

Bayesian Uncertainty Analysis for Complex Physical Models

Michael Goldstein
Durham University *

*Thanks to Basic Technology Initiative for funding the Managing Uncertainty for Complex Models consortium

The General Problem

Physical system has system value $y \in \mathcal{Y}$.

The simulator $f : \mathcal{X} \rightarrow \mathcal{Y}$, is a computer model for the system, where $x \in \mathcal{X}$ are uncertain model parameters.

Partition f into historic values and future values to be predicted, i.e. (f_h, f_p) corresponding to system values (y_h, y_p) .

We are particularly interested in cases where f is slow to evaluate and \mathcal{X}, \mathcal{Y} are high dimensional.

The General Problem

Physical system has system value $y \in \mathcal{Y}$.

The simulator $f : \mathcal{X} \rightarrow \mathcal{Y}$, is a computer model for the system, where $x \in \mathcal{X}$ are uncertain model parameters.

Partition f into historic values and future values to be predicted, i.e. (f_h, f_p) corresponding to system values (y_h, y_p) .

We are particularly interested in cases where f is slow to evaluate and \mathcal{X}, \mathcal{Y} are high dimensional.

We have n evaluations of the simulator at inputs $X \triangleq (x_1, \dots, x_n)$. Denote the evaluations as $F = (f(x_1), \dots, f(x_n))$.

We often have observations on y_h , denoted as z , where $z = y_h \oplus e$, where e is the measurement error, treated as independent of all other quantities.

The General Problem

Physical system has system value $y \in \mathcal{Y}$.

The simulator $f : \mathcal{X} \rightarrow \mathcal{Y}$, is a computer model for the system, where $x \in \mathcal{X}$ are uncertain model parameters.

Partition f into historic values and future values to be predicted, i.e. (f_h, f_p) corresponding to system values (y_h, y_p) .

We are particularly interested in cases where f is slow to evaluate and \mathcal{X}, \mathcal{Y} are high dimensional.

We have n evaluations of the simulator at inputs $X \triangleq (x_1, \dots, x_n)$. Denote the evaluations as $F = (f(x_1), \dots, f(x_n))$.

We often have observations on y_h , denoted as z , where $z = y_h \oplus e$, where e is the measurement error, treated as independent of all other quantities.

The model is not the system!

Simplest way to link model and system: suppose that $y = f(x^*) \oplus \epsilon$, for some x^* , where $\epsilon \perp\!\!\!\perp (f, x^*)$.

(ϵ is model or structural discrepancy error.)

Emulators

The emulator of f is a stochastic representation of the simulator expressing our current uncertainty about the value of $f(x)$ for each x (as updated by the collection of evaluations F).

For example, we might choose

$$f(x) = Bg(x) + r(x)$$

where B is a matrix of unknown coefficients, $g(x)$ is a vector of known functions of x (so $Bg(x)$ expresses global variation in f) and $r(x)$ is a residual process (for example, stationary Gaussian) representing local variation.

Emulators

The emulator of f is a stochastic representation of the simulator expressing our current uncertainty about the value of $f(x)$ for each x (as updated by the collection of evaluations F).

For example, we might choose

$$f(x) = Bg(x) + r(x)$$

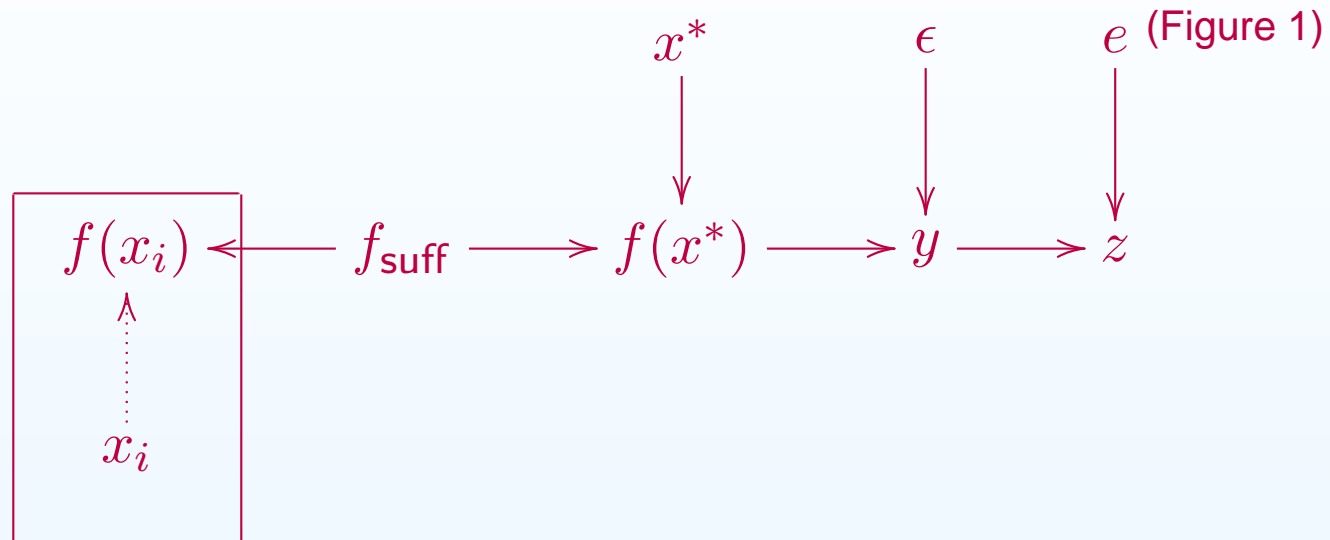
where B is a matrix of unknown coefficients, $g(x)$ is a vector of known functions of x (so $Bg(x)$ expresses global variation in f) and $r(x)$ is a residual process (for example, stationary Gaussian) representing local variation.

Fit multi-output emulator using all our favourite statistical methods.

Useful trick: build the prior for emulating the slow function f by analysing runs from a fast approximation to f and adjusting by a few runs of f itself.

COMMENT Often summarise simulator behaviour by a set of sufficient quantities, f_{suff} (for example f_{suff} might be B).

The graph



Independence graph for this specification

COMMENT Specifying and analysing this graph for high dimensional x and y , with slow to evaluate function f , is very challenging.

Bayes linear approach

The Bayes Linear approach is (relatively) simple in terms of belief specification and analysis, as it is based only on the mean, variance and covariance specification (made directly as primitive quantities - see de Finetti (1974)).

The key equations in the Bayes Linear approach are:

$$E_z[y] = E[y] + \text{Cov}[y, z]\text{Var}[z]^{-1}(z - E[z]),$$

$$\text{Var}_z[y] = \text{Var}[y] - \text{Cov}[y, z]\text{Var}[z]^{-1}\text{Cov}[z, y]$$

where $E_z[y]$ is the expectation for y adjusted by z , and $\text{Var}_z[y]$ is the variance of y adjusted by z

(Goldstein and Wooff (2007) Bayes linear statistics; theory and methods, Wiley)

Bayes linear approach

The Bayes Linear approach is (relatively) simple in terms of belief specification and analysis, as it is based only on the mean, variance and covariance specification (made directly as primitive quantities - see de Finetti (1974)).

The key equations in the Bayes Linear approach are:

$$E_z[y] = E[y] + \text{Cov}[y, z]\text{Var}[z]^{-1}(z - E[z]),$$

$$\text{Var}_z[y] = \text{Var}[y] - \text{Cov}[y, z]\text{Var}[z]^{-1}\text{Cov}[z, y]$$

where $E_z[y]$ is the expectation for y adjusted by z , and $\text{Var}_z[y]$ is the variance of y adjusted by z

(Goldstein and Wooff (2007) Bayes linear statistics; theory and methods, Wiley)

COMMENT

Full Bayes analysis gives full probabilistic output (which is great) but requires full probabilistic specification over all uncertainties (which is not so great).

Also, the full Bayes calculations are much more complicated and non-robust than the Bayes linear counterparts.

Forecasting

For computer models, the mean and variance of $f^* = f(x^*)$ are obtained from the mean function and variance function of the emulator for f , by conditioning on x^* then integrating over the distribution on x^* .

We compute the joint mean and variance of the collection (y, z) directly from the specification of $E[f^*]$, $\text{Var}[f^*]$, $\text{Var}[\epsilon]$, $\text{Var}[e]$ (as $z = y_h \oplus e$, $y = f^* \oplus \epsilon$).

We can now evaluate the forecast, i.e. the adjusted mean and variance for y_p adjusted by z using the Bayes linear adjustment formulae.

Forecasting

For computer models, the mean and variance of $f^* = f(x^*)$ are obtained from the mean function and variance function of the emulator for f , by conditioning on x^* then integrating over the distribution on x^* .

We compute the joint mean and variance of the collection (y, z) directly from the specification of $E[f^*]$, $\text{Var}[f^*]$, $\text{Var}[\epsilon]$, $\text{Var}[e]$ (as $z = y_h \oplus e$, $y = f^* \oplus \epsilon$).

We can now evaluate the forecast, i.e. the adjusted mean and variance for y_p adjusted by z using the Bayes linear adjustment formulae.

This analysis gives us forecasting without a preliminary calibration step, and therefore is tractable even for large systems.

(Bayes linear calibrated forecasting requires more computation but is still tractable.)

COMMENT: We may choose simulator evaluations F to minimise adjusted forecast variance.

History matching

History matching uses observed z to restrict parameter space for x^* .

Aim to rule out regions of $x^* \in \mathcal{X}$ unlikely to give rise to observed z .

Emulator gives $\mathbf{E}[f_h(x)]$ and $\text{Var}[f_h(x)]$ for each x . We calculate the **implausibility** $I_{(i)}(x) = |\mathbf{E}[f_i(x)] - z_i|^2 / \text{Var}[\mathbf{E}[f_i(x)] - z_i]$.

This calculation can be performed univariately, or over each sub-vector for which we have made a joint covariance specification.

History matching

History matching uses observed z to restrict parameter space for x^* .

Aim to rule out regions of $x^* \in \mathcal{X}$ unlikely to give rise to observed z .

Emulator gives $\mathbf{E}[f_h(x)]$ and $\text{Var}[f_h(x)]$ for each x . We calculate the **implausibility** $I_{(i)}(x) = |\mathbf{E}[f_i(x)] - z_i|^2 / \text{Var}[\mathbf{E}[f_i(x)] - z_i]$.

This calculation can be performed univariately, or over each sub-vector for which we have made a joint covariance specification.

The implausibilities are then combined, e.g. by $I_M(x) = \max_i I_{(i)}(x)$, and regions of x with large $I_M(x)$ are judged unlikely to be good choices for x^* .

We refocus our analysis on the ‘non-implausible’ regions of the input space, by resampling and refitting our emulator over such sub-regions and repeating the analysis.

This process is a form of iterative global search aimed at finding **all** choices of x^* which would give good fits to historical data.

Reification

Why should $y = f(x^*) \oplus \epsilon$?

What does simulator f really tell us about system y ?

How do we combine the information about y from simulators, (f, f', \dots) ?

Reification

Why should $y = f(x^*) \oplus \epsilon$?

What does simulator f really tell us about system y ?

How do we combine the information about y from simulators, (f, f', \dots) ?

Consider inputs x as an abstraction from real physical quantities and simulator f as a simplification (through approximations in physics and solution methods) to a much more realistic model f^* with the property that real, physical x^* would satisfy $y = f^*(x^*) \oplus \epsilon$.

Reification

Why should $y = f(x^*) \oplus \epsilon$?

What does simulator f really tell us about system y ?

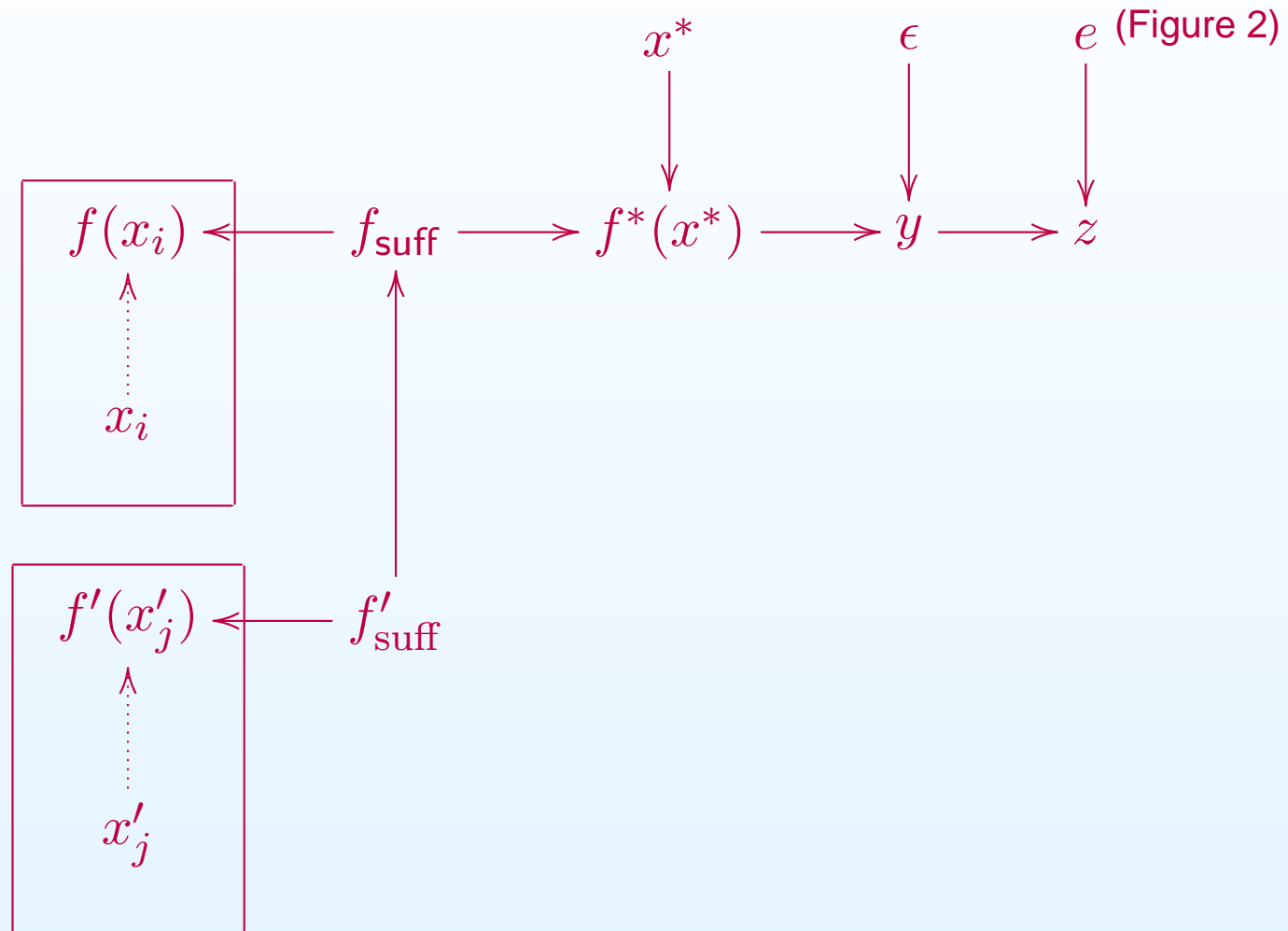
How do we combine the information about y from simulators, (f, f', \dots) ?

Consider inputs x as an abstraction from real physical quantities and simulator f as a simplification (through approximations in physics and solution methods) to a much more realistic model f^* with the property that real, physical x^* would satisfy $y = f^*(x^*) \oplus \epsilon$.

We call f^* the **reified model** (from reify: to treat an abstract concept as if it was real). An actual simulator f is informative for y because f is informative for f^* , as expressed through their linked emulators.

Advantages of reified modelling: straightens out the logic; provides a coherent unification of collections of models; allows us to make inferences about real, physical x^* ; allows us to incorporate qualitative and quantitative knowledge about model deficiencies in our representation of model discrepancy; tractable within Bayes linear approach.

The reified graph



Independence graph showing the reified simulator

References

- P.S. Craig, M. Goldstein, A.H. Seheult and J.A. Smith (1997), Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion), in Gatsonis et al. (eds), *Case Studies in Bayesian Statistics, Volume III*, Springer, pp 37–93
- J.Cumming and M. Goldstein (2010) Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments, in O'Hagan and West (eds), *Handbook of Applied Bayesian analysis*, OUP, to appear
- Bower, Vernon, Goldstein, Benson, Lacey, Baugh, Cole, Frenk, (2010) The parameter space of Galaxy formation, *Monthly notices of the Royal Astronomical Society, Main Journal*, to appear
- M. Goldstein and J.C.Rougier (2007), Reified Bayesian Modelling and Inference for Physical Systems, *Journal of Statistical Planning and Inference*, 139, 1221-1239

References

P.S. Craig, M. Goldstein, A.H. Seheult and J.A. Smith (1997), Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion), in Gatsonis et al. (eds), *Case Studies in Bayesian Statistics, Volume III*, Springer, pp 37–93

J.Cumming and M. Goldstein (2010) Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments, in O'Hagan and West (eds), *Handbook of Applied Bayesian analysis*, OUP, to appear

Bower, Vernon, Goldstein, Benson, Lacey, Baugh, Cole, Frenk, (2010) The parameter space of Galaxy formation, *Monthly notices of the Royal Astronomical Society, Main Journal*, to appear

M. Goldstein and J.C.Rougier (2007), Reified Bayesian Modelling and Inference for Physical Systems, *Journal of Statistical Planning and Inference*, 139, 1221-1239

And do check out the website for the Managing Uncertainty in Complex Models consortium, with lots of resources for dealing with these problems

<http://mucm.group.shef.ac.uk/index.html>