School of Informatics, University of Edinburgh

A Note on Noise-free Gaussian Process Prediction with Separable Covariance Functions and Grid Designs

by

Christopher K. I. Williams, Kian Ming A. Chai, Edwin V. Bonilla
School of Informatics, University of Edinburgh
5 Forrest Hill, Edinburgh EH1 2QL, UK
c.k.i.williams@ed.ac.uk, K.M.A.Chai@sms.ed.ac.uk,
edwin.bonilla@ed.ac.uk

Informatics Research Report 1228

School of Informatics **December 2007**
http://www.informatics.ed.ac.uk/

A Note on Noise-free Gaussian Process Prediction with Separable Covariance Functions and Grid Designs

Christopher K. I. Williams, Kian Ming A. Chai, Edwin V. Bonilla
School of Informatics, University of Edinburgh
5 Forrest Hill, Edinburgh EH1 2QL, UK
c.k.i.williams@ed.ac.uk, K.M.A.Chai@sms.ed.ac.uk,
edwin.bonilla@ed.ac.uk

Informatics Research Report 1228

SCHOOL *of* INFORMATICS

December 2007

Abstract : Consider a random function f with a separable (or tensor product) covariance function, i.e. where $\mathbf{x}$ is broken into D groups $(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^D)$ and the covariance function has the form $k(\mathbf{x}, \tilde{\mathbf{x}}) = \prod_{i=1}^{D} k^i(\mathbf{x}^i, \tilde{\mathbf{x}}^i)$. We also require that observations of f are made on a D-dimensional grid. We show how conditional independences for the Gaussian process prediction for $f(\mathbf{x}_*)$ (corresponding to an off-grid test input $\mathbf{x}_*$) depend on how $\mathbf{x}_*$ matches the observation grids. This generalizes results on autokrigeability (see, e.g. Wackernagel 1998, ch. 25) to $D > 2$.

Keywords : Gaussian process prediction, separable covariance function, autokrigeability

We consider Gaussian process regression or kriging (see e.g. Rasmussen and Williams, 2006). By way of notation let the dataset of input-output pairs be denoted by $\mathcal{D} = \{(\mathbf{x}_1, f_1), \ldots (\mathbf{x}_n, f_n)\}$, where $\mathbf{x}$ is a p-dimensional vector and f is a scalar value. Let $k(\mathbf{x}, \mathbf{x}')$ be a covariance (kernel) function. If the f's are noise-free observations of the underlying function, then the prediction at a new, unobserved location $\mathbf{x}_*$ has mean and variance given by

$$\overline{f}_* = \mathbf{k}_*^T K^{-1} \mathbf{f} \stackrel{def}{=} \boldsymbol{\alpha}_*^T \mathbf{f} \tag{1}$$

$$\text{var}(f_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T K^{-1} \mathbf{k}_* \tag{2}$$

where K is the $n \times n$ matrix with entries $k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}_*$ is the $n \times 1$ vector with entries $k(\mathbf{x}_i, \mathbf{x}_*)$, $\mathbf{f}$ is the vector $(f_1, \ldots, f_n)^T$, and $\boldsymbol{\alpha} = K^{-1}\mathbf{k}_*$.

Now let $\mathbf{x}$ be divided into D groups each of length ℓ_i, so that $\sum_{i=1}^D \ell_i = p$ and $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^D)$. Note the difference between a superscript (denoting a group in $\mathbf{x}$) and a subscript (which denotes the name of an observed datapoint). A *separable* covariance function has the form

$$k(\mathbf{x}, \tilde{\mathbf{x}}) = \prod_{i=1}^{D} k^i(\mathbf{x}^i, \tilde{\mathbf{x}}^i). \tag{3}$$

This might also be called a *tensor product* covariance function (see, e.g. Ritter, 2000, section VI.2). One example of a separable covariance function is the squared exponential (or Gaussian) covariance function

$$k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp(-\sum_{i=1}^{p} (x^i - \tilde{x}^i)^2). \tag{4}$$

In this case we can choose p groups each with length 1.

Let there be a group of observation locations $\mathcal{O}^i$ for each group i, with $\mathcal{O}^i = \{\mathbf{x}_1^i, \ldots, \mathbf{x}_{n^i}^i\}$, where n^i is the number of observations in $\mathcal{O}^i$. Let the observation locations in $\mathcal{D}$ be defined by the grid $\mathcal{O}^1 \otimes \mathcal{O}^2 \otimes \ldots \otimes \mathcal{O}^D$. Now consider making a prediction at a test point $\mathbf{x}_* = (\mathbf{x}_*^1, \mathbf{x}_*^2, \ldots, \mathbf{x}_*^D)$. Which datapoints in $\mathcal{D}$ will have non-zero weight (i.e. α coefficient) will depend on whether $\mathbf{x}_*^i \in \mathcal{O}^i$, for $i = 1, \ldots, D$. We will say that $\mathbf{x}_*^i$ *matches* $\mathcal{O}^i$ if $\mathbf{x}_*^i \in \mathcal{O}^i$. Wlog, we order the groups so that $\mathbf{x}_*^i \in \mathcal{O}^i$, for $i = 1, \ldots, i_*$, and $\mathbf{x}_*^i \notin \mathcal{O}^i$, for $i = i_* + 1, \ldots, D$, i.e. it is the first i_* groups that match. If a datapoint $\mathbf{x}_j$ matches $\mathbf{x}_*$ on groups $1, \ldots, i_*$, then we say that datapoint $\mathbf{x}_j$ matches $\mathbf{x}_*$. (Note that if $\mathbf{x}_*$ matches no groups, i.e. it differs on all dimensions from the grid, then all datapoints match $\mathbf{x}_*$.)

We can now state our first result:

Proposition 1 *For a separable covariance function with a grid design it is only those datapoints which match* $\mathbf{x}_*$ *that will have non-zero weights.*

Proof: The matrix K can be written as $K = K^1 \otimes \cdots \otimes K^D$, where $\otimes$ denotes the Kronecker product, and K^i is the $n^i \times n^i$ matrix corresponding to the ith factor in the covariance

function evaluated at the locations in $\mathcal{O}^i$. Similarly $\mathbf{k}_* = \mathbf{k}_*^1 \otimes \cdots \otimes \mathbf{k}_*^D$. Then using the identities $(A \otimes B)(C \otimes D) = AC \otimes BD$, $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ and $(A \otimes B)^T = A^T \otimes B^T$ repeatedly we obtain

$$\mathbf{k}_*^T K^{-1} = (\mathbf{k}_*^1)^T (K^1)^{-1} \otimes \cdots \otimes (\mathbf{k}_*^D)^T (K^D)^{-1}. \tag{5}$$

Now first consider a group where $\mathbf{x}_*^i$ matches $\mathcal{O}^i$. In this case $(\mathbf{k}_*^i)^T (K^i)^{-1}$ will give rise to a unit row vector $(0, \ldots, 0,\ 1,\ 0, \ldots, 0)$ with the 1 entry locating which of the entries in $\mathcal{O}^i$ it is that $\mathbf{x}_*^i$ matches. (This is easy to see as it is just one row from the identity $K^i(K^i)^{-1} = I_{n_i}$.) Thus all matched groups will give rise to such a unit vector in the Kronecker products of eq. 5. For the unmatched groups there is no special structure and those factors $(\mathbf{k}_*^j)^T (K^j)^{-1}$ for $j = i_* + 1, \ldots, D$ cannot be simplified. However, it is precisely the matching groups that give rise to our result as the Kronecker product of the unit vectors selects exactly those observations in $\mathbf{f}$ that match $\mathbf{x}_*$ according to our definition. Thus the number of points with non-zero weight will be $\prod_{j=i_*+1}^{D} n^j$.

Note that the predictive variance has the same dependence structure due to the presence of the same factor $\mathbf{k}_*^T K^{-1}$ in eq. 2.

Remark 1: This result is illustrated in Fig. 1(a) and Fig. 2. In Fig. 2(a) the test point matches $\mathcal{O}^1$ and $\mathcal{O}^2$, so it is only the datapoints on dimension x^3 that have non-zero weights (shown with larger open circles). In Fig. 2(b) when the test point is moved to a general position in the top layer, then it still matches against $\mathcal{O}^1$, but all datapoints in the top layer have non-zero weight. In Fig. 2(c) illustrates the fact that a "grid design" may not look completely grid-like. In this case dimensions x^2 and x^3 form the second (unmatched) group $\mathbf{x}^2$, and the locations of the datapoints in this group can be arbitrary. However, note that for every matched datapoint in the upper layer there is a corresponding datapoint in the lower (unmatched) layer.

Remark 2: One extreme version of this result is when all groups match so that $\mathbf{x}_*$ is equal to one of the $\mathbf{x}$'s in the training set; in this case (due to the noise-free observations) it is only the corresponding f that counts. At the other extreme, when $\mathbf{x}_*$ matches no groups then all datapoints (in general) have non-zero weight.

Proposition 2 *Proposition 1 still holds even if the non-matching points are only sparsely observed wrt the grid.*

Proof: A simple proof comes from the fact that even if the non-matching points on the grid were fully observed they would not have any effect on the prediction of $f(\mathbf{x}_*)$. There should be no difference between not observing a f-value at an unmatched grid point and observing a f-value there but then giving it zero weight.

Remark 3: This result is illustrated in Fig. 1(b), where non-matching datapoints have been thinned. However, Fig 1(c) illustrates that this thinning *cannot* be carried out on the *matched* points. Here the matched point at (x_2^1, x_2^2) has been removed and this means that all datapoints now have non-zero weight. The sparsely-observed grid design could be restored by deleting datapoints (x_1^1, x_2^2) and (x_3^1, x_2^2).

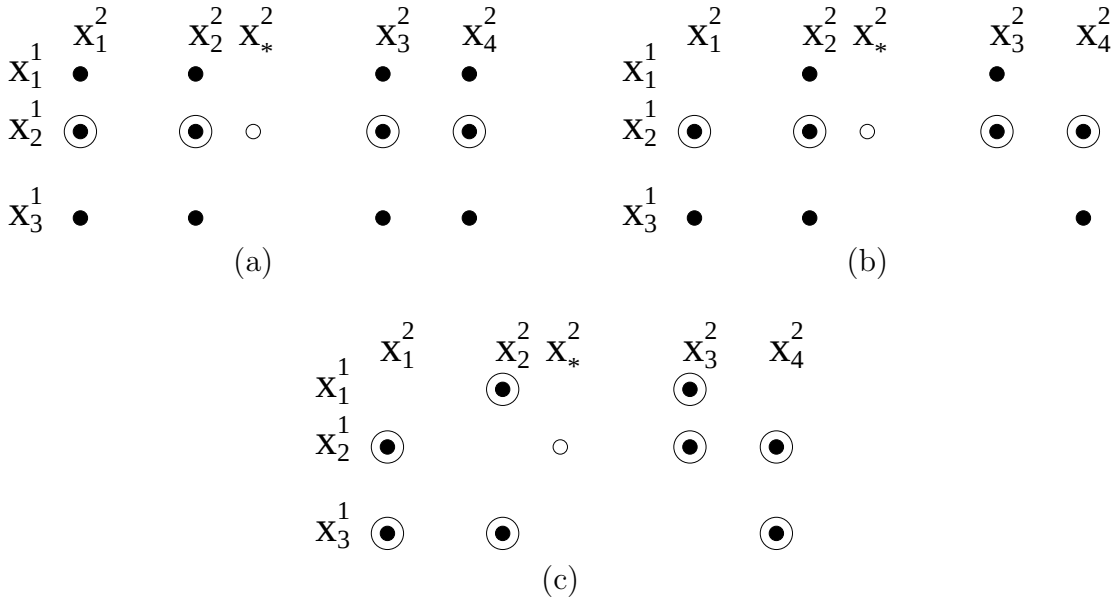

Figure 1: The observed data points are shown as black circles, and the test point is shown with an small open circle. Panel (a): The test point matches x_2^1 but not $\mathcal{O}^2$ so it is the datapoints indicated with a large open circles that have non-zero weights. Panel (b) illustrates the fact that unobserved datapoints in the grid do not change the result if they are unmatched. Panel (c) illustrates the crucial importance of a grid design for our results; if the matched point at (x_2^1, x_2^2) is removed from panel (b), then the conditional independence no longer holds and all datapoints have non-zero weight.

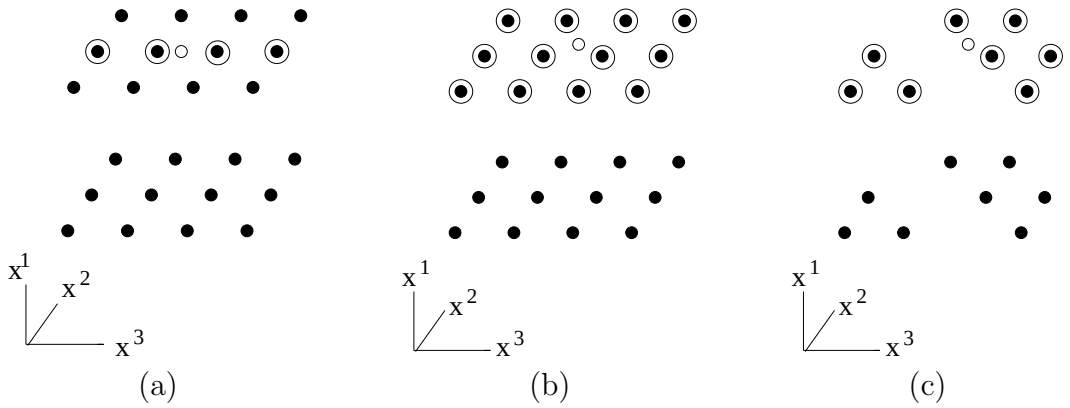

Figure 2: A 3-dimensional set up. The observed data points are shown as black circles, and the test point is shown with an small open circle. Panel (a): The test point matches $\mathcal{O}^1$ and $\mathcal{O}^2$, so it is only the datapoints on dimension x^3 that have non-zero weights; these are shown with larger open circles. Panel (b): When the test point is moved to a general position in the top layer, then it still matches against $\mathcal{O}^1$, but all datapoints in the top layer have non-zero weight. Panel (c) illustrates the fact that a "grid design" may not look completely grid-like. In this case dimensions x^2 and x^3 form the second (unmatched) group $\mathbf{x}^2$, and the locations of the datapoints in this group can be arbitrary. However, note that for every matched datapoint in the upper layer there is a corresponding datapoint in the lower (unmatched) layer.

Related Work

These results are known in the geostatistics literature for the case of $D = 2$ groups under the name of *autokrigeability* (Wackernagel, 1998, ch. 25). Here the set up is cokriging, where there is a m-variate stochastic process $\mathbf{Y}(\mathbf{x}) = (Y^1(\mathbf{x}), \ldots, Y^m(\mathbf{x}))^T$, where $\mathbf{x}$ is (say) a spatial variable and $Y^i(\mathbf{x})$ denotes the ith variable at location $\mathbf{x}$. (We might imagine measuring the concentration of minerals bearing copper, lead and zinc as the different variables.) The proportional correlation model defines the covariance $\mathrm{cov}(Y^i(\mathbf{x}), Y^j(\mathbf{x}')) = V_{ij}k(\mathbf{x}, \mathbf{x}')$, where V_{ij} is the i, jth element of a positive definite matrix V. This is equivalent to the separable covariance function defined above if we let $\mathbf{x}$ here be $\mathbf{x}^1$ and consider only a discrete set of m locations for $\mathbf{x}^2$. Under the proportional correlation model it is known that autokrigeability holds for isotopic data, i.e. if all variables are measured at the same sample locations then the prediction of $Y^i(\mathbf{x}_*)$ depends only on the observations of Y^i and not on the other Y^j's with $j \neq i$ (Wackernagel, 1998, ch. 25).

O'Hagan (1998) defines the symmetric Markov property as

$$k((\mathbf{x}^1, \mathbf{x}^2), (\tilde{\mathbf{x}}^1, \tilde{\mathbf{x}}^2))\, k((\mathbf{x}^1, \tilde{\mathbf{x}}^2), (\mathbf{x}^1, \tilde{\mathbf{x}}^2)) = k((\mathbf{x}^1, \mathbf{x}^2), (\mathbf{x}^1, \tilde{\mathbf{x}}^2))\, k((\mathbf{x}^1, \tilde{\mathbf{x}}^2), (\tilde{\mathbf{x}}^1, \tilde{\mathbf{x}}^2)) \quad (6)$$

for all $\mathbf{x}^1, \tilde{\mathbf{x}}^1 \in \mathcal{O}^1$ and for all $\mathbf{x}^2, \tilde{\mathbf{x}}^2 \in \mathcal{O}^2$. He then shows that the symmetric Markov property holds if and only if the covariance structure of the random variables $f(\mathbf{x}^1, \mathbf{x}^2)$ has a Kronecker product form. This property also states that $f(\mathbf{x}^1_i, \mathbf{x}^2_j)$ is independent of $f(\mathbf{x}^1_{i'}, \mathbf{x}^2_{j'})$ given *either* $f(\mathbf{x}^1_{i'}, \mathbf{x}^2_j)$ *or* $f(\mathbf{x}^1_i, \mathbf{x}^2_{j'})$, for $i' \neq i$ and $j' \neq j$. Taking Figure 1(a) as

an example, we have that $f(\mathbf{x}_2^1, \mathbf{x}_*^2)$ is independent of $f(\mathbf{x}_i^1, \mathbf{x}_j^2)$ given $f(\mathbf{x}_2^1, \mathbf{x}_j^2)$, for $i = 1, 3$, $j = 1, 2, 3, 4$; independence is made evident by the location of the zero-weights. Note that the symmetric Markov property is defined only wrt the case of two groups, i.e. $D = 2$.

Discussion

Our interest in this area came about from considering a multi-task learning problem, where $k^1(\mathbf{x}^1, \tilde{\mathbf{x}}^1)$ models similarity corresponding to inputs $\mathbf{x}^1$ and $\tilde{\mathbf{x}}^1$, while $\mathbf{x}^2$ is a task descriptor, so that $k^2(\mathbf{x}^2, \tilde{\mathbf{x}}^2)$ models the similarity between tasks. We initially rediscovered the autokrigeability result, which in that context means that given a grid design, for prediction on task i there is no benefit on making observations on task $j \neq i$, and then generalized it for $D > 2$.

Note that if the observations are noisy, these independence results will fail to hold and all datapoints will contribute to the prediction of f_*.

Acknowledgements

CW thanks Dan Cornford for pointing out the prior work on autokrigeability, Tony O'Hagan for comments on an earlier draft, and Manfred Opper for a helpful conversation.

References

O'Hagan, A. (1998). A Markov Property for Covariance Structures. Statistics Research Report 98-13, Nottingham University.

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

Ritter, K. (2000). *Average-Case Analysis of Numerical Problems*. Springer Verlag.

Wackernagel, H. (1998). *Multivariate Geostatistics*. Springer Verlag, second edition.