Testing for the independence of computer model outputs

Thomas E. Fricker, Jeremy E. Oakley

August 6, 2010

Abstract

When emulating a multiple output computer model, there are advantages in treating the outputs as independent and representing them with a collection of independent univariate Gaussian processes, rather than specifying a full covariance function. In this report we introduce methods for assessing the adequacy of the independent outputs approach in terms of producing joint-output predictions. A key point that sets the emulation problem apart from the general problem of fitting multivariate linear models with correlated residuals is that the variables of interest (the computer model outputs) are observed with no error. As a result, existing methods that make use of 'fitted residuals' (i.e. the difference between observations and the fitted model) are not applicable. Instead, we must validate the independent outputs approach using the emulator's predictions of future observations. Accordingly, we propose methods that partition the available data into training and validation sets, judging the emulator trained using the training data on its ability to make joint-output predictions of the validation data. We also also propose a cross-validation method, for use when the data are too sparse to allow the partitioning approach. We compare a variety of diagnostic quantities and tests using simulated data.

1 Introduction

An *emulator* is a statistical representation of a deterministic computer, used as a tool in the analysis of expensive computer experiments. Starting with the early work of Sacks et al. (1989); Currin et al. (1991); Haylock and O'Hagan (1996), a common approach has been developed in which a Gaussian process is used to represent the unknown function encoded by the computer model. The Gaussian process is conditioned on some training runs, and the resulting posterior process gives predictions of the output at any untried configuration of the inputs. Other methods have been used to create fast 'metamodels' of computer models, such as response surfaces (Dejean and Blanc, 1999; Rutherford et al., 2005), neural networks (El Tabach et al., 2007) and support vector machines (Drucker et al., 1997; Wang et al., 2005). The key difference between an emulator and these other approaches is that the output of an emulator is a full probability distribution, quantifying uncertainty in the predicted response of the computer model.

The original work in this field dealt with emulating a single output of a computer model. More recently there has been a recognition that most serious computer models produce a multitude of outputs, and that there is a need to emulate more than one output at a time. An important consideration when constructing an emulator for multiple outputs is how to deal with between-output dependencies. Perhaps the most simple approach is to assume that the outputs are a collection of unrelated functions, and emulate each output independently with a univariate Gaussian process. Doing so, however, may result in losing important information.

The alternative to assuming independent outputs is to model outputs jointly using a multivariate emulator. However, one generally has to make other some assumptions about the computer model in order to construct a valid dependency structure for the emulator. Sometimes these assumptions will result in poor marginal predictions of individual outputs. Thus, if the between-output dependencies are small (or non-existent), then best approach may be to use independent univariate emulators. The problem is deciding when this is the case. A comprehensive approach is to consider a variety of models with different dependency structures, including the independent outputs model, and to compare them, for example using Bayes factors. This can be time consuming though, and in this report we consider a quicker approach in which we fit an independent outputs model and assess its adequacy. We describe a number of methods for validating the joint predictions of an independent outputs emulator.

1.1 Tests for independence of autocorrelated processes

There is an extensive literature on the subject of testing for independence of variables. For categorical variables there are contingency table based tests such as the Pearson chi-squared test (and various extensions, see Rao and Scott, 1981) and exact tests (Agresti, 1992). For continuous variables there are various tests motivated by the fact that joint distribution function of independent variables is the product of the marginal distribution functions. The tests are then based on measures of distance between the estimated joint distribution function and the product of the estimated marginals (Hoeffding, 1948; Blum et al., 1961). Similar tests are available that, rather than using the distribution function, instead use the probability density function (Rosenblatt and Wahlen, 1992; Ahmad and Li, 1997) or the characteristic function (Csörgő, 1985; Kankainen and Ushakov, 1998). Related to testing for independence is testing for uncorrelatedness. Under the assumption of joint normality of the variables, independence and uncorrelatedness are equivalent. Tests for uncorrelatedness are generally based on correlation coefficients, often the Pearson product-moment correlation coefficient (Fisher, 1915; Hawkins, 1989), but rank correlation coefficients may also be used (Fieller et al., 1957).

A limitation of many of the tests alluded to above is that they assume that the observations of each variable are i.i.d., making them unsuitable for testing for independence of stochastic processes that exhibit autocorrelation (either serial or spatial correlation). The presence of autocorrelation will generally inflate the variance of test statistics (since the effective sample size is reduced), which will invalidate the tests if the observations are assumed i.i.d.. A modification may be made to the test's reference distribution to allow for the autocorrelation, but that will generally result in a lower powered test. Cerioli (2002) demonstrate this in the case of a Pearson chi-squared test for independence of categorical variables observed on a spatial field.

For continuous variables, the problem of autocorrelation is often encountered when the observations are time series. A popular approach to testing for independence between two time series is 'pre-whitening': univariate models are fitted each time series (usually ARMA-type models), then inverted to obtain residuals that are assumed to be white noise. Test statistics are then calculated using the white noise residuals. If the univariate models are well fitted, then the test recovers the power of the equivalent test for i.i.d. observation. Examples include Haugh (1976) and Hong (1996) who consider the sample cross-correlation of the white noise residuals, and Wu et al. (2009) who considers a distance-based test on pre-whitened data. (Karvanen, 2005) proposes an alternative to pre-whitening, in which it is assumed that there is a lag d beyond which observations may be considered independent. The procedure involves resampling small subsets of data which have intra-point lags greater than d, then invoking a distance-based test on the subsets.

2 Emulating multiple outputs

Consider a deterministic computer model that takes an input vector $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$, and returns a vector of outputs $\mathbf{y} \in \mathbb{R}^k$. We represent the input-output relationship of the computer model by the function $\boldsymbol{\eta}(.)^T = (\eta_1(.), ..., \eta_k(.))$. In the usual emulation framework the uncertainty in $\boldsymbol{\eta}(.)$ is represented by the model

$$\boldsymbol{\eta}(.) = \mathrm{B}^T \mathbf{h}(.) + \mathbf{z}(.),$$

where $\mathbf{h}(.)$ is a vector of q regressors, B is a $q \times k$ matrix of unknown coefficients, and $\mathbf{z}(.)$ is a residual k-variate stochastic process, independent of B. Prior knowledge about the functional form of the outputs is built into the emulator through the choice of regressors in $\mathbf{h}(.)$ and the prior $\pi(\mathrm{B})$, while the nonparametric residual $\mathbf{z}(.)$ ensures that the posterior process interpolates the data.

The emulator prior is completed by specifying distributions for B and $\mathbf{z}(.)$. A flexible and tractable choice for $\mathbf{z}(.)$ is a stationary zero-mean Gaussian process,

$$\mathbf{z}(.)|\boldsymbol{\theta} \sim GP_k[0, \mathrm{C}(.,.)],$$

where $\mathrm{C}(.,.)$ is a $k \times k$ multivariate covariance function. An important point to note is that the residual $\mathbf{z}(.)$ does not represent 'error' in the traditional sense. Observations of the computer model are error-free, since output values are observed exactly at any point at which it is run, so $\mathbf{z}(.)$ represents the uncertainty in the outputs given the regression function $\mathrm{B}^T\mathbf{h}(.)$. Often knowledge about the coefficients B is weak, so we assume that the improper prior $\pi(\mathrm{B}) \propto 1$ is used, implying that

$$\mathrm{cov}[\eta_i(\mathbf{x}), \eta_j(\mathbf{x}')] = \mathrm{C}(\mathbf{x}, \mathbf{x}'). \tag{1}$$

That is, the dependence between the outputs is expressed entirely through the Gaussian process covariance function $\mathrm{C}(.,.)$.

Since the Gaussian process is stationary, the elements of $\mathrm{C}(.,.)$ can be expressed as

$$\mathrm{C}_{ij}(\mathbf{x}, \mathbf{x}') = \Sigma_{ij} c_{ij}(\mathbf{x}, \mathbf{x}'), \tag{2}$$

where each $c_{ij}(.,.)$ is a stationary spatial correlation function and the Σ_{ij} are the elements of a $k \times k$ covariance matrix Σ. The spatial correlation functions describe the rate at which uncertainties in the outputs grow as a prediction point moves away from an observed input point. The matrix Σ gives the covariances between outputs at any given input point (the *between-outputs* covariance matrix). A key consideration is how we specify Σ and $\{c_{ij}(.,.) : i,j = 1,..k\}$, since arbitrary choices may not guarantee the positive-definiteness of $\mathrm{C}(.,.)$. There are a number of approaches to constructing valid covariance functions, each with its own set of assumptions and restrictions:

Separable covariance function It is assumed that the residual process for every output exhibits the same type of response to changes in the input, so all the outputs have the same spatial correlation function $c(.,.)$. Then $c_{ij}(.,.) = c(.,.)$ for all i, j, so the covariance has the separable form

$$\mathrm{C}(.,.) = \Sigma c(.,.). \tag{3}$$

The hyperparameters in this covariance function are $\boldsymbol{\theta} = \{\Sigma, \Phi.\}$, where Φ represents hyperparameters in the correlation function $c(.,.)$. This approach has considerable computational advantages, firstly because there is just one set of correlation length hyperparameters to infer, and secondly because it results in a simple Kronecker product representation of the data covariance matrix. Examples of emulators using the separable covariance function are found in Bhattacharya (2007), Rougier (2007), Kennedy et al. (2008), Rougier et al. (2009) and Conti and O'Hagan (2010).

Nonseparable covariance functions There are many ways of constructing separable covariance functions, and a review is found in Fricker et al. (2010). We briefly describe two here.

(a) Linear model of coregionalization (LMC)
The LMC was developed in the field of geostatistics as a tool to model multivariate spatial processes (Journel and Huijbregts, 1978, Wackernagel, 1995, Goulard and Voltz, 1992, Gelfand et al., 2004). The idea is to construct residual processes $\mathbf{z}(.)$ as linear combinations of a basis of k independent univariate Gaussian processes. The resulting covariance function is

$$\mathrm{C}(\mathbf{x}, \mathbf{x}') = \mathrm{R}[\mathrm{diag}\{\kappa_1(.,.), ..., \kappa_k(.,.)\}]\mathrm{R}^T, \tag{4}$$

where $\kappa_1(.,.), ..., \kappa_k(.,.)$ are the univariate correlation functions of the basis Gaussian processes, and R is a $k \times k$ matrix such that $\mathrm{RR}^T = \Sigma$. We refer to covariance functions of this type as *LMC* covariance functions.

(b) Convolution approach
A k-output Gaussian process can be constructed by, for each output j, choosing a smoothing kernel $\kappa_j(\mathbf{x})$ and convolving it with a common 'latent' Gaussian white noise process $w(\mathbf{x})$. The resulting covariance function is

$$C_{ij}(\mathbf{x}, \mathbf{x}') = \Sigma_{ij} \frac{\int_{\mathcal{X}} \kappa_i(\mathbf{u} - \mathbf{x})\kappa_j(\mathbf{u} - \mathbf{x}')\, d\mathbf{u}}{\int_{\mathcal{X}} \kappa_i(\mathbf{u})\kappa_j(\mathbf{u})\, d\mathbf{u}} \quad i, j = 1, ...k. \tag{5}$$

We refer to covariance functions of this type as *CONV* covariance functions.

Independent outputs This is a special case of the nonseparable covariance function , in which it is assumed that all aspects of the computer model that are common to the outputs are represented in the regressors $\mathbf{h}(.)$, so that the elements of $\mathbf{z}(.)$ are considered independent. The covariance function is then

$$\mathrm{C}_{ij}(\mathbf{x}, \mathbf{x}') = \begin{cases} \Sigma_{ij} c_{ij}(\mathbf{x}, \mathbf{x}') & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The hyperparameters in this covariance function are $\boldsymbol{\theta} = \{\Sigma_{1,1}, ..., \Sigma_{k,k}, \Phi_1, ..., \Phi_k\}$, where Φ_i represents hyperparameters in the correlation function $c_{ii}(.,.)$. With this approach, each output can be treated as a distinct problem, with a separate univariate emulator built for each one in turn. This means that a bespoke correlation function can be specified for each output without any additional restrictions to ensure they are going to fit together into a coherent covariance structure. As a result, we expect to obtain output-marginal predictions that are at least as good as those from an emulator using any of the above non-independent covariance functions, but with fewer hyperparameters to infer.

If the assumption of independent outputs does not hold, then the drawback of using this approach will be seen when joint-output predictions are made. In particular, if there is interest in a function that combines the outputs, then ignoring between-output dependencies may result in overconfident or underconfident predictions.

2.1 Prior versus posterior correlation

For any of the above choices of covariance structure, to build the emulator the computer model is run at training design points $X = (\mathbf{x_1}, ..., \mathbf{x_n})$, giving training data outputs $\mathbf{y}^T = (\mathbf{y_1}^T, ..., \mathbf{y_k}^T)$, where $\mathbf{y_j}^T = (\eta_j(\mathbf{x_1}), ..., \eta_j(\mathbf{x_n}))$ is the vector of data from the jth output. Conditioning on $\mathbf{y}$ and integrating over $\boldsymbol{\beta}$, we obtain

$$\boldsymbol{\eta}(.)|\mathbf{y}, \boldsymbol{\theta} \sim GP_k[\mathbf{m}^\ddagger(.), \mathbf{C}^\ddagger(., .)], \tag{7}$$

where, for a set of $\acute{n}$ new input points $\acute{X} = (\acute{\mathbf{x}}_\mathbf{1}, ..., \acute{\mathbf{x}}_\mathbf{n})$,

$$\mathbf{m}^\ddagger(\acute{X}) = \acute{\mathrm{H}}(\acute{X})\hat{\boldsymbol{\beta}} + \mathrm{F}(\acute{X})\mathrm{V}^{-1}(\mathbf{y} - \mathrm{H}\hat{\boldsymbol{\beta}}), \tag{8}$$

$$\mathbf{C}^\ddagger(\acute{X}, \acute{X}) = \mathbf{C}(\acute{X}, \acute{X}) - \mathrm{F}(\acute{X})\mathrm{V}^{-1}\mathrm{F}(\acute{X})^T + \tag{9}$$

$$(\acute{\mathrm{H}}(\acute{X}) - \mathrm{F}(\acute{X})\mathrm{V}^{-1}\mathrm{H})(\mathrm{H}^T\mathrm{V}^{-1}\mathrm{H})^{-1}(\acute{\mathrm{H}}(\acute{X}) - \mathrm{F}(\acute{X})\mathrm{V}^{-1}\mathrm{H})^T, \tag{10}$$

with $\hat{\boldsymbol{\beta}} = (\mathrm{H}^T\mathrm{V}^{-1}\mathrm{H})^{-1}\mathrm{H}^T\mathrm{V}^{-1}\mathbf{y}$. The notation here is as follows:

$$\begin{aligned}
\mathrm{H} &= \mathrm{I} \otimes \mathbf{h}(X)^T, \\
\acute{\mathrm{H}}(\acute{X}) &= \mathrm{I} \otimes \mathbf{h}(\acute{X})^T, \\
\mathrm{V} &= \mathbf{C}(X, X), \\
\mathrm{F}(\acute{X}) &= \mathbf{C}(\acute{X}, X).
\end{aligned}$$

In the case of the separable covariance function, Σ may be given a conjugate prior distribution and integrated out of the posterior. Similarly, when the independent outputs model is used, the individual output variances $\{\Sigma_{ii} : i = 1, .., k\}$ may be given independent conjugate prior distributions and integrated out of the posterior. However, the conditioning on the spatial correlation hyperparameters cannot, in general be removed analytically. With a non-independent, nonseparable covariance function, the conditioning on Σ cannot be removed analytically either. A fully Bayesian approach would update the remaining covariance function hyperparameters using MCMC, but then we no longer have a closed form for the posterior distribution, and predictions must be presented as a sample. This approach increases the computational burden, since each MCMC update requires the inversion of the full $nk \times nk$ matrix V, potentially making the emulator itself slow to use. We therefore follow Kennedy and O'Hagan (2001), Conti and O'Hagan (2010), Rougier et al. (2009), and others, we estimate the remaining covariance function hyperparameters and treat them as known.

We see from equation (10) that, while the prior process (1) is stationary given the mean function, the covariance of the posterior (7) is a function of the position in the input space. The posterior variance of any given output is exactly zero at any input that coincides with a training point, then grows as the input moves away from the training point. The posterior between-outputs correlation, given by

$$\rho^{\ddagger}_{ij}(\mathbf{x}) = \frac{\mathbf{C}^{\ddagger}_{ij}(\mathbf{x},\mathbf{x})}{\sqrt{\mathbf{C}^{\ddagger}_{ii}(\mathbf{x},\mathbf{x})\mathbf{C}^{\ddagger}_{jj}(\mathbf{x},\mathbf{x})}}, \quad i, j = 1, ..., k, \tag{11}$$

is also in general non-stationary. An exception is when a separable covariance function (3) is used, in which case the posterior covariance has the form

$$\mathbf{C}^{\ddagger}(\mathbf{x},\mathbf{x}') = \Sigma \otimes c^{\ddagger}(\mathbf{x},\mathbf{x}') \quad \mathbf{x},\mathbf{x}' \in \mathcal{X},$$

where $c^{\ddagger}(\mathbf{x},\mathbf{x}')$ is a scalar-valued function. Therefore, with a separable covariance function, the posterior between-outputs correlation is constant. Also, if the independent outputs model is used then the posterior between-outputs correlation is also constant, being zero everywhere.

When a non-separable covariance function is used, $\rho^{\ddagger}_{ij}(\mathbf{x})$ will generally be a nonconstant function of $\mathbf{x}$, whose value depends the proximity of $\mathbf{x}$ to the training data. However, due to the complicated form of (10) (which involves inversions of the data covariance matrix V), it is not easy to deduce qualitative information about the form of the function for a given nonseparable covariance function through inspection of (10) and (11). Instead we investigate empirically, using a two-output Gaussian process $GP_2[0, C(.,.)]$ with a two-dimensional input space $\mathcal{X} = [0,1]^2$. We use two choices of $C(.,.)$:

1. An *LMC* covariance function with $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ and basis correlation functions

$$\kappa_1(\mathbf{x}) = \exp\{-\mathbf{x}^T\mathrm{diag}(4,4)\mathbf{x}\},$$
$$\kappa_2(\mathbf{x}) = \exp\{-\mathbf{x}^T\mathrm{diag}(10,10)\mathbf{x}\}.$$

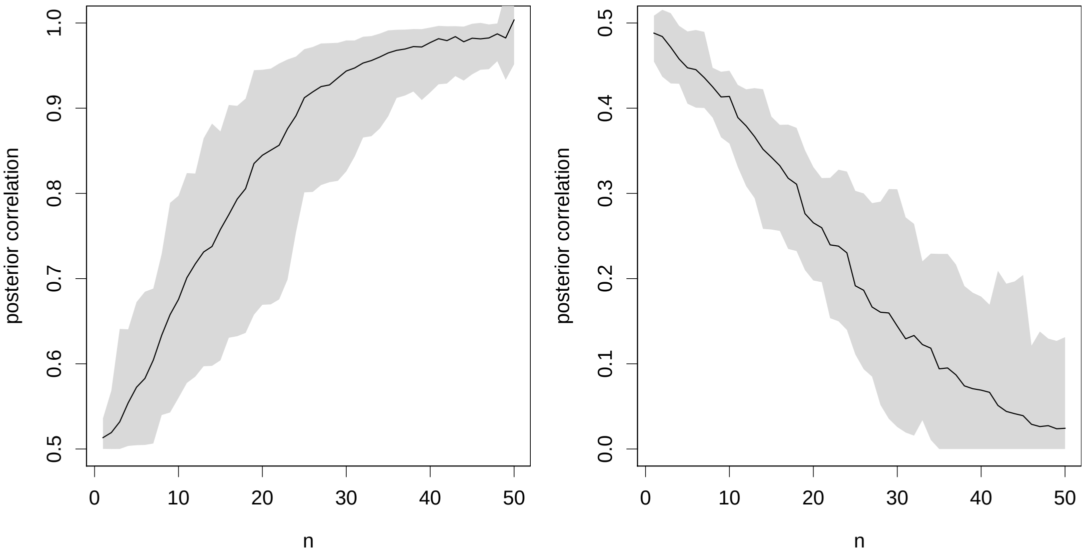
2. A *CONV* covariance function with $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ and smoothing kernels

$$\kappa_1(\mathbf{x}) = \exp\{-2\mathbf{x}^T\mathrm{diag}(4,4)\mathbf{x}\},$$
$$\kappa_2(\mathbf{x}) = \exp\{-2\mathbf{x}^T\mathrm{diag}(10,10)\mathbf{x}\}.$$

In both choices, the prior between-outputs correlation is 0.5. We investigate the posterior between-outputs correlation function $\rho^{\ddagger}_{12}(.)$ using various sizes of Sobol sequence designs (Kuipers and Niederreiter, 2006), as follows:

- For $n = 1, 2, ...,$
  1. Let $X_n = (\mathbf{x_1}, ..., \mathbf{x_n})$ be the first n points of a Sobol sequence on $\mathcal{X}$
  2. Let $X^* = (\mathbf{x^*_1}, ..., \mathbf{x^*_N})$ be the following N points of the same Sobol sequence.
  3. Using the given choice of $C(.,.)$, compute $\{\rho^{\ddagger}_{12}(\mathbf{x^*_i}) : i = 1, ..., N_s\}$.

We then use the sample $\{\rho^{\ddagger}_{12}(\mathbf{x^*_i}) : i = 1, ..., N_s\}$ to estimate the distribution of the posterior between-outputs correlation over the input space. Figure 1 shows the estimated mean and $(0.025, 0.975)$ quantiles of the posterior plotted against n. We see that with the *LMC* covariance function the posterior correlation tends to increase as the design size increases, approaching 1 everywhere by the time the input space is populated with 50 training data.



(a) *LMC*

(b) *CONV*

Figure 1: Posterior between-outputs correlation for a varying number of training data, for two choices of non-separable covariance function. Black line: $\rho^{\ddagger}_{12}(.)$ averaged over the input space. Grey regions: pointwise 95% quantiles of the distribution of $\rho^{\ddagger}_{12}(.)$ over the input space

When the *CONV* covariance function is used the posterior correlation tends to decrease with increasing design size, with its average value approaching zero at 50 training data.

These results about the posterior between-outputs correlation of two Gaussian processes have important implications. They show that in situations where two outputs are correlated but have different spatial correlation lengths, so they are best represented by a nonseparable covariance function, then conditioning on data changes the dependence structure of the outputs.

3 Testing for independence of computer model outputs

From the discussion in the previous sections, we can identify some important points we must consider when constructing a test for independence of computer model outputs:

1. We assume weak knowledge about the regression coefficients β, so the dependence structure of the outputs is specified by the covariance function of the residual process $\mathbf{z}(.)$.
2. The residual process $\mathbf{z}(.)$ represents uncertainty, rather than error. After updating the process in light of the data, $\mathbf{z}(.)$ is zero at any point at which we have evaluated the computer model.
3. The between-outputs dependencies, expressed by the cross-correlation functions of $\mathbf{z}(.)$, may change as we condition on data. They may increase or decrease, depending on the particular covariance structure of $\mathbf{z}(.)$.
4. If we erroneously treat the outputs as independent, then joint-output predictions will be misspecified.
5. Spatial autocorrelation is present in $\mathbf{z}(.)$.

The purpose of the emulator is to make predictions of the computer model outputs after conditioning on training data, and we require a method for identifying whether we may treat the uncertainty in those predictions as independent. Most current methods for assessing independence of autocorrelated variables (for example pre-whitening) use the differences between the fitted model and the observations to infer the dependence structure. These methods will not work in this setting, because point 2 tells us that the posterior residual process is zero at all points at which we have evaluated the computer model. In other words, after fitting the emulator we have no information left to use in the test (see Haslett and Hayes, 1998 for further discussion of this subject). Another issue is that point 3 tells us that the between-output correlations may change after conditioning on the data. Therefore we require a method that is based on the emulator's predictions of future data.

One solution to this problem is to partition the available data into a training set and a validation set (Bastos and O'Hagan, 2009). This can be useful as it enables us to examine the joint predictive distribution of a collection of points. However, in situations when the data are sparse we may not be able to afford to hold back data from the training set. In that case an alternative approach is cross-validation. We now describe schemes for assessing the adequacy of the independent outputs model using each of these two approaches. We describe the schemes for two outputs, $\boldsymbol{\eta}(.) = (\eta_1(.), \eta_2(.))^T$. For more than two outputs the schemes may be applied pairwise (but note that pairwise independence of $k > 2$ outputs does not imply mutual independence).

3.1 Data partitioning

We partition the data into a training set $\{X, \mathbf{y_1}, \mathbf{y_2}\}$ of size n and a validation set $\{\acute{X}, \acute{\mathbf{y}}_\mathbf{1}, \acute{\mathbf{y}}_\mathbf{2}\}$ of size $\acute{n}$. A univariate emulator is built for each output, giving predictive distributions

$$\eta_1(\acute{X})|\mathbf{y_1}, \boldsymbol{\theta}_1 \sim N[\mathbf{m}_1^\ddagger(\acute{X}), \mathbf{C}_1^\ddagger(\acute{X}, \acute{X})] \tag{12}$$

$$\eta_2(\acute{X})|\mathbf{y_2}, \boldsymbol{\theta}_2 \sim N[\mathbf{m}_2^\ddagger(\acute{X}), \mathbf{C}_2^\ddagger(\acute{X}, \acute{X})]. \tag{13}$$

Note that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ may represent either the complete set of covariance hyperparameters, or if we choose to use a conjugate prior for $(\Sigma_{11}, \Sigma_{22})$ and integrate over these hyperparameters then $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ represent just the spatial correlation function hyperparameters. The prediction errors are then

$$\mathbf{e_1} = \mathbf{m}_1^\ddagger(\acute{X}) - \acute{\mathbf{y}}_\mathbf{1} \tag{14}$$

$$\mathbf{e_2} = \mathbf{m}_2^\ddagger(\acute{X}) - \acute{\mathbf{y}}_\mathbf{2}. \tag{15}$$

The two univariate emulators should be individually validated to ensure that they are good output-marginal representations of the computer model (Bastos and O'Hagan, 2009). If the univariate emulators are well validated then any problems with their joint predictions are an indication of a problem with the assumption of their independence. We consider four methods for assessing the adequacy of the joint predictions:

1. The sample correlation coefficient calculated using the prediction errors:

$$r_\mathbf{e} = \frac{\mathbf{e_1}^T\mathbf{e_2}}{\sqrt{\mathbf{e_1}^T\mathbf{e_1}}\sqrt{\mathbf{e_2}^T\mathbf{e_2}}}. \tag{16}$$

   This can be used as an estimate of the true posterior between-outputs correlation. Alternatively, the distribution of $r_\mathbf{e}$ under the hypothesis of independence of $\eta_1(.)$ and $\eta_2(.)$, which has an expected value of zero, can be obtained by simulation, and a classical hypothesis can be performed.

2. A problem with using $r_{\mathbf{e}}$ to estimate the between-outputs correlation is that it does not account for the spatial correlation between the residuals. We could therefore consider a pre-whitening approach, in which we compute the de-correlated prediction errors,

$$\mathbf{s_1} = \mathbf{C}_1^{\ddagger}(\acute{X}, \acute{X})^{-\frac{1}{2}}\{\mathbf{m}_1^{\ddagger}(\acute{X}) - \acute{\mathbf{y}}_{\mathbf{1}}\} \tag{17}$$

$$\mathbf{s_1} = \mathbf{C}_2^{\ddagger}(\acute{X}, \acute{X})^{-\frac{1}{2}}\{\mathbf{m}_2^{\ddagger}(\acute{X}) - \acute{\mathbf{y}}_{\mathbf{2}}\}. \tag{18}$$

   We then look at the sample correlation coefficient calculated using the de-correlated prediction errors,

$$r_{\mathbf{s}} = \frac{\mathbf{s_1}^T\mathbf{s_2}}{\sqrt{\mathbf{s_1}^T\mathbf{s_1}}\sqrt{\mathbf{s_2}^T\mathbf{s_2}}}. \tag{19}$$

   This can be used in similar manner to $r_{\mathbf{e}}$, but with the advantage of having a smaller sampling variance.

3. We noted that a potential consequence of ignoring between-output correlation is that predictions of a function of the outputs may be overconfident or underconfident. Using this idea, we consider predicting $f(\mathbf{x}) = \alpha_1\eta_1(\mathbf{x}) + \alpha_2\eta_2(\mathbf{x})$, where (α_1, α_2) are suitably chosen constants to scale the outputs to comparable sizes. If the between-outputs covariance is zero then we expect the standardised errors

$$t_i = \frac{f(\acute{\mathbf{x}}_{\mathbf{i}})}{\sqrt{\alpha_1^2\mathbf{C}_1^{\ddagger}(\acute{\mathbf{x}}_{\mathbf{i}}, \acute{\mathbf{x}}_{\mathbf{i}}) + \alpha_2^2\mathbf{C}_2^{\ddagger}(\acute{\mathbf{x}}_{\mathbf{i}}, \acute{\mathbf{x}}_{\mathbf{i}})}} \quad i = 1, ..., \acute{n}, \tag{20}$$

   to have an approximately standard normal distribution (but note that $\{t_i : i = 1, ..., \acute{n}\}$ are not independent due to the spatial autocorrelation). If the individual output emulators are well validated, then unusually large or small standardised errors indicate that there is a problem with the joint covariance structure, and that the independent outputs model is inadequate.

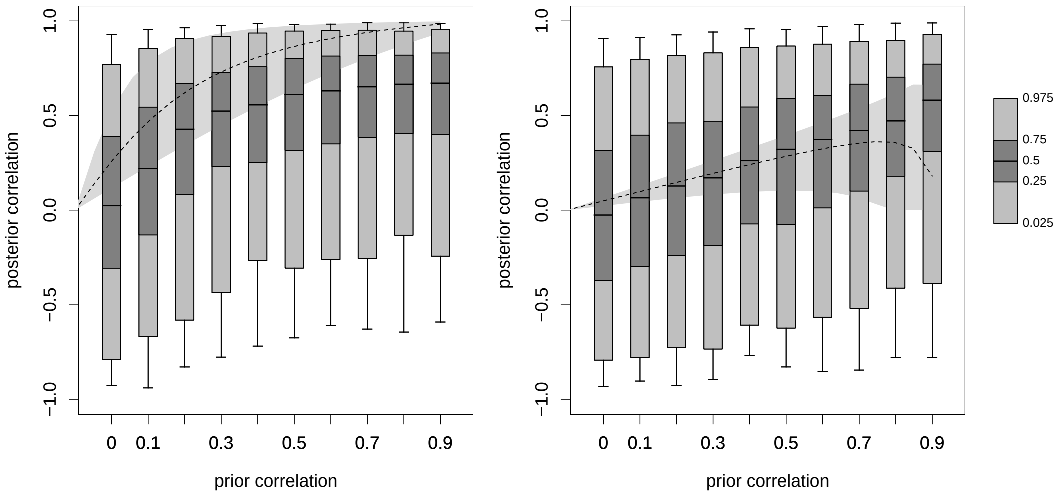
4. Under the hypothesis of independence, the (squared) Mahanalobis distance of the predictions is

$$D_{MD} = \mathbf{e_1}^T\mathbf{C}_1^{\ddagger}(\acute{X}, \acute{X})^{-1}\mathbf{e_1} + \mathbf{e_2}^T\mathbf{C}_2^{\ddagger}(\acute{X}, \acute{X})^{-1}\mathbf{e_2}. \tag{21}$$

   This has a theoretical $\chi^2_{2\acute{n}})$ distribution (or a scaled F distribution if $(\Sigma_{11}, \Sigma_{22})$ have been integrated out of the posterior, see Bastos and O'Hagan, 2009). A particularly large or small value of D_{MD} compared to the theoretical distribution indicates a problem with the independent outputs assumption. This diagnostic has the advantage that provides a single summary that takes into account the spatial autocorrelations.

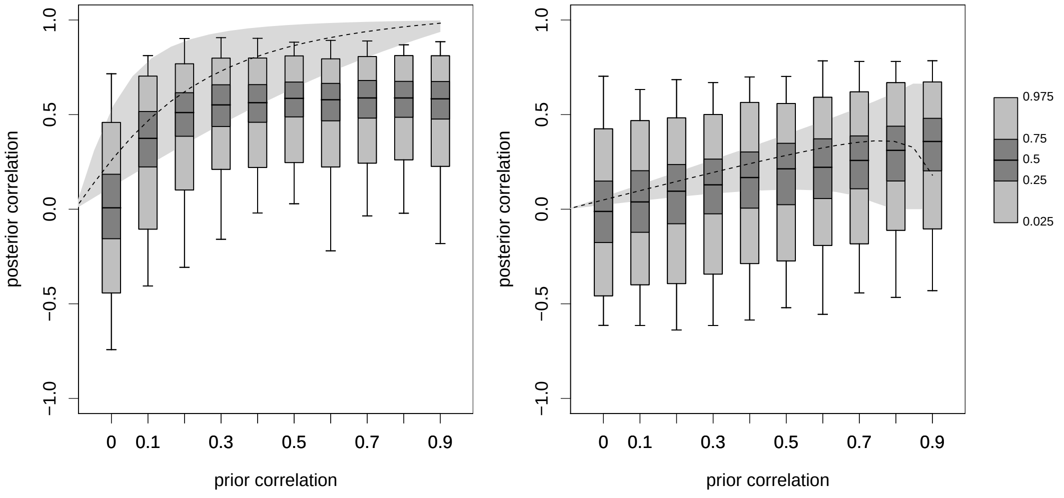
We demonstrate each of the four methods using data simulated from the same Gaussian processes as in section 2.1. We simulate observations at 40 points in a Sobol sequence design, taking the first 20 as training data and the second 20 as validation data. We generate a large number of data sets and calculate the diagnostic quantities described above.

Figures 2 and 3 show box plots of $r_{\mathbf{e}}$ and $r_{\mathbf{s}}$ respectively, with the distribution of the true posterior between-outputs correlation displayed in the background. Both methods tend to



(a) *LMC* (b) *CONV*

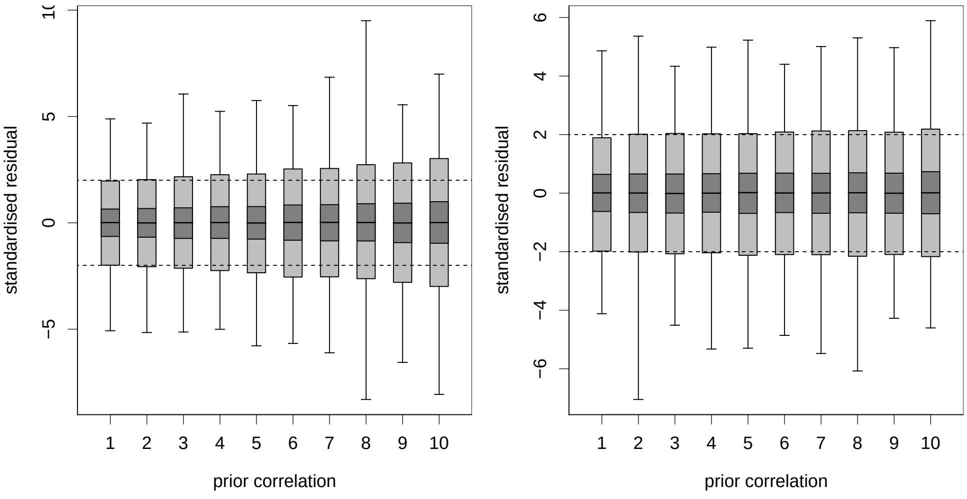
Figure 2: Box plots of sample correlation coefficient of prediction errors, using data simulated from Gaussian processes with *CONV* and *LMC* covariance functions. The legend to the right shows which quantiles are represented in the plots. Dashed line: true posterior correlation averaged over the input space. Grey regions: pointwise 95% quantiles of the distribution of the posterior correlation over the input space.



(a) *LMC* (b) *CONV*

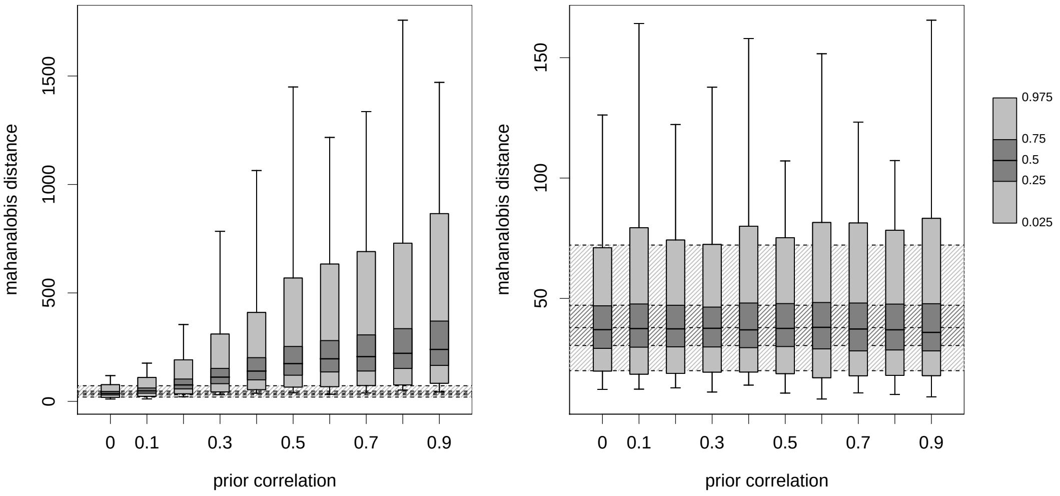
Figure 3: Box plots of sample correlation coefficient of spatially uncorrelated prediction errors, using the same data as in Figure 2.

underestimate the correlation when the true covariance function is an *LMC* type, but provide more accurate estimates when the true covariance function is a *CONV* type. As expected, the sampling variance of $r_{\mathbf{e}}$ is somewhat larger than that of $r_{\mathbf{s}}$. For calculating $\{t_i : i = 1, ..., \acute{n}\}$ we choose $\alpha_1 = \alpha_2 = 1$ since the data are generated from unit variance Gaussian processes. Figure 4 shows box plots of these standardised errors. We see that their variance increases as the true prior correlation increases, showing that unusually large values can be used as indication of positive between-outputs correlation. Figure 5 shows box plots of D_{MD}. The



(a) *LMC* (b) *CONV*

Figure 4: Box plots of $\{t_i : i = 1, ..., \acute{n}\}$, the standardised errors of predictions of a sum of the outputs, using data simulated from Gaussian processes with *CONV* and *LMC* covariance functions.

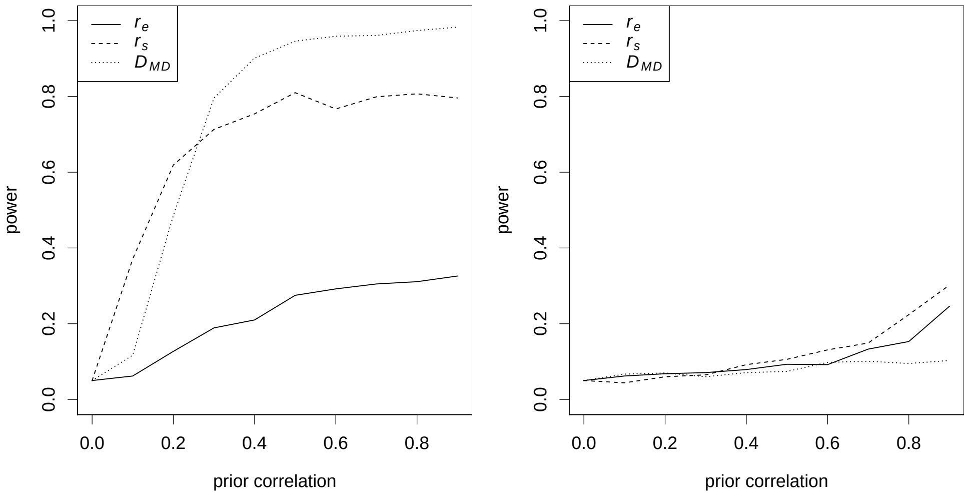


(a) *LMC* (b) *CONV*

Figure 5: Box plots of prediction Mahanalobis distances, using data simulated from Gaussian processes with *CONV* and *LMC* covariance functions. The legend to the right shows which quantiles are represented in the plots. Shaded horizontal bars in the background show the same quantiles for the theoretical distribution of the Mahanalobis distance for two independent outputs.

deviation from the reference distribution increases as the true prior correlation increases, showing that this too can be used as indication of positive between-outputs correlation.

The powers of hypothesis tests constructed using $r_{\mathbf{e}}$, $r_{\mathbf{s}}$ and D_{MD} are compared in Figure 6. The tests are conducted at the 5% level, and, for a given value of true between-outputs correlation, the power is the proportion of tests that reject the null hypothesis of uncorrelated outputs. When the true covariance function is an *LMC* type, the test using D_{MD} is the most



(a) *LMC* (b) *CONV*

Figure 6: Power of the $r_{\mathbf{e}}$, $r_{\mathbf{s}}$ and D_{MD} hypothesis tests, using data simulated from Gaussian processes with *CONV* and *LMC* covariance functions.

powerful for true correlations greater than 0.3, followed by the test using $r_{\mathbf{s}}$. The test using $r_{\mathbf{e}}$ proves to have very low power. When the true covariance function is an *CONV* type, none of the tests have high power. This is because the posterior correlation with this covariance function is generally small. Interestingly, the test using D_{MD} has the lowest power in this situation.

We conclude that when we have enough data to partition into training and validation sets, the between-outputs sample correlation coefficient calculated using de-correlated prediction errors, $r_{\mathbf{s}}$ seems to be the most useful diagnostic quantity. It can be used in a formal hypothesis test, or simply as an estimate of the between-outputs correlation coefficient. The Mahanalobis distance D_{MD} may also be used in a hypothesis test. The standardised errors of predictions of a sum of the outputs, $\{t_i : i = 1, ..., \acute{n}\}$ can also provide an indication of inadequacy of the independent outputs model.

3.2 Cross-validation

In many applications the number of available observations of the computer model may be such that we cannot afford to hold a number back from the training data for the purposes of validation. When this is the case, a solution is to validate using cross-validation techniques (see, for example, Rougier et al., 2009). Here we propose a leave-one-out (LOO) cross-validation scheme that computes a type of correlation coefficient statistic. We also propose a method for simulating a reference distribution for the statistic under the hypothesis of independence, providing means for carrying out a formal hypothesis test. The scheme is as follows.

- Given data $\mathbf{y_1}$ and $\mathbf{y_2}$, obtain estimates of the correlation hyperparameters for $\eta_1(.)$ and $\eta_2(.)$, $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ respectively.
- For $i = 1, 2, ..., n$ (where n is the total number of training data),
    1. Select the ith design point $\mathbf{x_i}$ and corresponding data points $(\mathbf{y_1})_i$ and $(\mathbf{y_2})_i$ and omit to form the reduced data denoted $\{X_{-i}, (\mathbf{y_1})_{-i}, (\mathbf{y_2})_{-i}\}$.

2. Obtain the posterior processes conditional on $\{X_{-i}, (\mathbf{y_1})_{-i}, (\mathbf{y_2})_{-i}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\}$
3. For $j = 1, 2$, let $(\mathbf{e}_j^{CV})_i = (\mathbf{y_j})_i - (\mathbf{m}^{\ddagger}_{-i}(\mathbf{x_i}))_j$, where $(\mathbf{m}^{\ddagger}_{-i}(\mathbf{x_i}))_j = \mathbb{E}[\eta_j(\mathbf{x_i})|(\mathbf{y_j})_{-i}, \hat{\boldsymbol{\theta}}_j]$.

- Let

$$r_{CV} = \frac{(\mathbf{e_1}^{CV})^T \mathbf{e_2}^{CV}}{\sqrt{(\mathbf{e_1}^{CV})^T \mathbf{e_1}^{CV}} \sqrt{(\mathbf{e_2}^{CV})^T \mathbf{e_2}^{CV}}}, \tag{22}$$

  where $\mathbf{e_1}^{CV}$ and $\mathbf{e_2}^{CV}$ are the vectors formed from the residuals $\{(\mathbf{e}_1^{CV})_i : i = 1, ..., n\}$ and $\{(\mathbf{e}_2^{CV})_i : i = 1, ..., n\}$ respectively.

We note that if we let

$$\tilde{\mathbf{y}}_\mathbf{1} = \frac{\mathbf{y_1} - \mathrm{H}\beta_1}{\tilde{\sigma}_1^2} \tag{23}$$

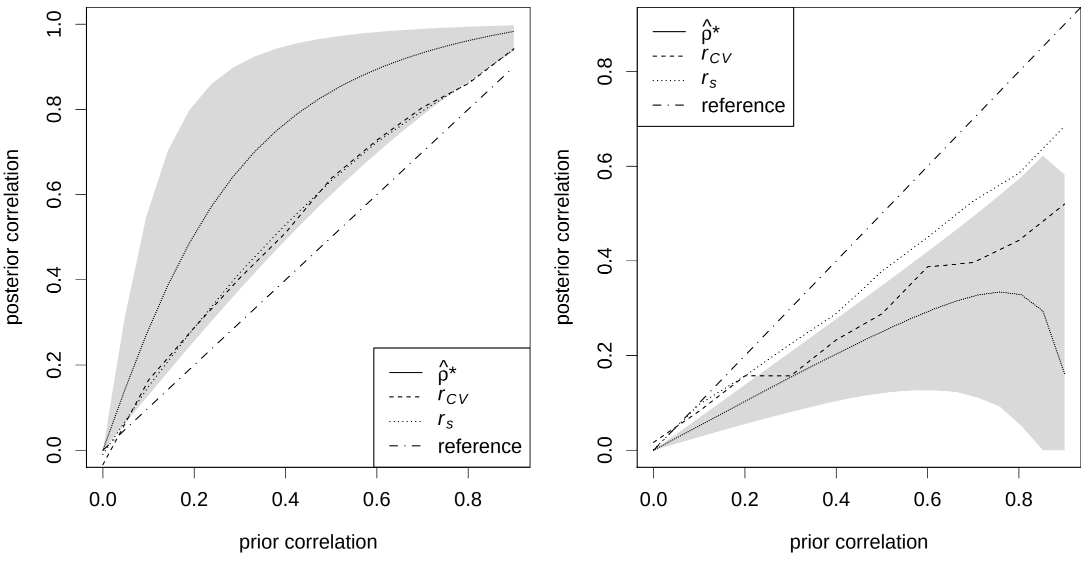
$$\tilde{\mathbf{y}}_\mathbf{2} = \frac{\mathbf{y_2} - \mathrm{H}\beta_2}{\tilde{\sigma}_2^2}, \tag{24}$$

where for $j = 1, 2,\ \tilde{\beta}_j$ is any $q \times 1$ vector and $\tilde{\sigma}_j^2$ is any positive scalar, then r_{CV} is unchanged if we replace $(\mathbf{y_1}, \mathbf{y_2})$ with $(\tilde{\mathbf{y}}_\mathbf{1}, \tilde{\mathbf{y}}_\mathbf{2})$ in the above scheme. That means we can obtain from the distribution of r_{CV} under the hypothesis of independence by simulating realisations from the models

$$\mathbf{z_1} \sim N_n[\mathbf{0}, c_1(X, X)], \tag{25}$$

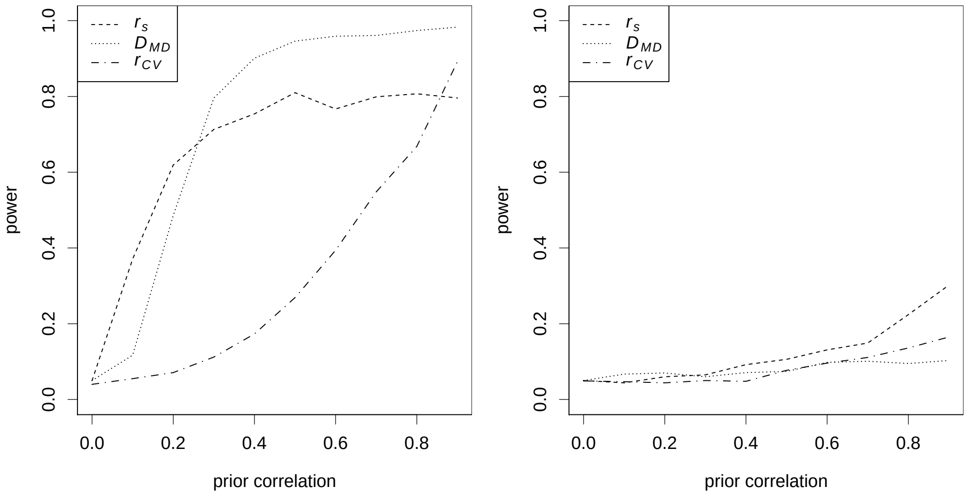
$$\mathbf{z_2} \sim N_n[\mathbf{0}, c_2(X, X)], \tag{26}$$

where $c_1(.,.)$ is the correlation function with hyperparameters $\hat{\boldsymbol{\theta}}_1$ and $c_2(.,.)$ is the correlation function with hyperparameters $\hat{\boldsymbol{\theta}}_2$, and compute r_{CV} for each realisation.



(a) *LMC* (b) *CONV*

Figure 7: The cross-validation correlation estimate r_{CV} (dashed line) with the true posterior correlation averaged over the input space (solid line)and pointwise 95% quantiles of the distribution of the posterior correlation over the input space (grey regions).



(a) *LMC*

(b) *CONV*

Figure 8: Power of the r_{CV}, $r_{\mathbf{s}}$ and D_{MD} hypothesis tests, using data simulated from Gaussian processes with *CONV* and *LMC* covariance functions.

We demonstrate this cross-validation scheme using data simulated from the same Gaussian processes as previously, using 20 training data in a Sobol sequence design. We generate a large number of data sets and compute r_{CV} for each. Figure 7 shows the mean value of r_{CV} for a range of prior between-outputs correlations. We see that on average, r_{CV} provides an estimate of the posterior between-outputs correlation that is correct for at least part of the input space. It is, however, some way off the spatially averaged between-outputs correlation. Figure 8 compares the power of the r_{CV} hypothesis test with the powers of the $r_{\mathbf{s}}$ and D_{MD} hypothesis tests from the last section. It compares favourably when the true covariance function is a *CONV* type, but the r_{CV} hypothesis test is considerably less powerful when the true covariance function is an *LMC* type. This is because r_{CV} is not adjusted to account for the spatial autocorrelation that is present in the outputs.

4 Conclusions

In this report we have introduced methods for assessing the adequacy of using independent univariate emulators for a multiple output computer model. A key point that sets this problem apart from the general problem of fitting multivariate linear models with correlated residuals is that the variables of interest (the computer model outputs) are observed with no error. As a result, existing methods that make use of 'fitted residuals' (i.e. the difference between observations and the fitted model) are not applicable. Instead, we must validate the independent outputs approach using the emulator's predictions of future observations.

Accordingly, we have proposed methods that partition the available data into training and validation sets, judging the emulator trained using the training data on its ability to make joint-output predictions of the validation data. Note that we have not discussed the question of how to decide the relative sizes of the training and validation sets; this is a subject for further research. We have also proposed a cross-validation method, for use when the data are too sparse to allow the partitioning approach.

We have found that when we implement the methods as formal hypothesis tests, it is important to choose test statistics that account for the autocorrelation that exists over the input space. Ignoring the autocorrelation results inflation of the sampling variance of the statistic and a loss of power of the test.

An interesting finding in this paper is that when a stationary multivariate Gaussian process has a nonseparable covariance function, the between-outputs correlation is no longer stationary after conditioning on observations. Moreover, the between-outputs correlation at a given point will change as we increase the number of observations, either increasing or decreasing depending on the structure of the covariance function. We found empirical evidence that the posterior between-outputs correlation increases with the number of observations for an *LMC* covariance function, and decreases with the number of observations for a *CONV* covariance function. Finding theoretical results to prove that this is the case in general, and to shed light on why it occurs, is an area for further work.

References

Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science*, **7 (1)**: 131–153.

Ahmad, I. and Li, Q. (1997). Testing independence by nonparametric kernel method, *Statistics & Probability Letters*, **34 (2)**: 201–210.

Bastos, L. S. and O'Hagan, A. (2009). Diagnostics for gaussian process emulators, *Technometrics*, **51 (4)**: 425–438.

Bhattacharya, S. (2007). A simulation approach to Bayesian emulation of complex dynamic computer models, *Bayesian Analysis*, **2**: 783–816.

Blum, J., Kiefer, J. and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function, *The annals of mathematical statistics*, **32 (2)**: 485–498.

Cerioli, A. (2002). Testing mutual independence between two discrete-valued spatial processes: A correction to Pearson chi-squared, *Biometrics*, **58 (4)**: 888–897.

Conti, S. and O'Hagan, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models, *Journal of statistical planning and inference*, **140 (3)**: 640–651.

Csörgő, S. (1985). Testing for independence by the empirical characteristic function, *Journal of Multivariate Analysis*, **16 (3)**: 290–299.

Currin, C., Mitchell, T., Morris, M. and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments, *Journal of the American Statistical Association*, **86**: 953–963.

Dejean, J. and Blanc, G. (1999). Managing uncertainties on production predictions using integrated statistical methods, in *Society of Petroleum Engineers Annual Technical Conference and Exhibition*.

Drucker, H., Burges, C., Kaufman, L., Smola, A. and Vapnik, V. (1997). Support vector regression machines, *Advances in neural information processing systems*, pp. 155–161.

El Tabach, E., Lancelot, L., Shahrour, I. and Najjar, Y. (2007). Use of artificial neural network simulation metamodelling to assess groundwater contamination in a road project, *Mathematical and Computer Modelling*, **45 (7-8)**: 766–776.

Fieller, E. C., Hartley, H. O. and Pearson, E. S. (1957). Tests for rank correlation coefficients. i, *Biometrika*, **44 (3/4)**: 470–481.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika*, **10 (4)**: 507–521.

Fricker, T. E., Oakley, J. and Urban, N. M. (2010). Multivariate Emulators with Nonseparable Covariance Structures, MUCM Technical Report.

Gelfand, A. E., Schmidt, A. M., Banerjee, S. and Sirmans, C. F. (2004). Nonstationary multivariate process modelling through spatially varying coregionalization, *Test*, **13 (2)**: 1–50.

Goulard, M. and Voltz, M. (1992). Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix, *Journal Mathematical Geology*, **21 (3)**: 269–286.

Haslett, J. and Hayes, K. (1998). Residuals for the linear model with general covariance structure, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60 (1)**: 201–215.

Haugh, L. (1976). Checking the independence of two covariance-stationary time series: a univariate residual cross-correlation approach, *Journal of the American Statistical Association*, **71 (354)**: 378–385.

Hawkins, D. L. (1989). Using u statistics to derive the asymptotic distribution of fisher's z statistic, *The American Statistician*, **43 (4)**: 235–237.

Haylock, R. G. and O'Hagan, A. (1996). *On Inference for Outputs of Computationally Expensive Algorithms with Uncertainty on the Inputs*, Bayesian Statistics 5, Oxford University Press.

Hoeffding, W. (1948). A non-parametric test of independence, *The Annals of Mathematical Statistics*, **19 (4)**: 546–557.

Hong, Y. (1996). Testing for independence between two covariance stationary time series, *Biometrika*, **83 (3)**: 615.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*, Academic Press.

Kankainen, A. and Ushakov, N. (1998). A consistent modification of a test for independence based on the empirical characteristic function, *Journal of Mathematical Sciences*, **89 (5)**: 1486–1494.

Karvanen, J. (2005). A resampling test for the total independence of stationary time series: Application to the performance evaluation of ica algorithms, *Neural Processing Letters*, **22 (3)**: 311–324.

Kennedy, M., Anderson, C., O'Hagan, A., Lomas, M., Woodward, I., Gosling, J. and Heinemeyer, A. (2008). Quantifying uncertainty in the biospheric carbon flux for England and Wales, *Journal of the Royal Statistical Society: Series A(Statistics in Society)*, **171 (1)**: 109–135.

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion), *Journal of the Royal Statistical Society. Series B*, **63**: 425–464.

Kuipers, L. and Niederreiter, H. (2006). *Uniform distribution of sequences*, Dover Publications.

Rao, J. and Scott, A. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables, *Journal of the American Statistical Association*, **76 (374)**: 221–230.

Rosenblatt, M. and Wahlen, B. (1992). A nonparametric measure of independence under a hypothesis of independent components, *Statistics & Probability Letters*, **15 (3)**: 245–252.

Rougier, J. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations, *Climatic Change*, **81 (3)**: 247–264.

Rougier, J., Guillas, S., Maute, A. and Richmond, A. (2009). Expert knowledge and multivariate emulation:the thermosphere-ionosphere electrodynamics general circulation model (tie-gcm), *Technometrics*, **51 (4)**: 414–424.

Rutherford, A., Inman, D., Park, G. and Hemez, F. (2005). Use of response surface metamodels for identification of stiffness and damping coefficients in a simple dynamic system, *Shock and Vibration*, **12 (5)**: 317–331.

Sacks, J., Welch, W., Mitchell, T. and Wynn, H. (1989). Design and analysis of computer experiments, *Statistical Science*, **4**: 409–435.

Wackernagel, H. (1995). *Multivariate Geostatistics*, Springer.

Wang, L., GmbH, S. and Springer, N. (2005). Support Vector Machines: theory and applications.

Wu, E., Yu, P. and Li, W. (2009). A smoothed bootstrap test for independence based on mutual information, *Computational Statistics & Data Analysis*, **53 (7)**: 2524–2536.