

Bayes Linear Uncertainty Analysis for Large Computer Models

Michael Goldstein
Durham University, UK
Part of the MUCM conspiracy
(thanks to Jonathan Cumming)

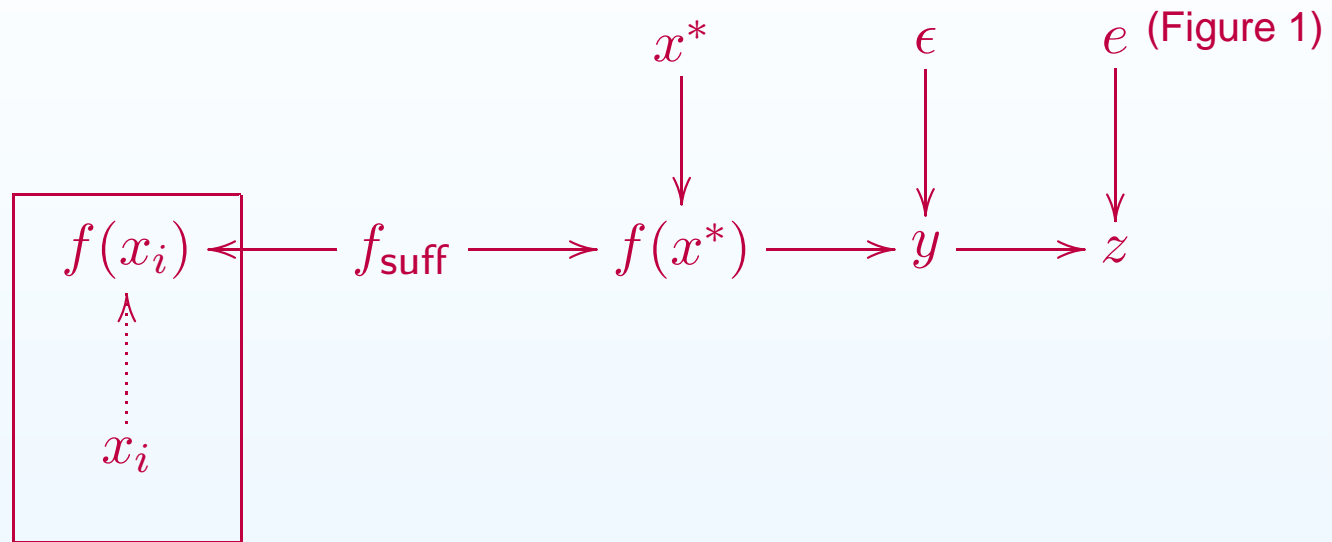
The General Problem

- We start with the physical system and denote the system value as $y \in \mathcal{Y}$. We often have observations on y , denoted as z , where $z = Hy + e$, where H is the incidence matrix, and e is the measurement error which is often treated as independent of all other quantities.
- The simulator is a deterministic complex computer model for the physical system. We denote the simulator as $f : \mathcal{X} \rightarrow \mathcal{Y}$, where $x \in \mathcal{X}$ are uncertain model parameters.
- We have n evaluations of the simulator at inputs $X \triangleq (x_1, \dots, x_n)$, and we denote the resulting evaluations as $F = (f(x_1), \dots, f(x_n))$.
- We often partition these quantities into historic values and future values to be predicted, i.e. $F = (f_h, f_p)$ corresponding to system values (y_h, y_p) .
- The Best Input Approach: We proceed as though there exists a value $x^* \in \mathcal{X}$ where $x^* \perp\!\!\!\perp f$, such that $f^* \triangleq f(x^*)$ separates y from f, x^* .
- Define the model discrepancy as $\epsilon \triangleq y - f^*$, where it follows that $\epsilon \perp\!\!\!\perp (f, x^*)$.
- We are particularly interested in cases where f is slow to evaluate and \mathcal{X}, \mathcal{Y} are high dimensional.

Emulators

- The emulator of f is a stochastic representation of the simulator updated by evaluations of that function at known inputs, expressing our current uncertainty about the value of $f(x)$ for each x .
- 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.
- For simplicity, we sometimes suppose that the simulator behaviour can be summarised in a set of sufficient quantities, f_{suff} , where we judge that $f(x_i) \perp\!\!\!\perp f(x_j) \mid f_{\text{suff}}$, for $x_i \neq x_j$. (For example f_{suff} might be the collection B .) We adjust f_{suff} by the simulator evaluations $(F; X)$.
- The emulator is summarised by the mean and variance functions $\mu(x) \triangleq \mathbf{E}[f(x)]$ and $\kappa(x, x') \triangleq \text{Cov}[f(x), f(x')]$, which we infer from the adjusted f_{suff} .
- Useful trick: build the prior for emulating the slow function f by analysing runs from a fast, simplified approximation to f .

The graph



Independence graph representing the Best Input Approach

Bayes linear Forecasting

- 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), Theory of Probability).
- 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 .

- The mean and variance of f^* are obtained from the mean function and variance function of the emulator for f , namely $\mu(x)$, $\kappa(x, x')$, which are the only features of the emulator that we are required to specify.
- Using these values, we compute the unconditional mean and variance of f^* by first conditioning on x^* and then integrating out with respect to the prior distribution on x^* .

Forecasting without calibration

- Given $E[f^*]$, $\text{Var}[f^*]$, $\text{Var}[\epsilon]$, $\text{Var}[e]$, it is then straightforward to compute the joint mean and variance of the collection (y, z) .
- We can now evaluate 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.
- The approach is likely to be effective when global variation outweighs local variation and the global functional forms $g(x)$ for f_h and f_p are similar.
- We may choose simulator evaluations X to minimise adjusted forecast variance.

Bayes or Bayes Linear?

- Full Bayesian Analysis
 - Gives a joint distribution for (x^*, y) ;
 - Requires full distributional choices for f , ϵ and e , as well as for x^* ;
 - In large problems, tractability often requires a Gaussian distribution for $\{f, \epsilon, e\}$. Even then, large problems can be prohibitively expensive;
 - Clever computation may be able to reduce this expense, if the application is important enough, but big questions about robustness/sensitivity are unavoidable, because of the complexity of the likelihood surface;
- Bayes Linear Analysis
 - Requires full specification for x^* , but only mean and covariance specification for $f(x)$, ϵ , e ;
 - Much more tractable for large problems;
 - Typically analysis is much more stable. Conclusions are derived directly from more genuine uncertainty specifications. Bayes linear interpretive and diagnostic methods used for careful checking of our inferences.

Calibration via Implausibility

- Calibration is learning about x^* using the simulator evaluations and z .
- Using the emulator we can obtain, for each set of inputs x , $E[f_h(x)]$ and $\text{Var}[f_h(x)]$.
- We seek to rule out regions of $x^* \in \mathcal{X}$ which are unlikely to have given rise to observed z .
- To achieve this, we calculate the implausibility:
$$I_{(i)}(x) = |E[f_i(x)] - z_i|^2 / \text{Var}[f_i(x) - z_i]$$
 for each component.
- 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, such as by using
$$I_M(x) = \max_i I_{(i)}(x)$$
, and can then be used to identify regions of x with large $I_M(x)$ as implausible, i.e. unlikely to be good choices for x^* .
- With this information, we can then refocus our analysis on the ‘non-implausible’ regions of the input space, by 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.

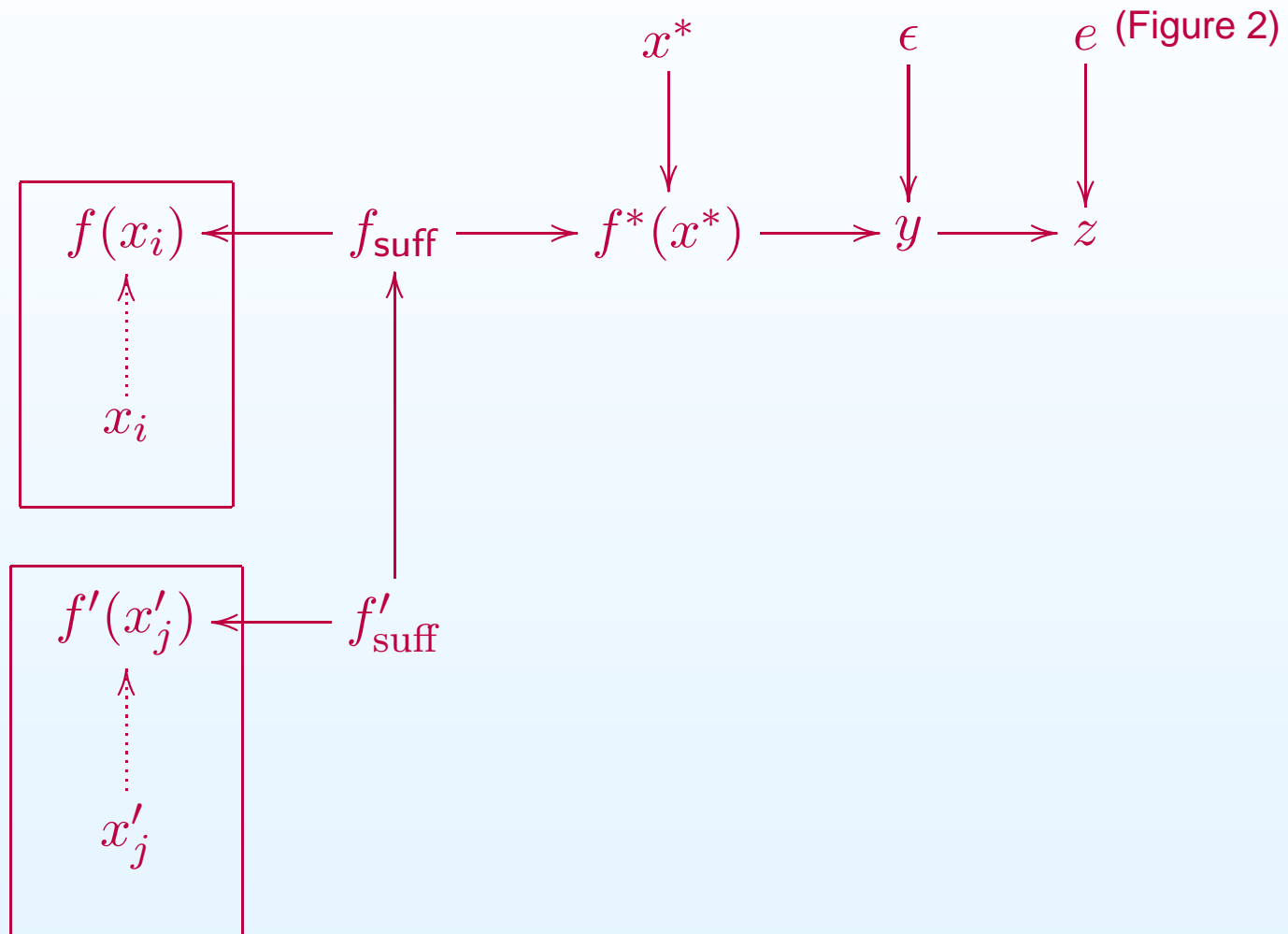
Calibrated forecasting

- Implausibility analysis is used to reduce the input space. We refit the emulator to the reduced subspace. However, some quantities that we wish to forecast may still have substantial local variation, so that further calibration will be informative for the forecast. In calibrated forecasting, we introduce a calibration stage before forecasting (whilst retaining tractability).
- We find the Bayes Linear estimate of the best input x^* by using z . This is $\hat{x} = E_z[x^*] = Wz + U$. We evaluate the simulator f at \hat{x} : $\hat{f} = f(\hat{x})$. This is called the hat run and introduces local knowledge about $f(x^*)$, as opposed to the global knowledge we get from general runs of the simulator.
- We can determine the mean and variance of \hat{f} and the covariance between \hat{f} and (z, y) by repeated use of the emulator for \hat{f} as $\hat{f} = f(\hat{x}) = f(Wz + U) = f(W(f(x^*) + \epsilon + e) + U) = f(W[Bg(x^*) + r(x^*)] + \epsilon + e + U)$, etc.
- Using this information, we can forecast y_p by finding the adjusted mean and variance for y_p using both z and \hat{f} .

Reification

- What does a simulator f really tell us about a physical system y ? Why should we believe that there is a 'best input' x^* ? How do we combine the information about y from a collection of simulators, (f, f', \dots) ?
- Idea: consider our inputs x as an abstraction from real physical quantities and our simulator f as a simplification (through approximations in physics, solution methods, level of detail) to a much more realistic simulator f^* with the property that real, physical x^* would be the best input to f^* , i.e. that $(y - f^*(x^*)) \perp\!\!\!\perp (x^*, f^*, f)$.
- We call f^* the reified simulator (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 (1996), Bayes Linear Strategies for Matching Hydrocarbon Reservoir History. *Bayesian Statistics 5*, 69–95

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

P.S. Craig, M. Goldstein, J.C. Rougier and A.H. Seheult (2001), Bayesian Forecasting for Complex Systems Using Computer Simulators. *JASA*, 96:717–729.

M. Goldstein and J.C. Rougier (2006), Bayes Linear Calibrated Prediction for Complex Systems, *JASA*, forthcoming.

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

M. Goldstein and D.A. Wooff (2007), Bayes linear statistics; theory and methods, Wiley, forthcoming.