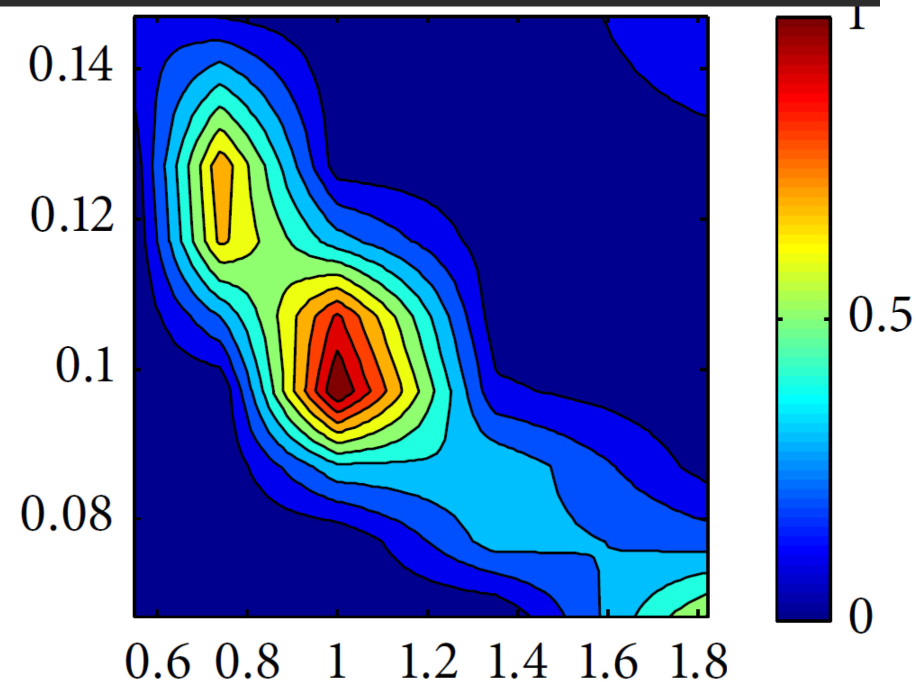
objective function
observations
mean
$\pm 1\sigma$
expected loss
next evaluation

horizontal axis: σ
vertical axis: λ

UNIVERSITY OF OXFORD

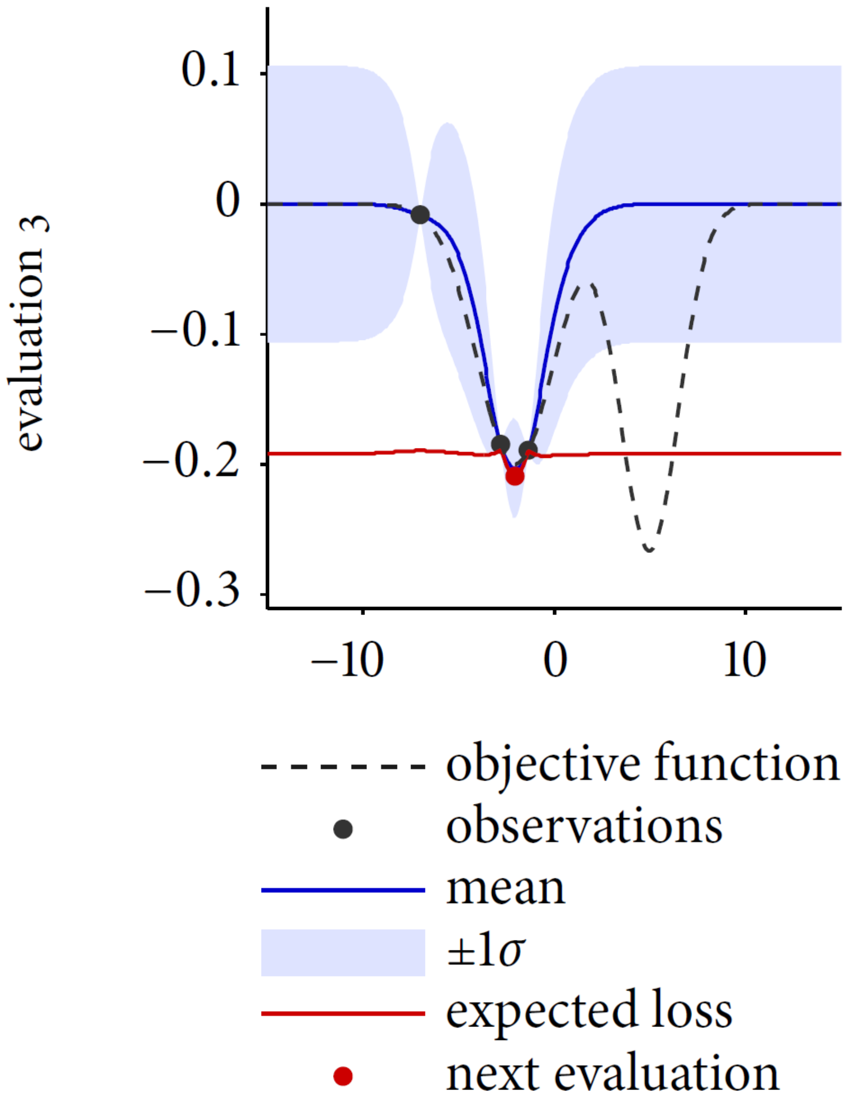
Management of hyperparameters is important for optimisation: we start with no data!

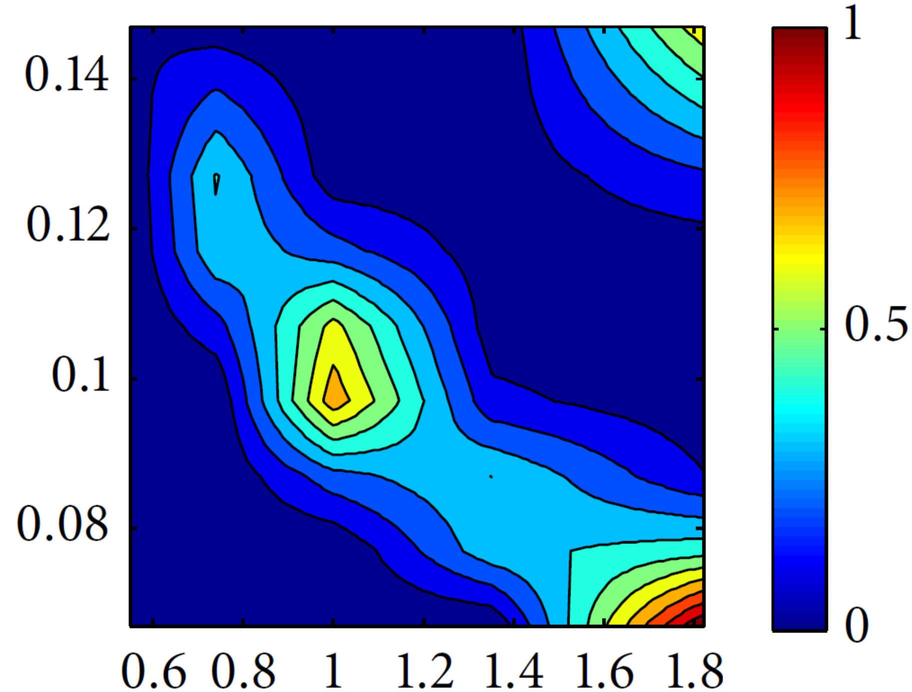




horizontal axis: σ
vertical axis: λ

Management of hyperparameters is important for optimisation: we start with no data!

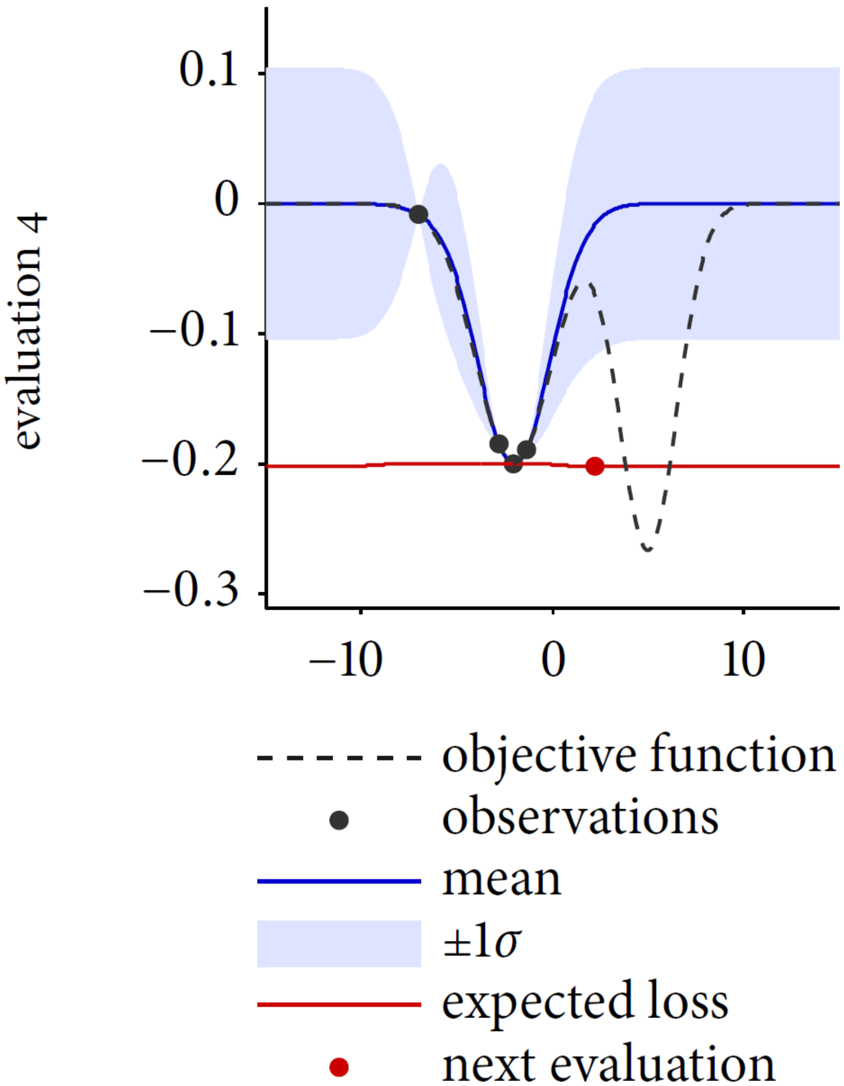


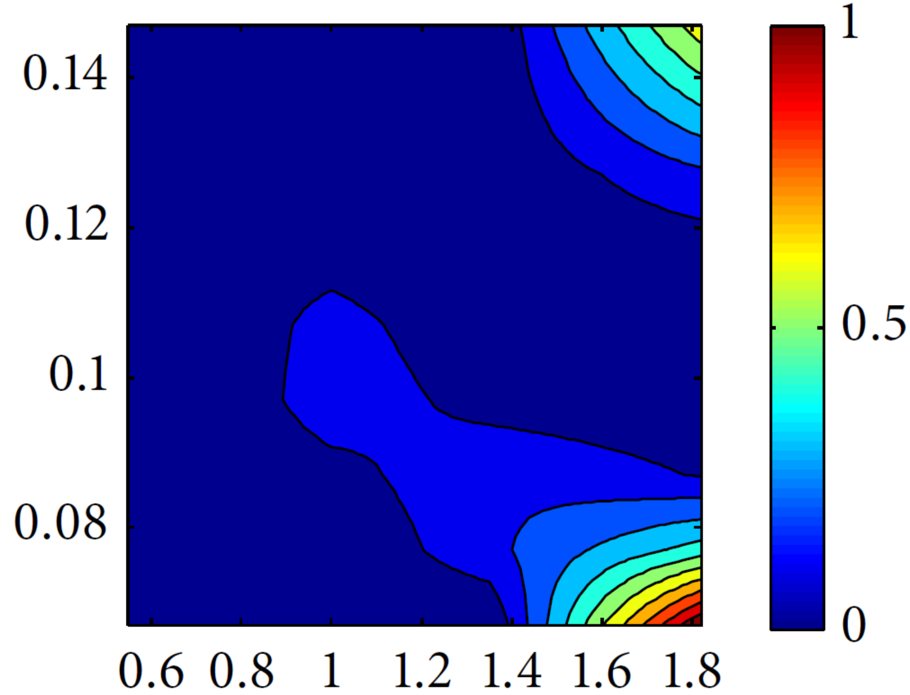


horizontal axis: σ
vertical axis: λ

UNIVERSITY OF OXFORD

Management of hyperparameters is important for optimisation: we start with no data!

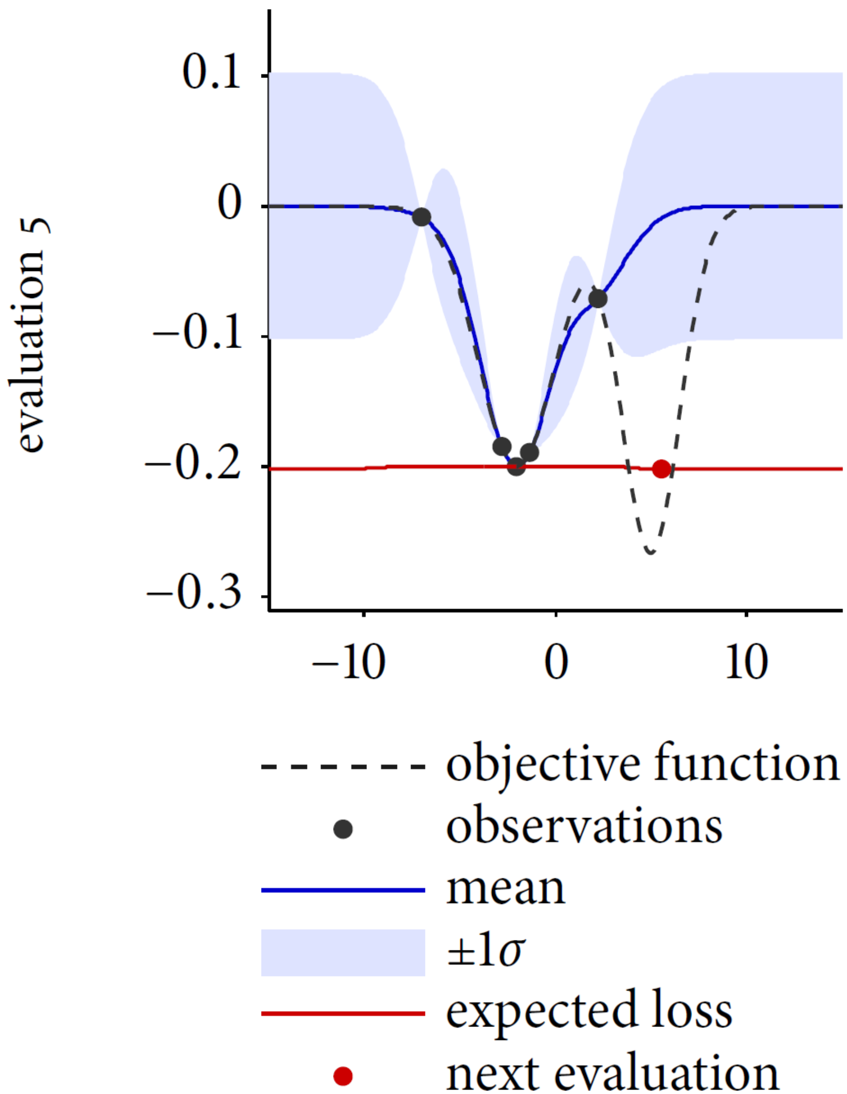


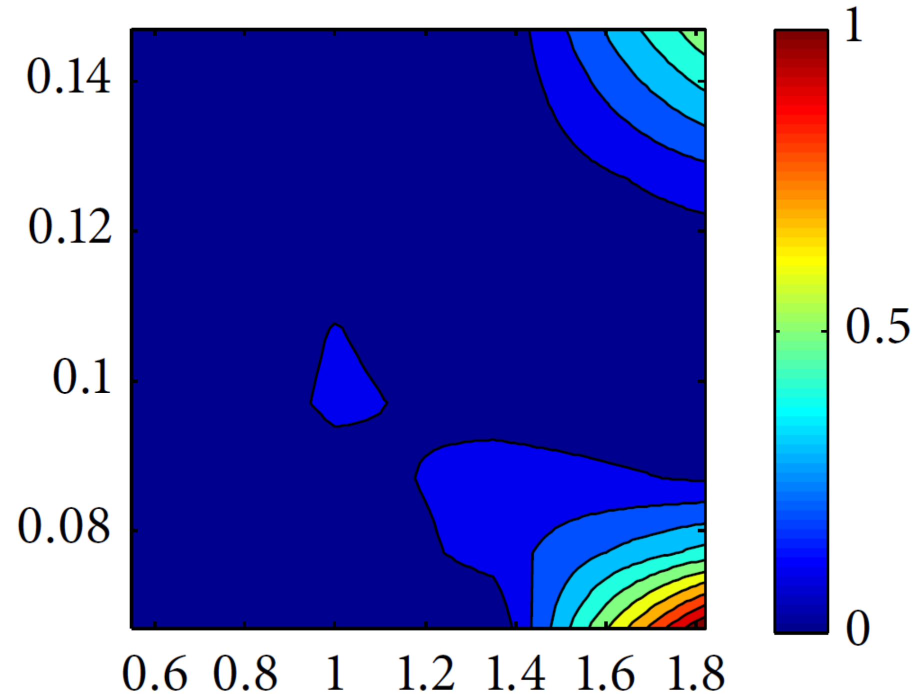


horizontal axis: σ
vertical axis: λ

UNIVERSITY OF OXFORD

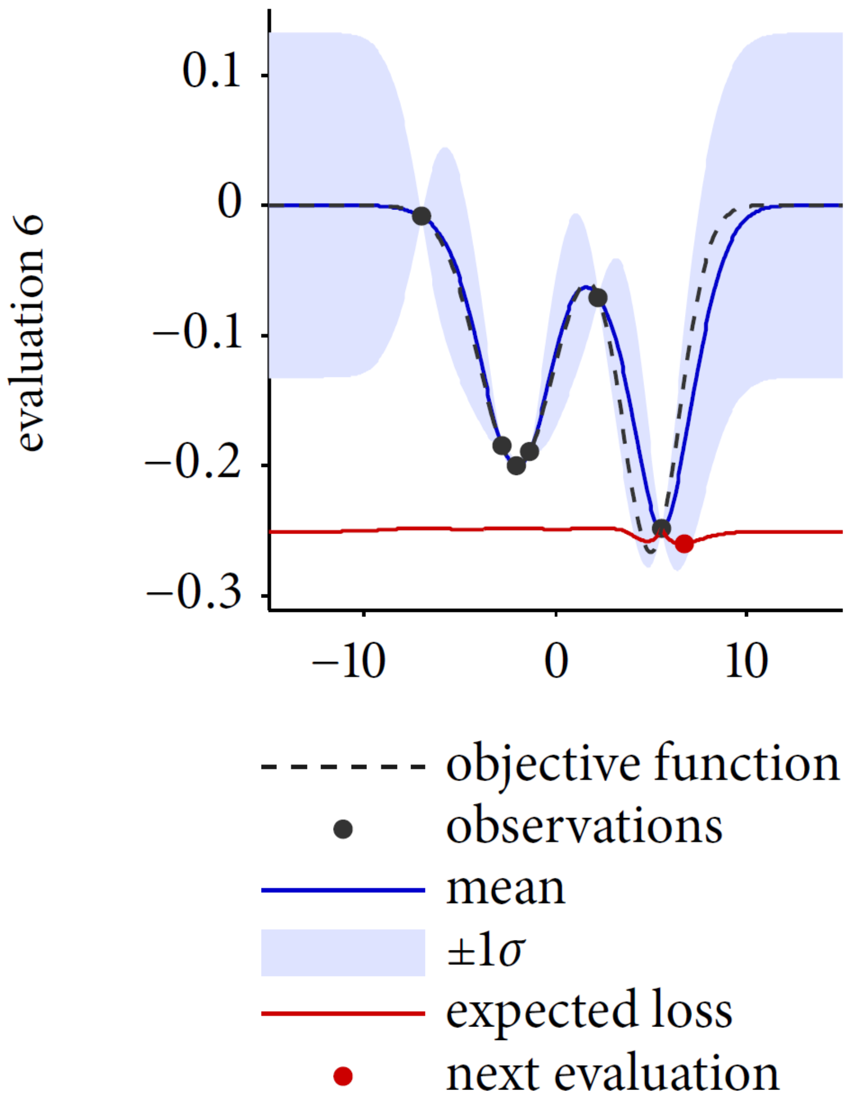
Management of hyperparameters is important for optimisation: we start with no data!

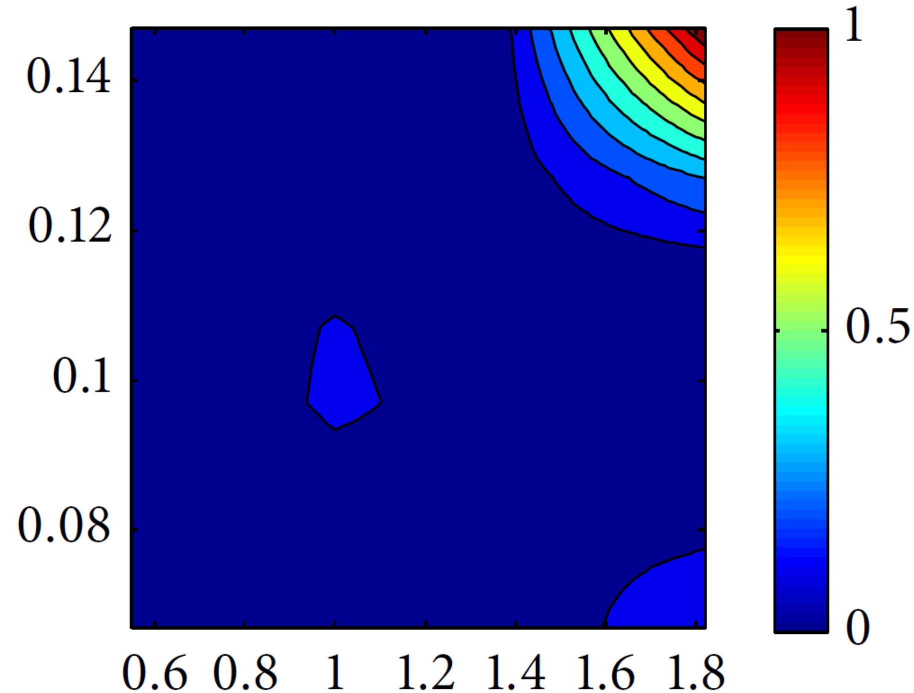




horizontal axis: σ
vertical axis: λ

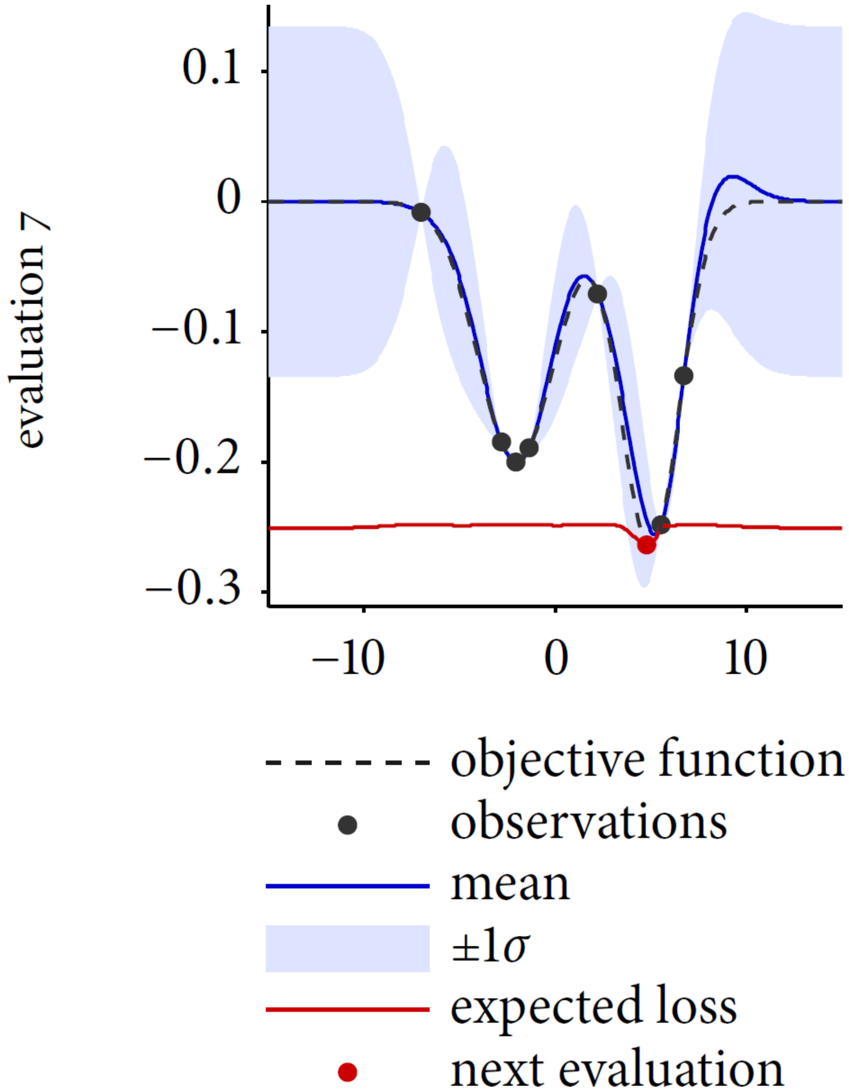
Management of hyperparameters is important for optimisation: we start with no data!

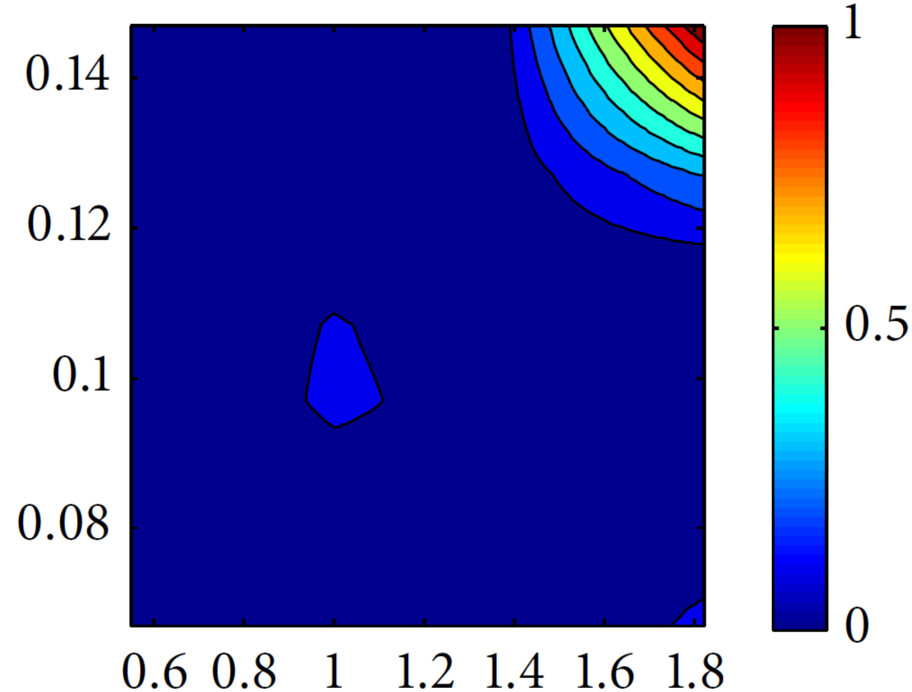




horizontal axis: σ
vertical axis: λ

Management of hyperparameters is important for optimisation: we start with no data!

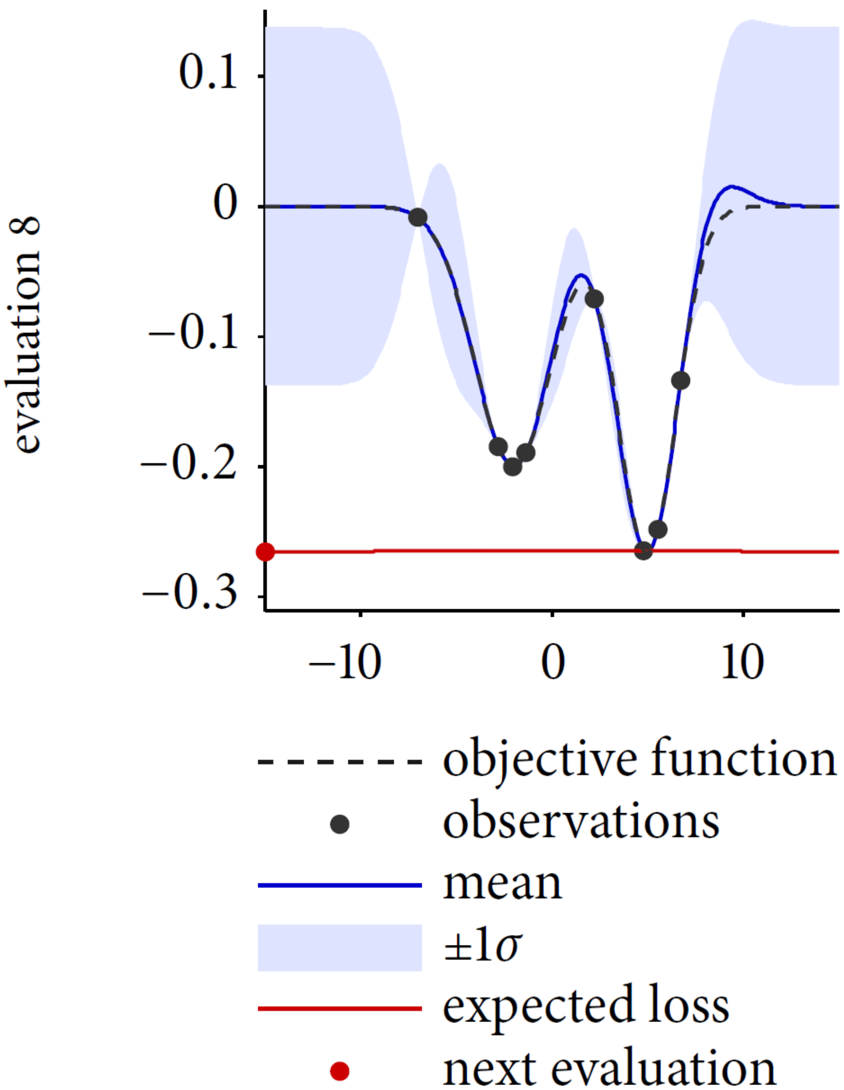


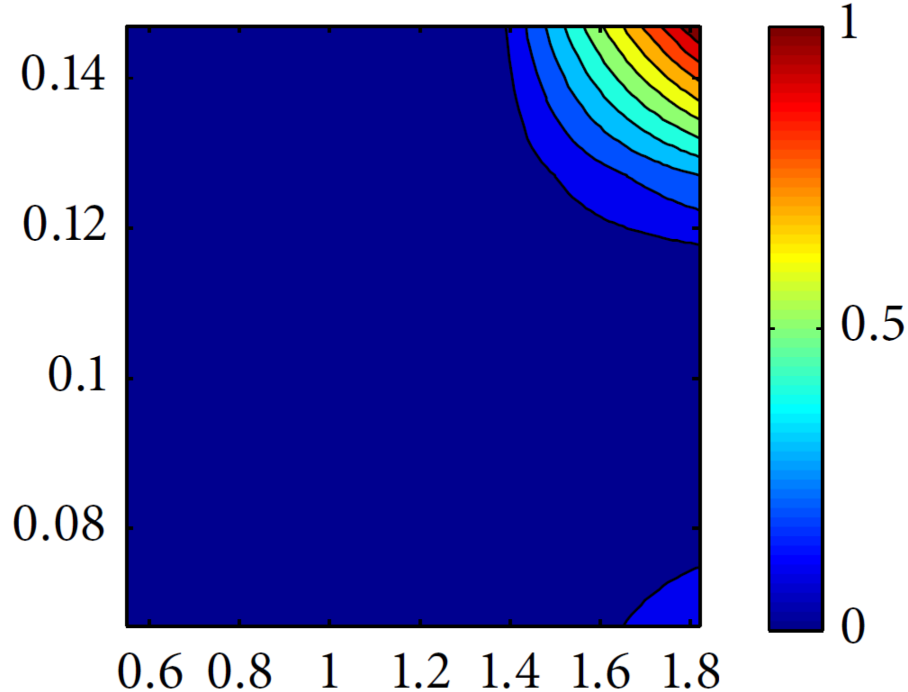


horizontal axis: σ
vertical axis: λ

UNIVERSITY OF OXFORD

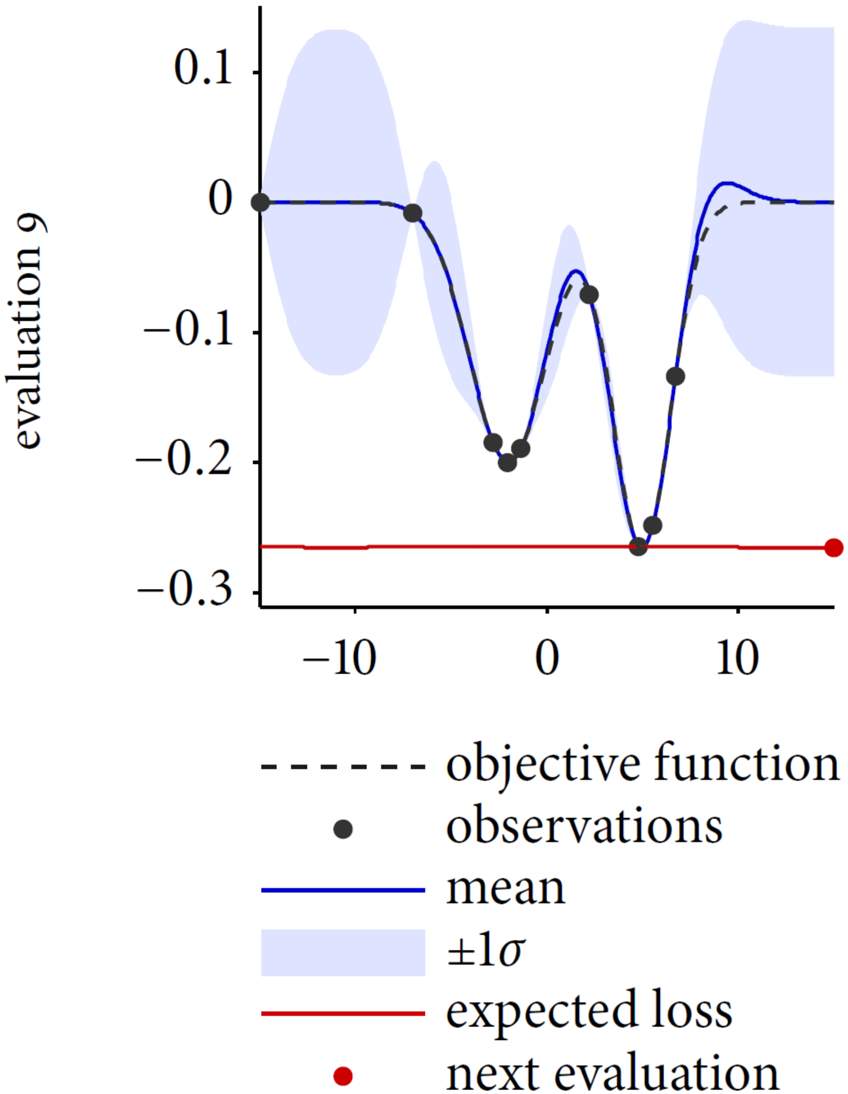
Management of hyperparameters is important for optimisation: we start with no data!

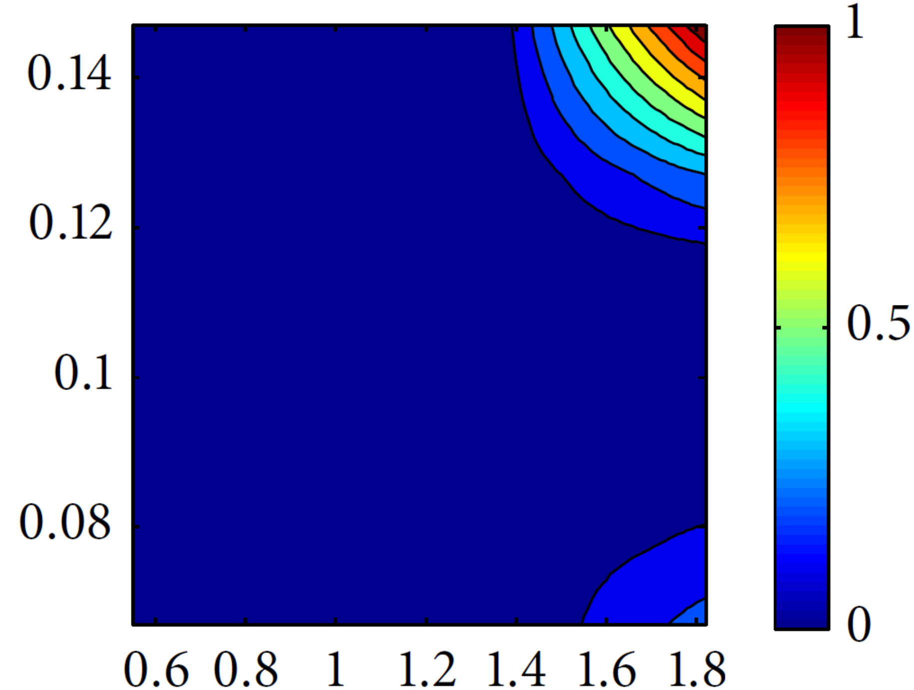




horizontal axis: σ
vertical axis: λ

Management of hyperparameters is important for optimisation: we start with no data!





horizontal axis: σ
vertical axis: λ