SUPPLEMENT TO "EFFICIENT AND INTERPRETABLE ADDITIVE GAUSSIAN PROCESS REGRESSION AND APPLICATION TO ANALYSIS OF HOURLY-RECORDED NO_2 CONCENTRATIONS IN LONDON"

BY SAHOKO ISHIDA[1,a] AND WICHER BERGSMA[1,b]

[1] *Department of Statistics, London School of Economics and Political Science,* [a]*s.ishida@lse.ac.uk;* [b]*w.p.bergsma@lse.ac.uk*

1. Kernels and Kernel properties.

1.1. *Different kernels.*

Common kernels. We introduced squared exponential kernel and fractional Brownian Motion kernel in the paper. The examples below are other kernels that are commonly used in the machine learning and spatial statistics literature. Note that all kernels have a scale parameter $\alpha > 0$ and some also share another parameter $\rho > 0$, called length-scale. Let $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$ and $t, t' \in \mathbb{R}$ and $c > 0$ be a constant.

1. Matérn class kernel (Matérn (1960), Stein (1999)):

$$k_{mat}(\mathbf{x}, \mathbf{x}') = \alpha^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||}{\rho} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||}{\rho} \right),$$

   where K_ν is a modified Bessel function. The parameter $\nu > 0$ determines the roughness of the corresponding process.
2. Periodic kernel:

$$k_{pr}(t, t') = \alpha^2 \exp\left(- \frac{2\sin^2(\frac{\pi|t-t'|}{p})}{\rho^2} \right),$$

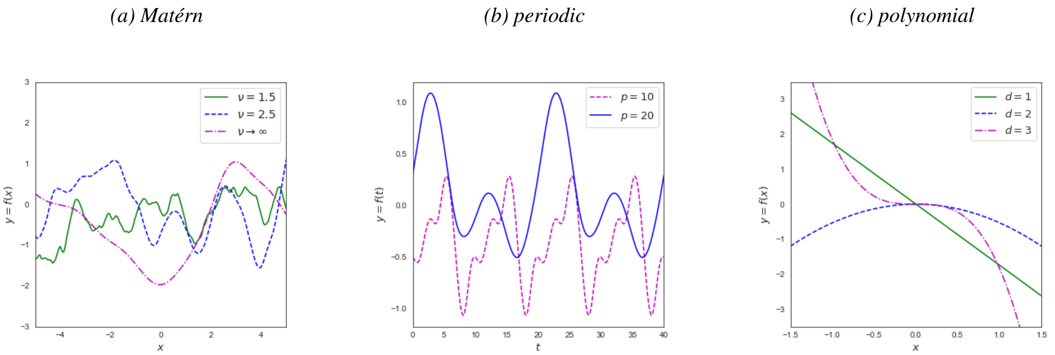
   where $p > 0$, period parameter, can be treated as known or unknown. Corresponding GP is a periodic function of period p.
3. Polynomial kernel

$$k_{pol}(\mathbf{x}, \mathbf{x}') = \alpha^2 \left(\mathbf{x}^\top \mathbf{x}' + c \right)^d,$$

   where $m \in \mathbb{N}$. With $d = 1$ we have linear kernel.
4. Constant kernel

$$k_{const}(\mathbf{x}, \mathbf{x}') = c \tag{1.1}$$

With Matérn class kernels, the smoothness of the process can be controlled by parameter ν. With $\nu = 1.5$ process is rough (see figure 1) compared with $\nu = 2.5$. It is worth noting that for $\nu \to \infty$, it equals S.E. kernel. Periodic kernel is useful when handling the continuous-time process that has a regular cycle, e.g. daily, weekly or annually. It can be derived from S.E. kernel; we have $k_{se}(\mathbf{u}, \mathbf{u}') = k_{pr}(t, t')$ where $\mathbf{u} = (\sin(\frac{p}{2\pi}t), \cos(\frac{p}{2\pi}t))^\top$. In fact any kernel k can be made periodic with this formulation. The constant kernel is usually used in combination with other kernels.


(a) Matérn

(b) periodic

(c) polynomial

Fig 1: Sample paths from zero mean Gaussian process with different kernels. For all panels, the scale parameter α is set to be 1, and the length-scale parameter $\rho = 1$ for (a) and (b). For polynomial kernel, $c = 0$. For the additional parameters see the legend of each panel.

Stationary and non-stationary kernel. SE kernel, Matérn kernel, and periodic kernel constructed from SE kernel are in the class of stationary kernel, more specifically isotropic kernel. A stationary kernel is a function of a lag vector $\tau = \mathbf{x} - \mathbf{x}'$ of two inputs. When the value of the function depends only on the norm of the two inputs $r = ||\tau||$, the kernel is said to be isotropic and the corresponding process is invariant under shift in time or space. While the assumption of isotropy or stationarity gives a nice interpretation of correlation structure, we need a class of non-stationary kernels in the case where this assumption does not hold. A few simple examples of non-stationary kernels include linear kernel and polynomial kernel. Using these kernels in Gaussian process regression corresponds with Bayesian linear or polynomial regression. Another useful non-stationary kernel is the fractional Brownian Motion kernel and kernels that are constructed from this kernel, such as its centred version.

1.2. *Kernel sums and products.* Given valid kernels k_1 and k_2 on $\mathcal{X}$, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ constructed as their sum or product

$$k(\mathbf{x}, \mathbf{x}) = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$$
$$k(\mathbf{x}, \mathbf{x}) = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$$

constitutes a positive definite kernel. The kernel k_1 or k_2 can be a (positive) constant kernel (1.1). Hence, adding a positive constant or multiplying by a positive constant gives a positive definite kernel. It is also important to note that it is not necessary that k_1 and k_2 are defined on the same set. For example, given $k_1 : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ and $k_2 : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$, then $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ given by

$$k((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_1', \mathbf{x}_2')) = 1 + k_1(\mathbf{x}_1, \mathbf{x}_1') + k_2(\mathbf{x}_2, \mathbf{x}_2') + k_1(\mathbf{x}_1, \mathbf{x}_1')k_2(\mathbf{x}_2, \mathbf{x}_2'),$$

where $\mathbf{x}_l, \mathbf{x}_l' \in \mathcal{X}_l$ for $l = 1, 2$, is a positive definite kernel.

2. Centring of kernels.

2.1. *Reproducing kernel Hilbert space.* Recall that Hilbert space is a complete inner product space equipped with a positive definite inner product. Let $\mathcal{H}$ be a Hilbert space of functions over a set $\mathcal{X}$ with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The Hilbert space $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) if and only if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying

1. $k(x,\cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$
2. $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$

The function k is called reproducing kernel. Note that using the two properties, we have that $k(x,x') = \langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}}$, hence k is positive definite. It can be shown by the Moore–Aronszajn theorem (Aronszajn (1950)) that a kernel defines a unique RKHS and vice versa. We write the norm of a function in f in $\mathcal{H}$ as $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle}$.

2.2. *Centring of kernel and functions in RKHS.* Let P be distribution over a non-empty set $\mathcal{X}$ and $X, X' \in \mathcal{X}$ are independent and follow P. We consider a kernel k on $\mathcal{X}$ and let $\mathcal{H}_k$ denote the RKHS induced by k. We can center this kernel by,

$$k_{cent}(x,x') = \langle k(x,\cdot) - \mu_P, k(x',\cdot) - \mu_P\rangle_{\mathcal{H}_k} \tag{2.1}$$

where μ_P is the kernel mean given by

$$\mu_P := \mathop{\mathbb{E}}_{X\sim P}[k(X,\cdot)] = \int_{\mathcal{X}} k(x,\cdot)dP(x).$$

Note that the expectation of any function $f \in \mathcal{H}_k$ can be computed as an inner product with μ_P:

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim P}[f(X)] &= \int_{\mathcal{X}} f(x)dP(x)\\ &= \int_{\mathcal{X}} \langle k(x,\cdot), f\rangle_{\mathcal{H}_k} dP(x)\\ &= \langle \int_{\mathcal{X}} k(x,\cdot)dP(x), f\rangle_{\mathcal{H}_k} = \langle \mu_P, f\rangle_{\mathcal{H}_k}.\end{aligned}$$

The centred kernel (2.1) is positive definite by construction. We can see that this corresponds with (2.9) in the main paper by

$$\begin{aligned}\langle k(x,\cdot) - \mu_P, k(x',\cdot) - \mu_P\rangle_{\mathcal{H}_k} &= \langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}_k} - \langle \mu_P, k(x',\cdot)\rangle_{\mathcal{H}_k} - \langle k(x,\cdot), \mu_P\rangle_{\mathcal{H}_k} + \langle \mu_P, \mu_P\rangle_{\mathcal{H}_k}\\ &= k(x,x') - \mathop{\mathbb{E}}_{X\sim P}[k(x',X)] - \mathop{\mathbb{E}}_{X'\sim P}[k(X',x)] + \mathop{\mathbb{E}}_{X,X'\sim P}[k(X,X')].\end{aligned}$$

Note that

$$\begin{aligned}\mathop{\mathbb{E}}_{X,X'\sim P}[k(X,X')] &= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')dP(x)dP(x')\\ &= \int_{\mathcal{X}}\int_{\mathcal{X}} \langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}_k} dP(x)dP(x')\\ &= \langle \int_{\mathcal{X}} k(x,\cdot)dP(x), \int_{\mathcal{X}} k(x',\cdot)dP(x')\rangle_{\mathcal{H}_k} = \langle \mu_P, \mu_P\rangle_{\mathcal{H}_k}.\end{aligned}$$

Given a sample $x_1, \dots x_n$ drawn from P, the kernel mean μ_P can be estimated empirically, by

$$\hat{\mu}_P = \frac{1}{n}\sum_{i=1}^{n} k(x_i,\cdot).$$

By replacing μ_P with $\hat{\mu_P}$, we get (2.10) of the main paper.

3. Kronecker products.

3.1. *Kronecker product and its properties.*

Kronecker product. Consider two matrices $\mathbf{A} = \{a_{i,j}\}_{1\leq i\leq n, 1\leq j\leq m}$ and $\mathbf{B} = \{b_{i,j}\}_{1\leq i\leq p, 1\leq j\leq q}$. The Kronecker product of the two matrices, $\mathbf{A} \otimes \mathbf{B}$, is the matrix of size $np \times mq$ given by

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} \dots & a_{1,m}\mathbf{B} \\ \vdots \quad \ddots & \vdots \\ a_{n,1}\mathbf{B} \dots & a_{n,m}\mathbf{B}. \end{bmatrix}$$

More generally, we denote the Kronecker product of $d \geq 2$ matrices, $\mathbf{A}_l$ where $l = 1, \dots d$ by

$$\mathbf{A} := \bigotimes_{l=1}^{d} \mathbf{A}_l$$

If each matrix $\mathbf{A}_l$ is size $n_l \times m_l$, the resulting Kronecker product matrix $\mathbf{A}$ has size $\prod_{l=1}^{d} n_l \times \prod_{l=1}^{D} m_l$.

Kronecker product properties. We list some of the properties of Kronecker product that we use in this paper. In addition to $\mathbf{A}_l$ defined above, let us assume we have, for $l = 1, \dots, d$, $\mathbf{B}_l$ of size $p_l \times q_l$, $\mathbf{B}'_l$ of size $p_l \times q_l$, $\mathbf{C}_l$ of size $h_l \times k_l$ and $\mathbf{D}_l$ of size $m_l \times p_l$. The size of matrices is given so that the operations $\mathbf{B}_l + \mathbf{B}'_l$ and $\mathbf{A}_l\mathbf{D}_l\mathbf{B}_l$ are allowed.

1. Bilinearity:
$$\mathbf{A}_l \otimes (\mathbf{B}_l + \mathbf{B}'_l) = \mathbf{A}_l \otimes \mathbf{B}_l + \mathbf{A}_l \otimes \mathbf{B}'_l$$
2. Associativity:
$$\mathbf{A}_l \otimes (\mathbf{B}_l \otimes \mathbf{C}_l) = (\mathbf{A}_l \otimes \mathbf{B}_l) \otimes \mathbf{C}_l$$
$$\alpha(\mathbf{A}_l \otimes \mathbf{B}_l) = (\alpha\mathbf{A}_l) \otimes \mathbf{B}_l = \mathbf{A}_l \otimes (\alpha\mathbf{B}_l)$$
where α is a scalar.
3. Transpose:
$$\left(\bigotimes_{l=1}^{d} \mathbf{A}_l\right)^{\top} = \bigotimes_{d=l}^{d} \mathbf{A}_l^{\top}$$
4. Inverse:
$$\left(\bigotimes_{l=1}^{d} \mathbf{A}_l\right)^{-1} = \bigotimes_{d=l}^{d} \mathbf{A}_l^{-1}$$
5. The mixed product properties:
$$\bigotimes_{l=1}^{d} (\mathbf{A}_l\mathbf{D}_l) = \left(\bigotimes_{l=1}^{d} \mathbf{A}_l\right)\left(\bigotimes_{l=1}^{d} \mathbf{D}_l\right)$$
This can be generalised further. For example,
$$\bigotimes_{l=1}^{d} (\mathbf{A}_l\mathbf{D}_l\mathbf{B}_l) = \left(\bigotimes_{l=1}^{d} \mathbf{A}_l\right)\left(\bigotimes_{l=1}^{d} \mathbf{D}_l\right)\left(\bigotimes_{l=1}^{d} \mathbf{B}_l\right)$$
6. Matrix vector product
$$(\mathbf{A}_l \otimes \mathbf{B}_l)\,\mathbf{v} = \text{vec}\left(\mathbf{B}_l\mathbf{V}\mathbf{A}_l^{\top}\right)$$
where $\mathbf{V} = \text{vec}^{-1}(\mathbf{v})$ is the inverse of the vectorization operator and $\mathbf{v}$ is a vector of length $m_l q_l$.

3.2. *Eigendecomposition of a Gram matrix with Kronecker product.* Assume a tensor product kernel

$$k(\mathbf{x},\mathbf{x}') = \prod_{l=1}^{d} k_d(\mathbf{x}_l,\mathbf{x}'_l)$$

over a multidimensional grid $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ where $\mathbf{x}_l \in \mathcal{X}_l$ and each k_l is defined on $\mathcal{X}_l$. Let n_l denote the dimension of each grid. Then the associated Gram matrix can be written as

$$\mathbf{K} = \bigotimes_{l=1}^{d} \mathbf{K}_l$$

where $\mathbf{K}_l$ is a $n_l \times n_l$ gram matrix for l input dimension, with i,j-th element given by $k_l(\mathbf{x}_{(l),i},\mathbf{x}_{(l),j})$. Let $\mathbf{K}_l = \mathbf{Q}_l \boldsymbol{\Lambda}_l \mathbf{Q}_l^\top$ be eigendecomposition of each matrix. Then the eigendecomposition of the matrix K is the following:

$$\begin{aligned} \mathbf{K} &= \bigotimes_{l=1}^{d} \left(\mathbf{Q}_l \boldsymbol{\Lambda}_l \mathbf{Q}_l^\top\right) \\ &= \bigotimes_{l=1}^{d} \mathbf{Q}_l \bigotimes_{l=1}^{d} \boldsymbol{\Lambda}_l \bigotimes_{l=1}^{d} \mathbf{Q}_l^\top. \end{aligned} \tag{3.1}$$

Note that $\mathbf{Q} \equiv \bigotimes_{l=1}^{d} \mathbf{Q}^l$ is orthonormal, i.e., $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_n$. We can confirm this by

$$\begin{aligned} \mathbf{Q}\mathbf{Q}^\top &= \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l\right)\left(\bigotimes_{l=1}^{d} \mathbf{Q}_l\right)^\top = \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l\right)\left(\bigotimes_{l=1}^{d} \mathbf{Q}_l^\top\right) \\ &= \bigotimes_{l=1}^{d} \mathbf{Q}_l\mathbf{Q}_l^\top = \bigotimes_{l=1}^{d} \mathbf{I}_{n_l} = \mathbf{I}_n. \end{aligned}$$

3.2.1. *An example with fractional Brownian motion kernel.* This decomposition leads to a particularly efficient algorithm when using fBM kernel or squared fBM kernel with a known Hurst coefficient γ_l. Let each k_l be a fBM$_{\gamma_l}$ kernel. This means that we have only one hyper-parameter (scale parameter) to estimate for each dimension l. We denote the corresponding gram matrix by $\mathbf{K}_l = \alpha_l \mathbf{K}'_l$ where $\mathbf{K}'_l$ is un-scaled gram matrix. Let $\mathbf{K}'_l = \mathbf{Q}_l \boldsymbol{\Lambda}_l \mathbf{Q}_l^\top$ be the eigendecomposition of the un-scaled matrix. Then eigendecomposition of $\mathbf{K}_l$ is

$$\mathbf{K}_l = \alpha_l \mathbf{K}'_l = \mathbf{Q}_l \left(\alpha_l \boldsymbol{\Lambda}_l\right) \mathbf{Q}_l^\top.$$

Using (3.1), we can write

$$\mathbf{K} = \bigotimes_{l=1}^{d} \mathbf{K}'_l = \bigotimes_{l=1}^{d} \mathbf{Q}_l \bigotimes_{l=1}^{d} \left(\alpha_l \boldsymbol{\Lambda}_l\right) \bigotimes_{l=1}^{d} \mathbf{Q}_l^\top.$$

This means that when estimating the hyper-parameters by maximising the marginal likelihood or by MCMC, we do not have to apply eigendecomposition at each iteration. By simply multiplying each eigenvalue by the scale parameters, the inverse and the determinant can be updated.

3.3. *Row-wise Kronecker product.* Consider two matrices $\mathbf{A} = \{a_{i,j}\}_{1\leq i\leq n, 1\leq j\leq m}$ and $\mathbf{B} = \{b_{i,j}\}_{1\leq i\leq n, 1\leq j\leq q}$. Let $\mathbf{A}_i$ and $\mathbf{B}_i$ be the $i-$th row of the matrices $\mathbf{A}$ and $\mathbf{B}$ respectively. The row-wise Kronecker product of the two matrices, $\mathbf{A} \bullet \mathbf{B}$, is the matrix of size $n \times mq$ given by

$$\mathbf{A} \bullet \mathbf{B} = \begin{bmatrix} \mathbf{A}_1 \otimes \mathbf{B}_1 \\ \mathbf{A}_2 \otimes \mathbf{B}_2 \\ \vdots \\ \mathbf{A}_n \otimes \mathbf{B}_n \end{bmatrix}$$

where $\otimes$ is Kronecker product. Row-wise Kroncker product may also be called face-splitting product. Let $\mathbf{v}$ be a vector of length n. Then we have

$$\mathbf{A} \bullet \mathbf{v} = \mathbf{v} \bullet \mathbf{A} = \mathbf{V}_d \mathbf{A}$$

where $\mathbf{V}_d = \text{diag}(\mathbf{v})$, a diagonal matrix with its diagonal elements given by $\mathbf{v}$.

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society* **68** 337–404.

MATÉRN, B. (1960). *Spatial variation*. Allmänna Förlaget.

STEIN, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, New York.