Bayes Linear Methods for Multilevel Emulation of Complex Physical Systems

Jonathan Cumming and Michael Goldstein

Department of Mathematical Sciences, Durham University,
Science Laboratories, South Road, Durham, DH1 3LE
e-mail: j.a.cumming@durham.ac.uk

1 Computer Models

- We start with a complex physical system, and express the system behaviour as a vector of system attributes $y = (y_H, y_P)$ where y_H is a collection of historical values, and y_P is a collection of values we wish to predict.
- We have a vector of observations, z_H on the system values y_H, where $z_H = y_H + e_H$, and the observational error $e_H \perp\!\!\!\perp y_H$.
- The simulator is a deterministic computer model which embodies the laws of nature which govern the behaviour of the physical system. We write the simulator as $F(x) = (F_H(x), F_P(x))$.
- The Best Input Approach: We proceed as though that there exists a value x^* where $x^* \perp\!\!\!\perp F$, such that $F^* = F(x^*)$ summarises all the information that the simulator conveys about the system.
- In practice, the simulator F simplifies the physics and approximates the solutions of the resulting equations and so there is a discrepancy between the simulator output and the true values of the system attributes.
- We define model discrepancy as $\epsilon = y - F^*$, where it follows that $\epsilon \perp\!\!\!\perp (F, x^*)$.

2 Hydrocarbon Reservoir Models

- A hydrocarbon reservoir is an underground region of porous rock which contains and oil and/or gas.
- The purpose of the simulator is to model the flows and distributions of fluids within the reservoir over time.
- The model is based on a grid of size $38 \times 27 \times 25$.
- The inputs to the reservoir model include: permeability fields, porosity field, fault transmissibilities, aquifer features, and fluid saturation properties.
- The model outputs comprise time series of attributes of well behaviour, such as pressures and production rates and totals, for each well and injector in the reservoir.
- We consider oil production rate for a 3-year window from a model of the Norwegian Gullfaks field (see Figure 1).
- The simulator takes between 1.5 and 3 hours to evaluate for a given choice of input parameters.

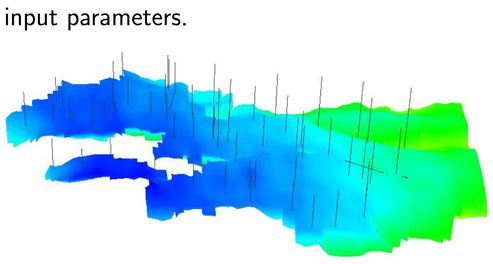

Figure 1: Graphical representation of the Gullfaks model

3 Multiscale Models

- For some problems, an approximate version of the computer model is also available. This coarse simulator $F^c(x)$ can be evaluated in substantially less time and for substantially less cost than the accurate simulator $F^a(x)$, albeit with a lower degree of accuracy.
- The accurate simulator $F^a(x)$ is thus more expensive to evaluate, but more informative about the behaviour of the physical system, such that $y \perp\!\!\!\perp F^c | F^a$.
- Both F^a and F^c are models of the same physical system, therefore we expect strong qualitative similarities between the two models.
- We have n evaluations of the coarse simulator at inputs $X^c = x^c_{1:n}$ denoted $F^c_{[n]}$. Similarly $F^a_{[m]}$ is a collection of m evaluations of F^a at inputs X^a. Since F^c is less expensive to evaluate than F^a then, for a given budget, $n \gg m$.
- Further, we suppose that the simulators can be summarised in a set of sufficient statistics F_{suff} for which we judge that $F(x_i) \perp\!\!\!\perp F(x_j) | F_{\text{suff}}$.
- The relationships between the simulators, the system and the observational data are depicted in the graphical model in Figure 2.

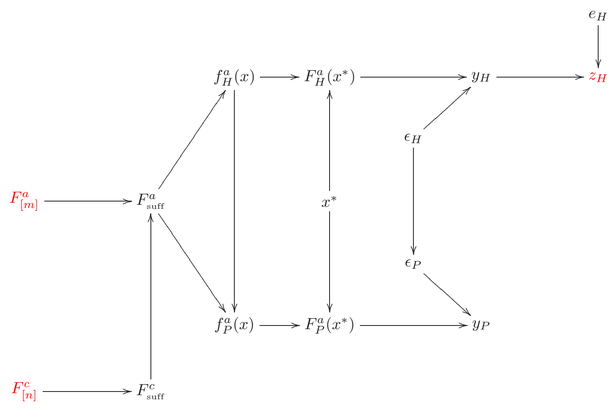

Figure 2: Graphical model of relationships between the simulators, the system and the observational data. Observed quantities are red.

4 Full Bayes or Bayes Linear?

- Full Bayesian Method
  - Gives full posterior probability distributions
  - Requires a prior probability distribution for F and x^*, a probabilistic discrepancy measure relating $F(x^*)$ to y, and a likelihood relating historical data z_H to y_H;
  - Requires strong distributional choices for F, ϵ, e, and x^*;
  - It becomes increasingly difficult to obtain meaningful full prior probability specifications when the dimension of x^* and F grow large;
  - In all but small problems, tractability requires a simple form, such as a Gaussian, for $\{F, \epsilon, e\}$;
  - Even then, the computations for learning from data become technically difficult and computationally expensive;
  - The likelihood surface tends to be highly complex, meaning full Bayes calculations may become highly non-robust.
- Bayes Linear Approach
  - Considers expectation as the primitive quantification of uncertainty;
  - Requires a full specification for x^*, but only mean and covariance specification for $F(x^*)$, y, and F_{suff};
  - Much more tractable for larger problems.
  - The Bayes Linear approach is simpler in terms of belief specification and prediction.
  - The key equations in the Bayes Linear approach are:

$$\mathrm{E}_D(X) = \mathrm{E}(X) + \mathrm{Cov}(X, D)\mathrm{Var}(D)^{-1}(D - \mathrm{E}(D)),$$
$$\mathrm{Var}_D(X) = \mathrm{Var}(X) - \mathrm{Cov}(X, D)\mathrm{Var}(D)^{-1}\mathrm{Cov}(D, X)$$

where $\mathrm{E}_D(X)$ is the expectation for X adjusted by D, and $\mathrm{Var}_D(X)$ is the variance of X adjusted by D.

  - Can be viewed as a simple approximation to a full Bayes analysis, or as the appropriate analysis given a partial specification based on expectation.

5 Emulation

- The simulator is often sufficiently slow to evaluate that we may only make a relatively small number of evaluations of F at different choices of input values, x.
- We express our uncertainty about the value of the function for all input values at which we have not evaluated the simulator via an emulator.
- The emulator, f, of F is a representation of our uncertainty about the simulator updated by evaluations of that function at known inputs.
- Often the global variation of the simulator component F_i is determined by a relatively small subset $x_{[i]}$ of the inputs – the active variables.
- We write the emulator, f^c_i, for component i of the coarse simulator F^c as

$$f^c_i(x) = g_i(x_{[i]})^T \beta^c_i + u^c_i(x)$$

- $B = \{\beta^c_{ij}\}$ are unknown scalars and g_{ij} are known deterministic functions of x. Thus $Bg(x)$ expresses the global variation in F.
- $u^c(x)$ is a weakly stationary process that expresses the residual local variation in $F^c(x)$ not captured by the trend, and where $u^c \perp\!\!\!\perp B$.
- Since n is large, we determine appropriate forms and values for $\{g(\cdot), B, u^c(x)\}$ from the model runs $F^c_{[n]}$.
- When the variation in the residual is small and the number of active variables is not large, this active variable approach enormously reduces the dimension of the computations whilst only having a small impact on the accuracy of our results.
- For each well in the Gullfaks model, we emulate oil production rate at 12 time points corresponding to y_H.

6 Multilevel Emulation

- Using the coarse simulator runs and the methods described above, we construct a coarse emulator $f^c(x)$ of $F^c(x)$.
- We use $f^c(x)$ as a basis for constructing an appropriate prior specification for the accurate emulator $f^a(x)$ of $F^a(x)$.
- We write the multilevel accurate emulator as

$$f^a_i(x) = g_i(x_{[i]})^T \beta^a_i + \beta^a_{ui} u^c_i(x) + u^a_i(x)$$

- We link the coefficients of the two emulators by specifying $\beta^a_{ij} = \rho_{ij}\beta^c_{ij} + \gamma_{ij}$, where ρ and γ are unknown scalars and $\rho \perp\!\!\!\perp \gamma$.
- We specify a prior expectation and covariance for $(\rho_{ij}, \gamma_{ij}, u^a_i(x))$ including temporal correlations between coefficients between emulators of output at the same well but separated through time.
- We carefully choose a small number of additional design points at which to evaluate the fine emulator, exploiting active variable structure via the border-block decomposition.
- Then update $(\beta^a_{ij}, \beta^a_u, u^c)$ by a small number of model evaluations to obtain the adjusted accurate emulator.
- For the Gullfaks model, the accurate model is the original simulation and the coarse model is obtained by vertically coarsening the gridding of the reservoir.

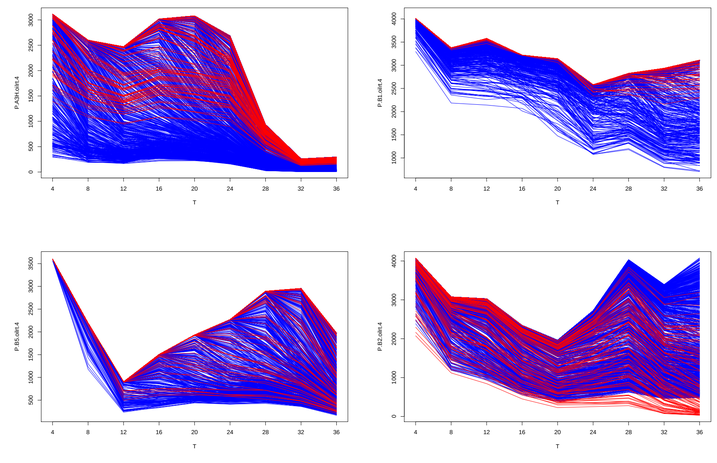

Figure 3: Plots of oil rate at four wells from accurate and coarse runs of the hydrocarbon simulator.

7 History Matching

- History matching is identifying the set X^* of possible choices of x^* for the simulator by comparing the simulator output with historical observations.
- Using the emulator we obtain, for x, $\mathrm{E}(F^a_H(x))$ and $\mathrm{Var}(F^a_H(x))$ and we rule out regions of x space for which we expect that $F^a_H(x)$ is likely to be a very poor match to z_H.
- We specify a prior for the model discrepancy.
- We assess match quality of a given x for a collection of q outputs of the accurate simulator F^a_H by the implausiblity of x

$$I(x) = (\mathrm{E}(F^a_H(x)) - z)^T \mathrm{Var}(F^a_H(x) - z)^{-1} (\mathrm{E}(F^a_H(x)) - z)/q$$

- Large implausibility corresponds to a poor match quality. Low implausibility corresponds to either good match quality or high uncertainty.

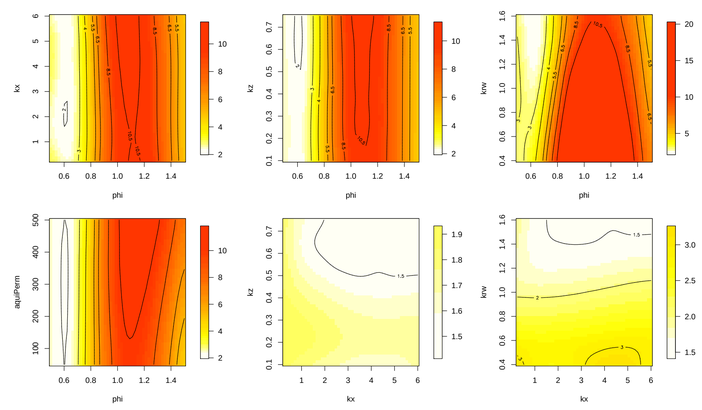

Figure 4: Implausibility plots for a subset of active variables

- At each production well, we history match the Gullfaks reservoir to the time series of 12 oil production rates.
- Application to the Gullfaks model show that values of porosity $\phi > 0.79$ are implausible history matches to the observed history.
- With this information, we can then refocus our analysis and re-emulate within the non-implausible regions, X^+, of the input space.

8 Forecasting

- The mean and variance of $F^a(x^*)$ are obtained from the mean function and variance function of the emulator for F^a over X^+ (denoted $\mu(x) = \mathrm{E}(F^a(x))$ and $\Sigma(x, x') = \mathrm{Var}(F^a(x), F^a(x'))$), which are the only features of the emulator that we are required to specify.
- Using these values we can then compute the unconditional mean (μ^*) and variance (Σ^*) of $F^a(x^*)$ by first conditioning on x^* and then integrating out with respect to the prior distribution.
- Given μ^* and Σ^*, it is then straightforward to compute the joint mean and variance of the collection (y_P, z_H) with no Gaussian requirement on the error terms.
- For the Gullfaks model, we consider forecasting an additional time point located one-year beyond the end of the original time series.
- We can now evaluate the forecast by the adjusted mean and variance for y_P adjusted by z_H using the Bayes linear adjustment formulae.

$$\mathrm{E}_{z_H}(y_P) = \mu^*_P + (\Sigma^*_{PH} + \Sigma^\epsilon_{PH})(\Sigma^*_H + \Sigma^\epsilon_H + \Sigma^e_H)^{-1}(z_H - \mu^*_H),$$
$$\mathrm{Var}_{z_H}(y_P) = (\Sigma^*_P + \Sigma^\epsilon_P) - (\Sigma^*_{PH} + \Sigma^\epsilon_{PH})(\Sigma^*_H + \Sigma^\epsilon_H + \Sigma^e_H)^{-1}(\Sigma^*_{HP} + \Sigma^\epsilon_{HP}).$$

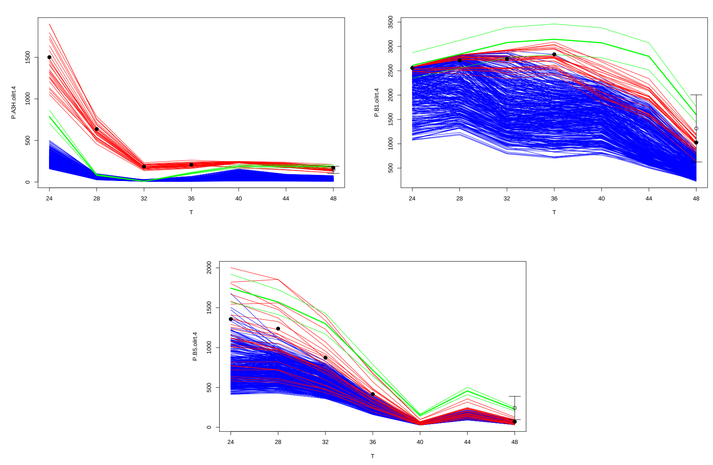

Figure 5: Plots of coarse and accurate runs with observations, μ^* ($\bullet$) and system forecasts (Φ) for the three active wells

- Using a Uniform prior on x^* and our multiscale emulator for $F^a(x)$ we obtain the forecasts for y_P illustrated in 5.
- Observe that whilst μ^*_P may differ substantially from z_P, all our forecast intervals for y_P are within measurement error of the observed production at that point.

References

Craig, P. S., Goldstein, M., Rougier, J. C., and Seheult, A. H. (2001), "Bayesian forecasting for complex systems using computer simulators," *Journal of the American Statistical Association*, 96, 717–729.

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1997), "Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments," in *Case Studies in Bayesian Statistics*, eds. Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P., and Singpurwalla, N. D., New York: Springer-Verlag, vol. 3, pp. 36–93.

Cumming, J. A. and Goldstein, M. (2008), "Small-sample Designs for Complex High-Dimensional Models Based on Fast Approximations," *In submission*.

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

Goldstein, M. and Wooff, D. A. (2007), *Bayes Linear Statistics: Theory and Methods*, Wiley.