Log Contrast Multivariate Indicator Cokriging

Alexis Boukouvalas, Dan Cornford

Abstract

In this report we present an outline of a conditional density emulator. Our approach is based on indicator cokriging developed in the geostatistics literature. A discrete approximation to the continuous conditional density is calculated using an independent emulator for each discretized "slice". To ensure the output is always a valid distribution and to make the independence assumption between slices more realistic, the simplex space defining the probability densities are transformed to a log contrast space where the emulators operate. The back transformation to the simplex space is accomplished using a scaled unscented transformation.

Contents

1 Introduction

In this report we tackle the problem of conditional density estimation. In our previous work we have proposed a heteroscedastic emulator which allows for more flexible modelling of the variance. However this model still assumes the output distribution is Gaussian and the Rabies disease model we are investigating, exhibits highly skewed distributions and in some regions of the parameter space multi-modality. Our proposed method is non-parametric and makes no assumptions on the form of the output distribution.

In terms of notation we denote the stochastic scalar output code simulator f; in our motivating example this is the Rabies model. For a given input design point $\mathbf{x}$ the conditional distribution is $p(f|\mathbf{x})$. The marginal distribution across all input design space is $p(f)$.

We wish to describe non-parametrically the conditional distribution $p(f|\mathbf{x})$. We use the indicator kriging approach (Pardo-Iguzquiza and Dowd, 2005) where we discretize the distribution into k slices and emulate the slices using a Gaussian Process (GP) either independently or jointly (cokriging). Let the probabilities be denoted by p_k. For example in the Rabies model with three slices separated at 10 and 100 years:

- $p_1(\leq 10 years)$ probability of disease extinction within 10 years

- $p_2(> 10 years \ AND \leq 100 years)$

- $p_3(> 100 years)$ probability of disease extinction after 100 years. This includes the cases where the disease does not become extinct.

The slices define a simplex:

$$p_i \ \geq \ 0 \tag{1}$$

$$\sum_{i=1}^{k} p_i \ = \ 1 \tag{2}$$

We could emulate each slice independently using a GP. However the slices are highly correlated, in fact being only $k-1$ free parameters. An alternative would be to build a multi-output GP emulator. The problem with this approach has to do with the simplex constraints defined in Equations (1) and (2). Building a multi-output GP that respects the non-negative constraint is straightforward (Equation (1)); e.g. using a log transformation. However to guarantee the sum to unity constraint (Equation (2)) is non-trivial and most closely related to the problem of multi-class classification in the machine learning community (Chapter 3 (Williams and Rasmussen, 2006)). We are not aware of an efficient solution that satisfies the unity constraint.

We investigate another approach to transform the slices into $k-1$ independent variables ϕ_i.

2   Existing approaches

In this section we briefly describe existing approaches that have been proposed in the literature to address the conditional density estimation problem.

1. *Indicator cokriging* (Pardo-Iguzquiza and Dowd, 2005). In traditional indicator cokriging, a multivariate GP is used to estimate all k slices jointly, followed by a smoothing step where all cumulative distribution function order violations are corrected.

2. *Mixture Density Networks* (Bishop, 1994). This is a Gaussian Mixture Model (GMM) where the mixture parameters are predicted using a neural network. Traditionally plug-in estimates are used for the parameters of the neural network. The problem with the Mixture Density Networks (MDN) approach is that in order to obtain the predictive uncertainty, a Bayesian approach is needed, typically requiring expensive Monte Carlo.

3. *Warping* (Snelson et al., 2004). Warping allows the transformation of the GP output by a certain class of non-linear functions. The function parameters are jointly estimated with the GP parameters. In this framework non-Gaussian output distributions can be modelled. However in the original framework there are no constraints imposed to ensure the output is a valid probability.

4. *Copulas* (Bardossy and Li, 2008). Copulas link multivariate distributions to their one-dimensional marginals. They have been successfully applied to indicator kriging since they naturally give rise to valid distributions. The main issue is computational cost. In particular for the multiplied Gaussian and central χ^2 copulas which are commonly used in spatial statistics, the predictive distribution requires the evaluation of 2^n terms where n is the number of training points. In practice, only local prediction is performed where the closest m neighbouring points to the prediction site are used where m is set according to the computational budget available.

5. *Kernel Quantile Regression* (Takeuchi et al., 2009). In this work, the cumulative distribution is predicted in a piecewise-linear form using Kernel Quantile Regression (KQR). The approach however requires smoothing of the predictions to correct for order violations.

3 Outline

We now provide an overview of our method and our experimental plan.

3.1 Description of approach

1. Discretize the cumulative distribution (CDF) or probability density (PDF) into k slices.

2. Transform into $k - 1$ log contrasts using orthonormal coefficient vectors to ensure contrasts are independent (assuming a uniform prior Dirichlet - see Section 4).

3. Construct $k - 1$ independent emulators.

4. Back transform using generalized softmax (Section 7). The transformation from log-contrast space to the simplex is non-trivial. The simplest would be a Monte Carlo scheme. However since we only require the first two moments for each k a more efficient sampling can be achieved by using unscented methods. Another possibility inspired by multi-class classification in the machine learning community would be to employ a deterministic approximation to the transformation such as expectation propagation. We initially plan to use unscented methods due to their simplicity and theoretical performance guarantees.

We note that our method suffers from the curse of dimensionality. The number of slices k required for a fixed level of accuracy increases exponentially with the dimensionality of the input.

Our proposal is computationally efficient compared to the MDN and Copula approaches and does not require smoothing to correct order violations unlike the KQR and traditional indicator kriging methods. In cases where we do not a priori know the loss function exactly; the model we build is general and can accommodate different loss functions.

3.2 Planned Experiments

We plan to test our method on both synthetic and application data and compare its performance against existing approaches. Specifically we plan the following experiments:

- Synthetic datasets as can be found in (Takeuchi et al., 2009).

- For the application we will use the Rabies model with 1 output (time to extinction).

- We will compare independent and multivariate (separable) GPs in log contrast space to check the effectiveness of using log contrasts to achieve independence.

- Lastly, we will compare the performance of our method with MDN and possibly copula type approaches. We are also considering a comparison to a warped heteroscedastic GP (Boukouvalas and Cornford, 2009). This will relax the Gaussian assumption made by the heteroscedastic model but will require post-process smoothing to correct order violations.

We proceed in the following sections to explain in detail each of these stages.

4 Log contrasts

In Section 12.7 of (O'Hagan, 2004) log contrasts are described as a way to compare probabilities.

In the indicator approach we approximate a possibly continuous distribution by a discrete set[1].

In the notation of (O'Hagan, 2004) the p_i are written as θ_i and we will follow the book's notation from this point forward.

Given n independent observations the likelihood is multinomial:

$$f(\mathbf{n}|\theta) = \frac{n!}{n_1!n_2!\dots n_k!}\Pi_{j=1}^k\theta_j^{n_j}$$

where n_j the number of observations that correspond to the outcome θ_j and the vectors $\mathbf{n} = [n_1\dots n_k]$, $\theta = [\theta_1\dots\theta_k]$. The conjugate prior distribution is the Dirichlet:

$$f(\theta) = B(\mathbf{a})^{-1}\Pi_{j=1}^k\theta_j^{a_j-1}$$

where $\mathbf{a} = [a_1,\dots,a_k]$ and

$$B(\mathbf{a}) = B(a_1\dots a_k) = \frac{\Pi_{j=1}^k\Gamma(a_j)}{\Gamma(\sum_{j=1}^k a_j)}$$

Given a Dirichlet prior $f(\theta) = D(\mathbf{a})$ and a multinomial likelihood on the data Y, the posterior is $f(\theta|Y) = D(\mathbf{a}+\mathbf{n})$.

A log contrast is specified as

$$\phi = \sum_{j=1}^k c_j log(\theta_j) \tag{3}$$

where c_j are the linear coefficients subject to the constraint $\sum_{j=1}^k c_j = 0$.

If we suppose a Dirichlet prior for θ, i.e. $f(\theta) = D(\mathbf{a})$, then in Section 12.8 of (O'Hagan, 2004) the moment generating function and first two moments of ϕ are given. A series expansion of the moment generating function of ϕ shows that it can be approximated with a normal distribution and a remainder term that is of order a_j^{-2}. Thus given sufficiently large a_j's we can approximate the posterior $p(\phi|\theta)$ with a normal distribution.

Given this approximation, the mean and variance of a log contrast $\phi_\alpha = \sum_{j=1}^k c_j log(\theta_j)$ and covariance with another log contrast $\phi_\beta = \sum_{j=1}^k d_j log(\theta_j)$ are:

---

[1]In fact for the Rabies model the original distribution is discrete but infinite since it is the number of time steps to extinction. In practice it's not infinite since we arbitrarily truncate to a maximum number of number of time steps.

$$E(\phi_\alpha) = \sum_{j=1}^{k} c_j log(a_j) - \sum_{j=1}^{k} \frac{c_j}{2a_j} \tag{4}$$

$$var(\phi_\alpha) = \sum_{j=1}^{k} \frac{c_j^2}{a_j} \tag{5}$$

$$cov(\phi_\alpha, \phi_\beta) = cov(\sum_{j=1}^{k} c_j log(\theta_j), \sum_{j=1}^{k} d_j log(\theta_j)) = \sum_{j=1}^{k} \frac{c_j d_j}{a_j} \tag{6}$$

Note that exact analytic expressions exist for the mean, variance and covariance which require the evaluation of the digamma/trigamma functions.

Example 12.1 in (O'Hagan, 2004) shows how to get independence between log contrasts. Let $f(\theta) = D(a_0, \ldots a_0)$ so all a_j's are equal and suppose a_0 is large so that the approximation holds reasonable well. Let $\phi_1, \phi_2, \ldots, \phi_{k-1}$ log contrasts which are *orthonormal*, i.e.

$$\sum_{j=1}^{k} c_j = 0$$

$$\sum_{j=1}^{k} c_j^2 = 1$$

$$\sum_{j=1}^{k} c_{ij} c_{hj} = 0$$

for all i and $h \neq i$.

The above constraints can be written in vector form. Let vector $\mathbf{c^i} = [c_1 c_2 \ldots c_k]$:

$$|\mathbf{c}^i|_{L1} = 0$$
$$||\mathbf{c}^i||_{L2} = 1$$
$$(\mathbf{c}^i)^T \mathbf{c}^h = 0$$

where $L1$ and $L2$ are the respective norms and T the transpose..

Given these constraints, the log contrasts ϕ_i are approximately independent $N(0, a_0^{-1})$ random variables. This can be seen clearly if we substitute a_0 with a_j in Equations (4), (5) and (6). The independence stems from zero covariance given in Equation (6) which can be achieved in two ways:

- As in the example $a_j = a_0$ for all j and the $\mathbf{c}_i$ and $\mathbf{c}_j$ coefficient vectors are orthogonal.

- For the covariance to be zero $\sum_{j=1}^{k} \frac{c_j d_j}{a_j} = 0$ for any two coefficient vectors $\mathbf{c}$ and $\mathbf{d}$. This condition is more awkward since it is no longer simple orthogonality between two vectors but allows for a more flexible distribution specification for θ.

We follow the first approach since it is mathematically simpler. However we may look into how to enforce the latter condition to allow for more flexible distributions.

One problem with assuming $\theta \sim D(a_0 \ldots a_0)$ is the imposition of a fixed distribution on the individual θ_j's:

$$
\begin{aligned}
E(\theta_j) &= \frac{1}{k} \\
var(\theta_j) &= \frac{k-1}{a_0(ka_0+1)} \\
cov(\theta_j, \theta_l) &= -\frac{1}{k^2(ka_0+1)}
\end{aligned}
$$

If we go with the second approach and can specify arbitrary a_j then we can set a prior mean for each θ_j and overall prior strength $a = \sum_{i=1}^{k} a_j$ - see paragraph 12.11 in (O'Hagan, 2004) for more information.

To emulate using contrasts we have to tackle the following issues:

- *Section 5*: Given k slices, we need to construct $k-1$ orthogonal log contrast functions.

- *Section 6*: Having done that we can build $k-1$ independent emulators to predict ϕ_i.

- *Section 7*: In prediction the outputs are back transformed to k θ_i's. This is non-trivial.

We first present some early experimental results to check for independence in the log contrast space.

4.1 Experiment: ϕ's are uncorrelated

The experiment was conducted as follows:

- Generate N 3-D θ vectors subject to sum to unity and non-negative constraints where N is the number of samples shown in the x axis of Figure 1.

- Generate two orthonormal coefficient vectors $\mathbf{c}^1 = [0.5, -0.8090, 0.3090]^T$ and $\mathbf{c}^2 = [0.6454, 0.11027, -0.75576]^T$.

- Calculate $\phi^1 = (\mathbf{c}^1)^T log(\theta)$ and $\phi^2 = (\mathbf{c}^2)^T log(\theta)$ for all samples N.

- Calculate empirical correlation between ϕ^1 and ϕ^2.

We repeat this cycle 1000 times to get reliable estimated of the mean and variance of the correlation. The individual θ_j can be highly correlated. For example for 10^4 number of sample, we have $corr(\theta_1, \theta_2) = -0.6537$, $corr(\theta_1, \theta_3) = -0.6525$ and $corr(\theta_2, \theta_3) = -0.1469$. The results are shown in Figure 1. We observe that as the number of samples increases, the correlations become consistently quite small.
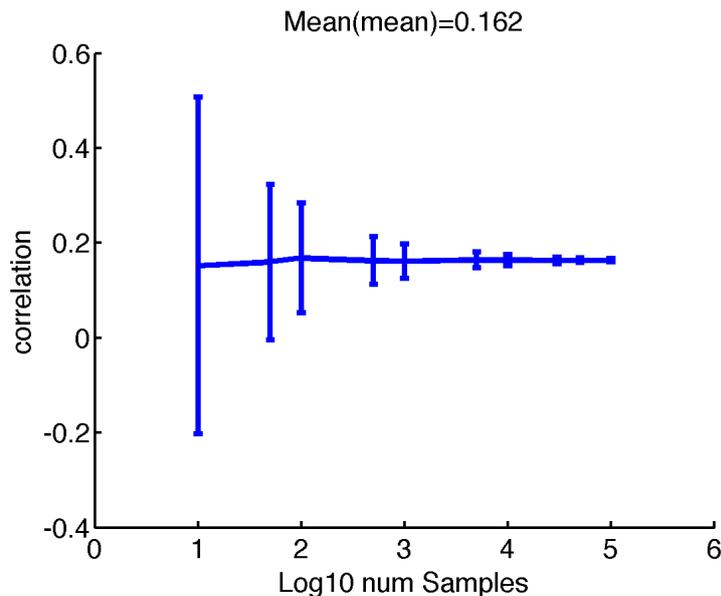
Figure 1: Correlation of ϕ. 1000 repetitions for each sample size. θ is 3D and ϕ is 2D.

5   Finding the basis

In order to assume independence among log contrasts we require orthonormal coefficient basis vectors. We now describe how to obtain these vectors for any k. The computation is performed at the beginning of the emulation construction.

The following constraints have to be satisfied:

$$|\mathbf{c}^i|_{L1} = 0 \tag{7}$$
$$||\mathbf{c}^i||_{L2} = 1 \tag{8}$$
$$(\mathbf{c}^i)^T \mathbf{c}^h = 0 \tag{9}$$

We adapt the proof in (Schey, 1985) to state the following theorem which shows how to construct the required $k-1$ log contrast coefficient vectors.

Theorem 1: Let $\phi_0 = [(1/k)]$ k dimensional vector. The $k-1$ eigenvectors corresponding to the $k-1$ non-zero eigenvalues of the matrix $\mathbf{W} = \mathbf{I}_k - \phi_0\phi_0^T$ form a orthonormal log contrast set.

Proof

1. Construct the matrix $\mathbf{W} = \mathbf{I}_k - \phi_0\phi_0^T$ where $\mathbf{I}_k$ is the $k \times k$ identity matrix. We note that $\phi_0^T \phi_0 = 1$ and that $\mathbf{W}$ is a symmetric idempotent matrix (by construction).

2. ϕ_0 is the corresponding eigenvector for the zero eigenvalue of $\mathbf{W}$. Proof:

$$\mathbf{W}\phi_0 = (\mathbf{I}_k - \phi_0\phi_0^T)\phi_0$$
$$= \phi_0 - \phi_0(\phi_0^T \phi_0)$$
$$= 0$$

3. Find the orthonormal eigenvectors $\phi_1 \ldots \phi_{k-1}$ for all other eigenvalues of $\mathbf{W}$. Since $\mathbf{W}$ is idempotent all other eigenvalues exist and are equal to 1.

4. Note that $\phi_j^T \phi_0 = 0$ for $j \neq 0$ since they correspond to different eigenvalues of $\mathbf{W}$.

5. Since ϕ_j orthogonal to ϕ_0 for $j \neq 0$:

$$
\begin{aligned}
\phi_j^T \phi_0 &= 0 \Leftrightarrow \\
\sum_{i=1}^{k} \phi_{ji} \phi_{0i} &= 0 \Leftrightarrow \\
\sum_{i=1}^{k} \phi_{ji} \frac{1}{\sqrt{k}} &= 0 \Leftrightarrow \\
\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \phi_{ji} &= 0 \Leftrightarrow \\
\sum_{i=1}^{k} \phi_{ji} &= 0
\end{aligned}
$$

6. Hence the $k-1$ $\phi_1 \ldots \phi_{k-1}$ eigenvectors of $\mathbf{W}$ satisfy the log contrast constraint (7).

6 Training the emulators

Having fixed the coefficient vectors, we may proceed with the inference of the $k-1$ emulators.

We estimate the discrete cumulative distribution θ from data and calculate $k-1$ ϕ_i's using Equation (3). The experimental design used to estimate θ remains an open issue (Section 8). Initially we propose to use a standard space filling design with many replicates per design point to allow for a reasonable at-a-point estimation of the CDF.

A technical issue that has to be handled during the transformation is the case of zero probabilities which we propose to substitute with a small value ϵ. Assuming sample size N, the smallest value per slice is $log(1/N)$ so we pick ϵ some order of magnitudes smaller.

7 Prediction - Finding θ

Taking the $k-1$ emulator predictions for ϕ_i we have to solve the following system to get the original θ:

$$
\begin{aligned}
\phi_1 &= (\mathbf{c^1})^T log(\theta) \\
\phi_{k-1} &= (\mathbf{c^{k-1}})^T log(\theta) \\
|\theta|_{L1} &= 1
\end{aligned}
$$

In scalar notation:

$$\phi_1 = \sum_{i=1}^{k} \mathbf{c_i^1} log(\theta_i)$$

$$\phi_{k-1} = \sum_{i=1}^{k} \mathbf{c_i^{k-1}} log(\theta_i)$$

$$\sum_{i=1}^{k} \theta_i = 1$$

In 12.15 of (O'Hagan, 2004) the solution is given as:

$$\theta_{\mathbf{M}} = \frac{exp(\mathbf{C}^{-1}\phi)}{1 + \mathbf{1}'exp(\mathbf{C}^{-1}\phi)} \tag{10}$$

where $\theta_{\mathbf{M}}$ the vector $\theta_1, \ldots, \theta_{k-1}$, $\mathbf{C}$ the matrix of $k - 1$ coefficients, and $\mathbf{1}$ a $(k-1) \times 1$ vectors of ones. We refer to this solution as generalized softmax.

During prediction for a set of test points X_* given a training set D, for each ϕ_i, $i \in [1, \ldots, k-1]$, we obtain a mean prediction $E[\phi_i|X_*, D]$ and a variance $Var[\phi_i|X_*, D]$ using $k - 1$ independent GPs. To propagate these moments through the mapping defined by Equation (10) we propose to use unscented methods to perform a deterministic sampling (van der Merwe et al., 2001).

The estimates of the mean and covariance obtained using the Unscented Transformation (UT) are accurate to the second order for any non-linear function (van der Merwe et al., 2001). Let n_* be the number of test points, i.e. locations where we wish to estimate the conditional density distribution.

We propose to use the Scaled Unscented Transformation (SUT) which ensures the predicted covariance is positive semi-definite. The method is described in Figure 2 and is adapted from (van der Merwe et al., 2001).

Three parameters control the SUT algorithm. $\kappa \geq 0$ is a scaling parameter which is typically set to 0. $\alpha \in [0, 1]$ controls the spread of the sigma point distribution. $\beta \geq 0$ is typically used to incorporate prior knowledge of higher order moments such as kurtosis. We typically use $\beta = 2$ which is appropriate for a Gaussian prior.

Lastly, we note than in many cases we may not wish to estimate the variance in the simplex space. If we wish to display error bars, propagating quantiles through the monotonic transformation defined by Equation (10) is more appropriate since variance provides a symmetric measure of uncertainty which can violate the simplex constraints. If the density estimation is performed as part of a decision problem, we may wish to sample valid output distributions. In our opinion, the most efficient way to achieve such a sampling is to directly sample from the emulators in the log contrast space and deterministically propagate the sample distribution through the transformation.

8 Conclusions and Open Questions

In this report we have provided an overview of our conditional density estimation method which is based on the usage of orthonormal coefficient vectors for the log contrasts to guar-

The Scaled Unscented Transformation (SUT) algorithm.

Data Input: A training set D, a set of test points X_* of size n_*, for each ϕ_i, $i \in [1, \ldots, k-1]$, a mean prediction $M_i^\phi = E[\phi_i | X_*, D]$ and a variance $V_i^\phi = Var[\phi_i | X_*, D]$ using $k-1$ independent GPs, a non-linear map function g defined by Equation (10).

Algorithm Parameters: Three scalar parameters are required $\{\alpha, \beta, \kappa\}$

Output: The propagated mean M_i^θ and variance V_i^θ in the simplex space, θ_i, $i \in [1, \ldots, k]$. Note that the mapping in Equation (10) defines only the first $k-1$ θ's with $\theta_k = 1 - \sum_{i=1}^{k-1}$. Therefore $V_k^\theta = 0$.

A. *Preprocessing*

- For log contrast ϕ_i calculate the Cholesky decomposition L of the covariance matrix, i.e. $V_i^\phi = LL^T$.

- Set $\lambda = a^2(n_* k) - n_*$.

B. *Sigma Points*: Define $2n_* + 1$ sigma points to propagate through the non-linear mapping g.

$$\mathbf{x}_0 = M_i^\phi \tag{11}$$

$$\mathbf{x}_i = M_i^\phi + (\sqrt{n_* + \lambda}L)_i, i = 1, \ldots, n_* \tag{12}$$

$$\mathbf{x}_i = M_i^\phi - (\sqrt{n_* + \lambda}L)_i, i = n_* + 1, \ldots, 2n_* \tag{13}$$

where $(\sqrt{n_* + \lambda}L)_i$ the ith row of the symmetric matrix.

C. *Weightings*: For each sigma point define a weighting with the property that $\sum_{i=0} 2n_* W_i = 1$.

- For the first sigma point two weightings are defined, for the mean (m) and variance (v) reconstruction respectively.

$$W_0^m = \frac{\lambda}{n_* + \lambda} \tag{14}$$

$$W_0^v = \frac{\lambda}{n_* + \lambda} + (1 - \alpha^2 + \beta) \tag{15}$$

- For all other sigma points $W_i = \frac{1}{2(n_* + \lambda)}$.

D. *Propagation*: Propagate each sigma point through the non-linear function $\mathbf{y}_i = g(\mathbf{x}_i)$.

E. *Reconstruction*: The estimates of the mean and variance in the simplex space are:

$$M_i^\theta = W_0^m \mathbf{y}_0 + \sum_{i=1}^{2n_*} W_i \mathbf{y}_i \tag{16}$$

$$V_i^\theta = W_0^v (\mathbf{y}_0 - M_0^\theta)(\mathbf{y}_0 - M_0^\theta)^T + \sum_{i=1}^{2n_*} W_i (\mathbf{y}_i - M_i^\theta)(\mathbf{y}_i - M_i^\theta)^T \tag{17}$$

F. *Goto A until all ϕ_i have been processed.*

Figure 2: The Scaled Unscented Transformation (SUT) algorithm used to propagate the mean and variance from the log contrast space ϕ to the simplex space θ.

antee output of valid probability distributions without the need for post process smoothing as well as allowing us to reduce the computational complexity by making a simplifying independence assumption in the log contrast space.

There are a number of unresolved issues not addressed in this report:

1. *How to determine the number of slices k?* Setting aside the issue of computational complexity, for simple distributions discretization can be quite coarse. This also raises the issue of heterogeneity where the number of slices is non-constant and could be a function of the input space.

2. *How to deal with zero probabilities?* The log transformation requires an explicit method to handle this situation. Our current thinking is to avoid the issue by substituting the zeroes with a small number.

3. *Zero uncertainty.* The last slice is currently predicted with zero variance which is undesirable since it is an artifact of our technique rather than a true reflection of uncertainty. However what variance to assign to θ_k is unclear.

4. *Validation & Diagnostics.* How to compare the predicted probability distribution to the true distribution is not immediately obvious. When examining at-a-point predictions, we can compare the predictive to the empirical CDF using either the Kullback Leibler (KL) divergence or the Receiver Operating Characteristic (ROC) curve. The KL is meaningful only in a relative sense while the Area Under the Curve (AUC) for the ROC curve is informative as an absolute measure. However both measures ignore our predictive uncertainty of the CDF. One possibility would be to sample from the emulator and compute the distribution of the AUC. Another issue is how to summarize the model performance across all validation points. The average AUC is one possibility but other more informative measure may exist.

5. *Meaning of uncertainty on CDF.* Apart from validation, a question of interpretation of this uncertainty arises in the context of further utilization of the emulation results.

6. *Experimental design.* We are currently investigating the question of experimental design for the heteroscedastic model (Boukouvalas and Cornford, 2009). However the additional flexibility of the log contrast non-parametric method will likely complicate the design question further.

Acknowledgements

References

Bardossy, A. and J. Li (2008). Geostatistical interpolation using copulas. *Water Resources Research 44* (W07412).

Bishop, C. M. (1994). Mixture density networks. Technical report, University of Aston.

Boukouvalas, A. and D. Cornford (2009). Learning heteroscedastic gaussian processes for complex datasets. Technical report, University of Aston.

O'Hagan, T. (2004). *Bayesian Inference*. Kendall's Advanced Theory of Statistics. Arnold.

Pardo-Iguzquiza, E. and P. A. Dowd (2005). Multiple indicator cokriging with application to optimal sampling for environmental monitoring. *Computers & Geosciences 31*, 1–13.

Schey, H. M. (1985). A geometric description of orthogonal contrasts in one-way analysis of variance. *The American Statistician 39*(2), 104–106.

Snelson, E., C. Rasmussen, and Z. Ghahrmani (2004). Warped gaussian processes.

Takeuchi, I., K. Nomura, and T. Kanamori (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Comput. 21*(2), 533–559.

van der Merwe, R., N. de Freitas, A. Doucet, and E. Wan (2001, Nov). The unscented particle filter. In *Advances in Neural Information Processing Systems 13*.

Williams, C. K. I. and C. E. Rasmussen (2006). *Gaussian Processes for Machine Learning*. MIT Press.