Multivariate Emulation: Is it Worth the Trouble?

Tom Fricker and Jeremy Oakley

University of Sheffield

Statistics and Machine Learning Interface Meeting
24th July 2009

Emulators for computer models

We want to emulate a p-input, k-output deterministic computer model.

- Treat the computer model as an unknown function $\boldsymbol{\eta} : \mathcal{X} \subset \mathbb{R}^p \mapsto \mathbb{R}^k$
- Prior:

$$\boldsymbol{\eta}(.)|\beta, \Sigma, \Phi \sim GP_k[\mathbf{m}(.), \mathrm{C}(., .)]$$

- $\mathbf{m}(\mathbf{x}) = (\mathbf{1}\ \mathbf{x}^T)\beta$: we use a linear trend
- $\mathrm{C}(\mathbf{x}, \mathbf{x}')$: a $k \times k$ matrix covariance function with hyperparameters (Σ, Φ)
  - ▷ A more complex regression structure may reduce the importance of the covariance function (cf J. Rougier)
  - ▷ But only if it is a good representation of the structure of the computer model.

The covariance function

We assume there is little knowledge about structure of $\boldsymbol{\eta}(.)$. The focus of our work is the multivariate covariance function $\mathrm{C}(.,.)$.

- Represents 2 types of correlation in **our beliefs about the residuals** (after subtracting the trend):
  - correlation between different outputs
  - correlation over input-space - $\boldsymbol{\eta}(.)$ is smooth
- Remember: there is no *'true'* correlation between the outputs.

How do we go about specifying and combining the 2 types of correlation?

1. Independent outputs (*IND*)

Most straightforward:

Ignore any between-output correlation, treat outputs as being **independent**

$$\mathrm{cov}[\eta_i(\mathbf{x}), \eta_j(\mathbf{x}')] = \delta_{ij}\sigma_j^2 c_j(\mathbf{x}, \mathbf{x}')$$

- Build a univariate GP emulator for each output
- Each output has its own spatial correlation function
- Train the emulator for output j using only data from output j.

2. Separable covariance (*SEP*)

Easiest way to define a multivariate covariance function:

Treat the two types of correlation as **separable**

(e.g. Conti & O'Hagan, 2007)

$$\mathrm{C}(\mathbf{x}, \mathbf{x}') = \Sigma c(\mathbf{x}, \mathbf{x}')$$

- Σ : between-outputs covariance matrix
- $c(\mathbf{x}, \mathbf{x}')$: spatial correlation function

Disadvantage: all outputs share the same spatial correlation function $c(\mathbf{x}, \mathbf{x}')$

3. Non-separable covariance

Somewhere between *IND* and *SEP*:

The Linear Model of Coregionalization (*LMC*)

(e.g. Wackernagel, 1995; Gelfand *et al.*, 2004)

- Outputs are linear combination of independent univariate GPs in vector $\mathbf{Z}(.)$:

$$\boldsymbol{\eta}(.) = \beta\mathbf{h}(.) + \mathrm{R}\mathbf{Z}(.)$$

$$Z_j(.) \sim GP[0, \kappa_j(., .)] \qquad j = 1, ..., k$$

  - ▷ we use squared exponentials for $\kappa_j(., .)$

- Between-output covariance at any given input is $\Sigma = \mathrm{R}\mathrm{R}^T$

$$\boldsymbol{\eta}(.) = \beta\mathbf{h}(.) + \mathrm{R}\mathbf{Z}(.), \qquad Z_j(.) \sim GP[0, \kappa_j(.,.)]$$

$$\Rightarrow \qquad \mathrm{C}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=1}^{k} \mathrm{T}_\ell \kappa_\ell(\mathbf{x}, \mathbf{x}'), \qquad \mathrm{T}_\ell = \mathrm{R}_{\bullet\ell}\mathrm{R}_{\bullet\ell}$$

This is a special case of the 'nested covariance' model,

$$\mathrm{C}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=1}^{S} \mathrm{T}_\ell \kappa_\ell(\mathbf{x}, \mathbf{x}')$$

- Taking $S = k$ and $\mathrm{T}_\ell = \mathrm{R}_{\bullet\ell}\mathrm{R}_{\bullet\ell}$ is a 'natural' way of ensuring the T_ℓ are positive semi-def:
  - parameterise by $\Sigma = \mathrm{cov}[\boldsymbol{\eta}(\mathbf{x}), \boldsymbol{\eta}(\mathbf{x})]$
  - decompose as $\Sigma = \mathrm{R}\mathrm{R}^T$
  - the correlation function for an individual output is a weighted sum of 'basis' functions $\kappa_j(.,.)$.
  - if no between-output correlation, then $\mathrm{corr}[\eta_j(\mathbf{x}), \eta_j(\mathbf{x}')] = \kappa_j(\mathbf{x}, \mathbf{x}')$, i.e. equivalent to *IND*.

Inference for hyperparameters

Hyperparameters in the GP prior,
$\boldsymbol{\eta}(.)|\beta, \Sigma, \Phi \sim GP_k[\mathbf{m}(.), \mathrm{C}(., .)]$:

- β, regression coefficients
  - conjugate prior, integrated out
- Σ, between-output covariance
  - *SEP*/*IND*: conjugate prior, integrated out
  - *LMC*: analytic integration not possible
- Φ, spatial correlation function parameters
  - analytic integration not possible for any of the emulators

For hyperparameters that cannot be analytically integrated:
we **estimate** by MLE and treat as **fixed**.

Regular outputs

We make the assumption that the computer model has *regular outputs*:

- The set of outputs is finite and fixed.
- Every output is observed at every input point (cf. *isotopic data* in geostatistics)

For *SEP*, this implies that the posterior for output j is a function only of data from output j:

$$\eta_j(.)|y_j \perp y_i \quad \forall i \neq j$$

Does a multivariate specification ever help?

Case Study 1: Simple Climate Model

(Work with Nathan Urban)

- 5 inputs
- We shall focus on 2 univariate outputs:
  - $\mathrm{CO_2}$ flux in the year 2000 (CO_2)
  - Surface temperature in the year 2000 (*temp*)
- Data: 60 training runs in an Latin hypercube design.
- Validation: a further 100 model runs.
- Emulators:
  - *SEP*, a separable emulator
    - 1 squared-exponential correlation function
  - *LMC*, an LMC emulator
    - 2 squared-exponential basis correlation functions
  - *IND*, 2 independent univariate emulators
    - each with 1 squared-exponential correlation function

CO_2

MSPE

SEP	LMC	IND
82.4	19.0	15.2

$Temp$

MSPE

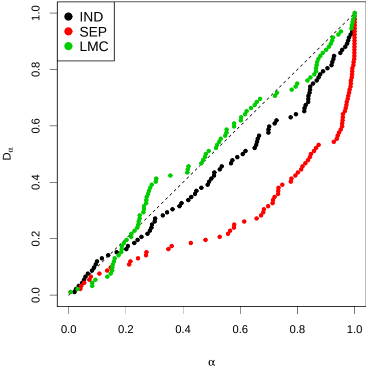
SEP	LMC	IND
7.4	4.0	3.0

CO_2

MSPE

SEP	LMC	IND
82.4	19.0	15.2

% of CIs containing true values

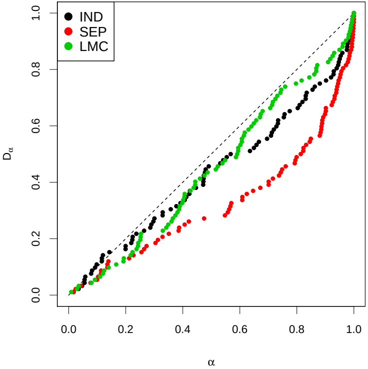


$Temp$

MSPE

SEP	LMC	IND
7.4	4.0	3.0

% of CIs containing true values



Independent emulators do just as well as LMC - So why bother with the multivariate specification?

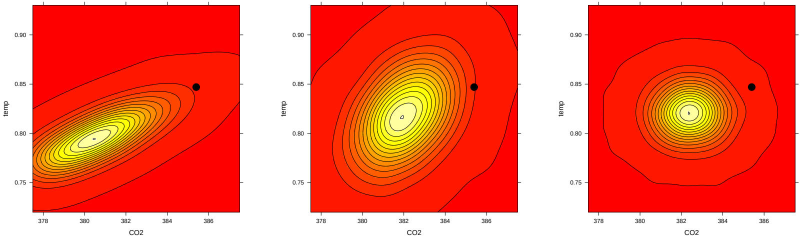
Example: Gross Primary Productivity (GPP), Π, a univariate function of the outputs

$$\Pi = \Pi_{max}\left[\frac{\boldsymbol{CO_2}}{(\boldsymbol{CO_2} + C)} + (T_{opt} \times \boldsymbol{temp} + 0.5 \times \boldsymbol{temp}^2\right]$$

What is the predictive distribution Π?

- simulate from the joint posterior of $(\boldsymbol{CO_2}, \boldsymbol{Temp})$

Joint posterior of $(\boldsymbol{CO_2}, \boldsymbol{Temp})$ at one particular validation point

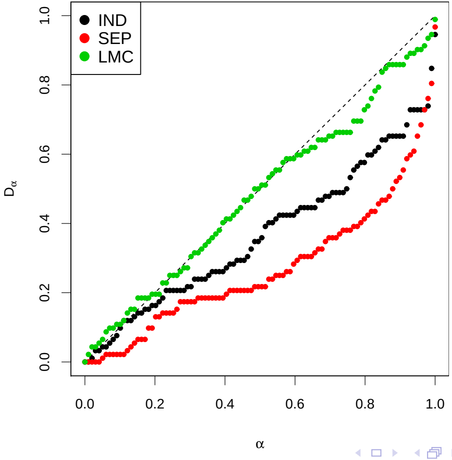


SEP LMC IND

GPP

MSPE	SEP	LMC	IND
	9.35	1.97	2.13

% of CIs containing true values



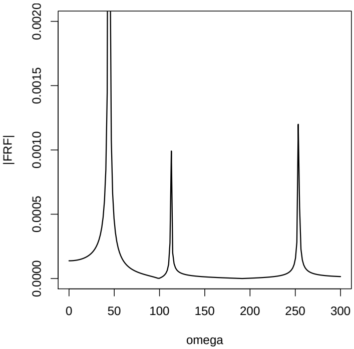
Case Study 2: A Finite Element Model

A simple finite element model for an aeroplane (Work with Neil Sims)

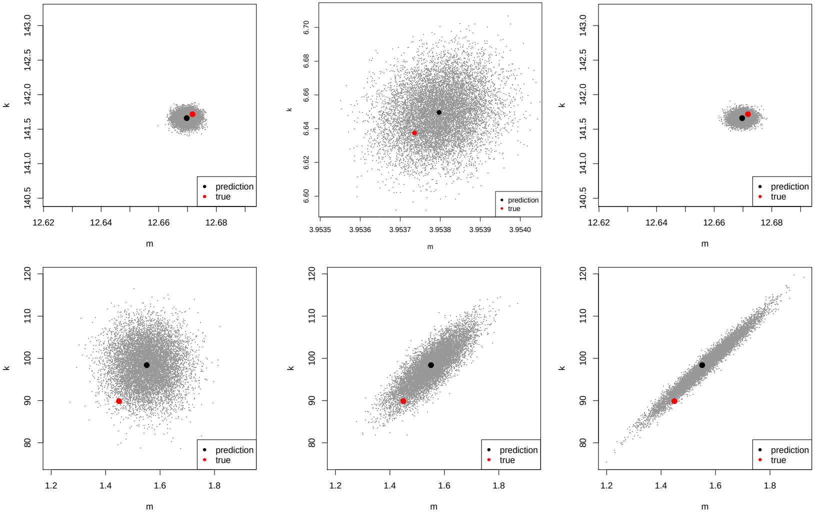
- The structure is represented by a large number of nodes.
  - The structure is represented by a large number of nodes.
  - A smaller number of parameters are used to set the overall physical properties of the structure - e.g. wing length, fuselage thickness, etc.
  - Select 5 as the variable inputs
- Outputs:
  - 3 pairs of mass and stiffness 'modal parameters', (m_i, k_i).
- The outputs are then combined to form the coefficients in a **frequency response function**,

$$FRF(\omega) = \sum_{i=1}^{3} \frac{1}{k_i - \omega^2 m_i}$$

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \xrightarrow{\boldsymbol{\eta}} \begin{pmatrix} m_1 \\ k_1 \\ m_2 \\ k_2 \\ m_3 \\ k_3 \end{pmatrix} \to FRF(\omega) = \sum_{i=1}^{3} \frac{1}{k_i - \omega^2 m_i}$$

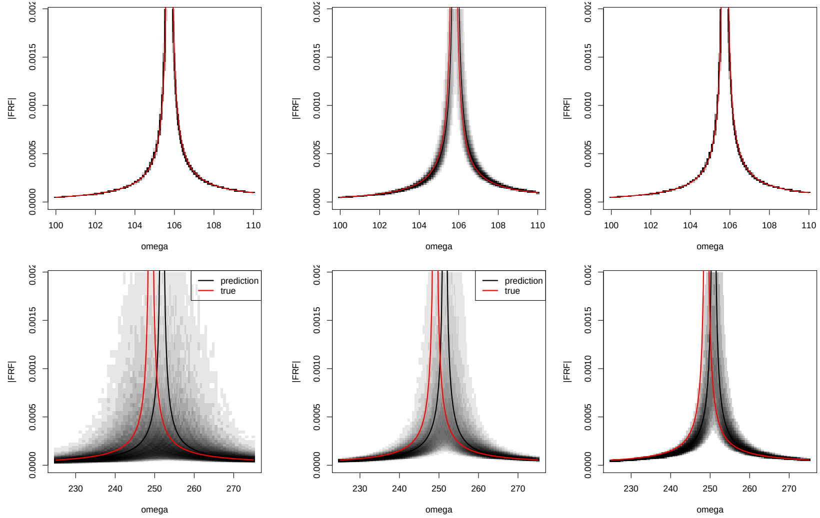


Single validation point, m v. k



Independent Separable LMC

Single validation point, $FRF(\omega)$



Independent Separable LMC

Conclusions

- I have not found any circumstances where a multivariate emulator outperforms independent univariate emulators *if we are only interested in marginal predictions of individual outputs.*
- But it does not seem uncommon for multiple outputs of a computer model to be used jointly.
- In this case, a multivariate specification can be important for propagating the uncertainty surrounding the joint predictions.
- A non-separable covariance structure can lead to better predictions by allowing different spatial correlation functions for different outputs.

Acknowledgements

Many thanks to Dr. Nathan Urban (Geosciences, Penn State university) for providing the Simple Climate Model data, and Neil Sims (Dept. Mechanical Engineering, University of Sheffield) for providing the FEM data.

References

- Conti, S. and O'Hagan, A. (2007). Bayesian emulation of complex multi-output and dynamic computer models, *Journal of Statistical Planning and Inference*. In review.
- Wackernagel, H. (1995). *Multivariate Geostatistics*, Springer.
- Gelfand, A. E., Schmidt A. M., Banerjee, S., and Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization (with discussion), *Test*, v. 13, no. 2, p. 1-50.
- Urban, N. M. and Keller, K. (2008). Probabilistic hindcasts and projections of the coupled climate, carbon cycle, and Atlantic meridional overturning circulation systems: A Bayesian fusion of century-scale observations with a simple model, *Tellus A*, In review.