Notes: Empirical Mode Decomposition & Gaussian Processes

*

Version: June 20, 2019

Abstract. Extension 1, Estimation: Treat each IMF as a separate Gaussian process and then represent the signal using multi-kernel representation of the Gaussian Process.
Extension 2, Forecasting: GP representation does not ensures itself that the predicted function from a given Gaussian process is IMF , that is, it satisfies (I1)-(I2). Therefore, we explore the formulation of IMFs as an analogue of Brownian Bridge.

Contents

1. Gaussian Processes and EMD: IMFs as Gaussian Processes with non-stationary kernels

We treat each IMF as a separate Gaussian process and then represent the signal using multi-kernel representation of the Gaussian Process.

1.1. The Stochastic Representation by Gaussian Processes

Let $s(t)$ for $t \in [0, \infty]$ be a continuous true signal which is observed on discrete grid of points in the interval $[0, T]$, $t = (t_1 < \cdots < t_N) = \{t_i\}_{i=1:N}$, where the subscripts represent the sampling index times. The observed values of the true signal $s(t)$ might be exact or be perturbed. The noisy observation are not uncommon situation

in practice. The perturbation of the true signal can be either deterministic (ie an chaotic system, not stable) or stochastic. If the realisations of the signal $s(t)$ are corrupted with some stochastic error term, the process which we observe is represented as follows

$$y(t) = s(t) + \epsilon, \text{ for } \epsilon \in \mathcal{N}(0, \sigma^2). \tag{1}$$

Therefore, our observation set consists of pairs $\{t_n, y_n\}$ where $y_n = y(t_n)$ for $t_n \in [0, T]$.

We would like to find an EMD decoposition of the signal $s(t)$. For EMD to exists, the input signal needs to be approximated by a continuous representation; therefore, the discrete signal $s(t)$ is converted back into a continuous analog signal. The background on the EMD decomposition is given in Subsection **??**. If we assume that the observations of the signal are exact, the approximation of the signal $s(t)$ might be carried out by a spline interpolation as described in Section **??** in Equation . In the presence of noisy environment, the approximation $S(t)$ might be obtained by a filtering techniques which would account for the perturbation of the true values. We describe in detail this case in Section 2.

Either way, we specify a continuous approximation $S(t)$ to the discrete realisation of the true signal $s(t)$ and define the representation of $S(t)$ given by EMD into M intrinsic mode functions (IMFs) as follows

$$S(t) = \sum_{m=1}^{M} \gamma_m(t) + r(t) = \sum_{m=1}^{M} \text{Re}\Big\{ A_m(t) e^{i\theta_m(t)} \Big\} + r(t). \tag{2}$$

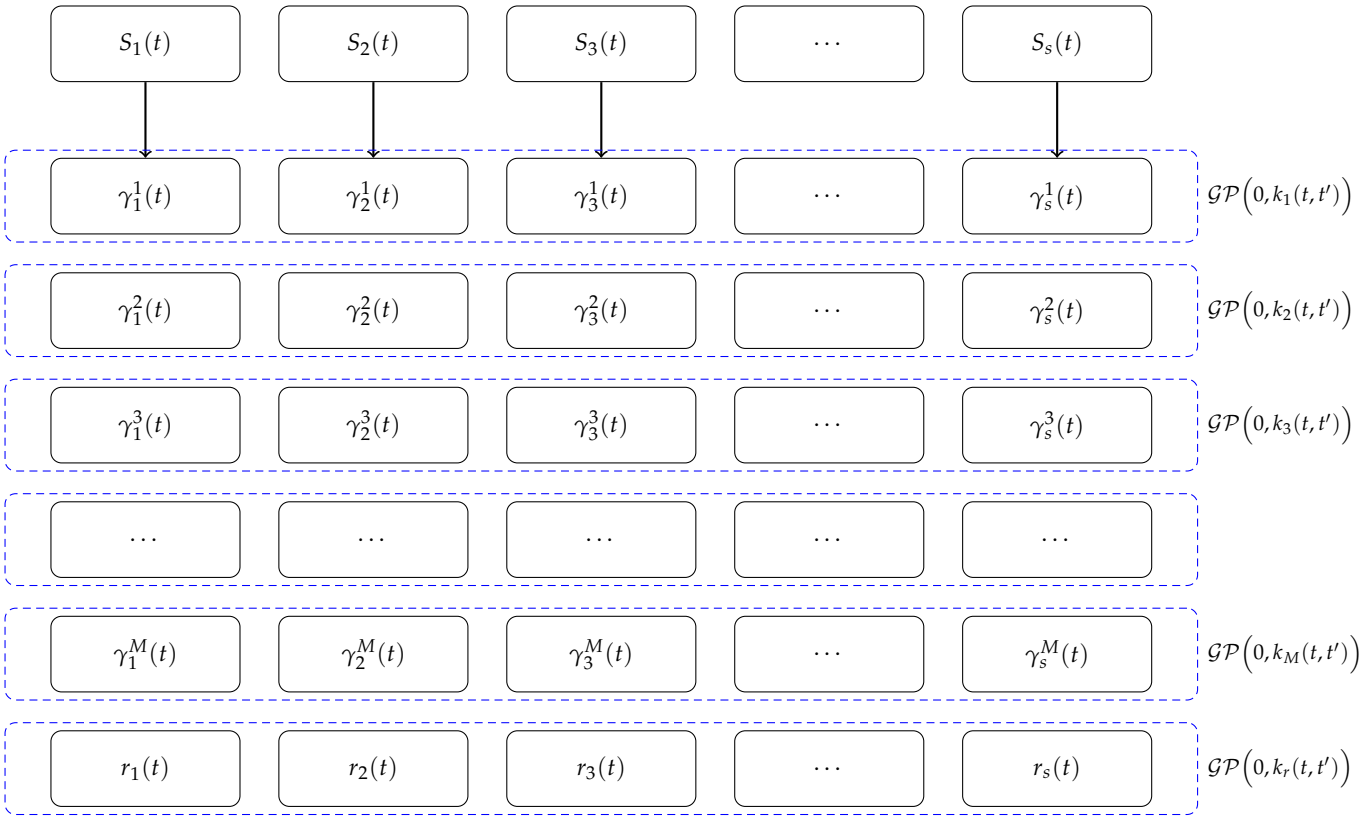
where $r(t)$ represents a tendency which does not have much of oscillation and therefore characterize the low frequency tend of $S(t)$.

1.1.1. Continuous signal $S(t)$ as a Gaussian Process

Our goal is to obtain a stochastic representation of the continuous signal $S(t)$ by a Gaussian Process. We postulate that each IMFs function, $\gamma_m(t)$, is a Gaussian process

$$\gamma_m(t) \sim \mathcal{GP}\Big(0, k_m(t, t')\Big), \tag{3}$$

where $k_m(t, t')$ is a positive definite covariance kernel which is parametrized by a set of parameters Ψ_m. discussion with dorota



Let us assume that we sample $S(t)$, and functions $\gamma_m(t)$ for $m = 1, \ldots, M$ at the N time points $t_1 < \ldots < t_N$. We denote by $\mathbf{t}$ the vector of points t_n for $n = 1, \ldots, N$.

Therefore, given the observations $\gamma_m(\mathbf{t})'' = [\gamma_m(t_1), \ldots, \gamma_m(t_N)]$, we would like to predict the values of $\gamma_m(t)$ at the argument s that is $\gamma_m(s)$, given the collected information in the observation set. Since $\gamma(t)$ is a Gaussian Process, the random variable $\gamma_m(s)|\gamma_m(\mathbf{t}), \mathbf{t}$ is a Gaussian Process with the conditional mean

$$\mu_m(s) := \mathbb{E}_{\gamma_m(t)|\gamma_m(\mathbf{t}),\mathbf{t}}\left[\gamma_m(s)\right] = \mathbf{k}_m(s,\mathbf{t})\mathbf{K}_m(\mathbf{t},\mathbf{t})^{-1}\gamma_m(\mathbf{t})$$

and the conditional covariance matrix given by

$$\tilde{k}_m(s,s') := \mathbb{E}_{\gamma_m(t)|\gamma_m(\mathbf{t}),\mathbf{t}}\left[(\gamma_m(s) - \mu_m(s))(\gamma_m(s') - \mu_m(s'))\right] = k_m(s,s') - \mathbf{k}_m(s,\mathbf{t})\mathbf{K}_m(\mathbf{t},\mathbf{t})^{-1}\mathbf{k}_m(\mathbf{t},s')^T$$

where

$$\mathbf{K}_m(\mathbf{t},\mathbf{t}) := \begin{bmatrix} k_m(t_1,t_1) & k_m(t_1,t_2) & \cdots & k_m(t_1,t_N) \\ k_m(t_2,t_1) & k_m(t_2,t_2) & \cdots & k_m(t_2,t_N) \\ \vdots & \vdots & \ddots & \vdots \\ k_m(t_N^{(i)},t_1) & k_m(t_N,t_2) & \cdots & k_m(t_N,t_N) \end{bmatrix}_{N\times N}$$

and

$$\mathbf{k}_m(s,\mathbf{t}) := \begin{bmatrix} k_m(s,t_1) & k_m(s,t_2) & \cdots & k_m(s,t_N) \end{bmatrix}_{1\times N}.$$

TODO: If we would like to regularize the Gram matrix of $\gamma_m(t)$, the mean function and kernel of the conditional distribution would be the following

$$\mu_m(s) := \mathbb{E}_{\gamma_m(t)|\gamma_m(\mathbf{t}),\mathbf{t}}\left[\gamma_m(s)\right] = \mathbf{k}_m(s,\mathbf{t})\left(\mathbf{K}_m(\mathbf{t},\mathbf{t}) + \sigma_k^2\right)^{-1}\gamma_m(\mathbf{t})$$

and the conditional covariance matrix given by

$$\tilde{k}_m(s,s') := \mathbb{E}_{\gamma_m(t)|\gamma_m(\mathbf{t}),\mathbf{t}}\Big[(\gamma_m(s)-\mu_m(s))(\gamma_m(s')-\mu_m(s'))\Big] = k_m(s,s') - \mathbf{k}_m(s,\mathbf{t})\Big(\mathbf{K}_m(\mathbf{t},\mathbf{t})+\sigma_k^2\Big)^{-1}\mathbf{k}_m(\mathbf{t},s')^T$$

what is equivalent to the accounting for the artificial noise component in the model in Equation (3).

Multikernel Representation of $S(t)$

The tendency component $r(t)$ can be modelled as a Gaussian Process itself or one can assume that $S(t)$ is a Gaussian Process conditioned on $r(t)$, that is

$$S(t)|r(t) \sim \mathcal{GP}\Big(r(t), k(t,t')\Big). \tag{4}$$

where $k(t,t')$ is a function of the kernels $k_m(t,t')$ for $m \in \{1,\ldots,M\}$ These two approaches provide an unconditional and conditional stochastic representation of $S(t)$, respectively, and determine two different estimators of the out-of-sample forecast for $S(t)$. The later is a more convenient assumption to preserve the monotonicity of the $r(t)$ which is a desired property of a residual function in the decomposition in Equation (2). To ensure the function $r(t)$ to have only single convexity change, $r(t)$ might be extrapolated by a power law which stays monotonic (ie a polynomial up to the second order). Then, the out-of-sample forecast of $S(t)$ would be conditioned on the extrapolation of $r(t)$. In order to preserve the monotonicity property of the tendency function $r(t)$ in the out-of-sample prediction, the extrapolation from a low order spline representation of $r(t)$, which is deterministic, is excepted to behaves better than the forecast from a Gaussian Process since the later would most plausibly wiggle around a trend and, consequently, would loose the monotonicity of $r(t)$. In the following work we would like to guarantee the out-of-sample monotonicity of $r(t)$ obtained by construction in the in-ample set, and therefore, we chose to work with the conditional representation of $x(t)$ given in Equation (4). TODO: derive the properties of these two estimators..

Given the Gaussian Process model of the $\gamma_m(t)$ in Equation (3), the distribution of $S(t)$ can be formulated as a uniform mixture of Gaussian Processes with different kernels. If we assume that the processes $\gamma_m(t)$ are independent, then, the stochastic representation of $S(t)$ from Equation (4) can be formulated as follows

$$S(t)|r(t) \sim GP\Big(r(t); \sum_{m=1}^{M} k_m(t,t')\Big) \tag{5}$$

If we denote by $k(t,t') := \sum_{m=1}^{M} k_m(t,t')$, then predictive distribution of $S(t)$ is given by

$$\mu(s) := \mathbb{E}_{S(t)|r(t),\mathbf{s},\mathbf{t}}[S(\mathbf{s})] = r(\mathbf{s}) + \sum_{m=1}^{M} \mu_m(s) \tag{6}$$

and the covariance matrix given by

$$\tilde{k}(s,s') := \mathbb{E}_{S(t)|S(\mathbf{t}),\mathbf{t}}\Big[(S_m(s)-\mu(s))(S(s')-\mu(s'))\Big] = \sum_{m=1}^{M} \tilde{k}(s,s') \tag{7}$$

If the processes of $\gamma_m(t)$ are not independent, the Gram matrix of the model for $S(t)$ contain additional elements which provide the correlation structure between different IMFs

$$s(t)|r(t) \sim \mathcal{GP}\Big(r(t); \sum_{m=1}^{M} k_m(t,t') + 2\sum_{m_1,m_2=1,m_1<m_2}^{M} k_{m1,m2}(t,t')\Big) \tag{8}$$

where $k_{m1,m2}(t,t')$ defines the dependence structure between $\gamma_{m_1}(t)$ and $\gamma_{m_2}(t)$.

2. Multiple Trials Setting: How to Construct $S(t)$ in Noisy Environment

We consider the following experiment setup. Let J represents the number of trials in our experiment which characterize the set of sample that we collected, ie. realisations of $y(t)$ on the discrete subsets of the interval $\in [0, T]$, which can be specified by **random or deterministic sub-sampling**. We assume that each trial has a set of N^i samples and N^i varies over trials for $i = 1, \ldots, J$. Let $\mathbf{y}^i$ and $\mathbf{t}^i$ denote N^i-dimensional vectors which represent the N^i observed values in the ith trial and the N_i corresponding time points being a subsample of $[0, T]$, respectively. Given that, $\mathbf{y}^i := y(\mathbf{t}^i) = [y(t^i_1), \ldots, y(t^i_{N^i})]$. **We remark that it is not ensured that for the same time point $t_0 \in [0, T]$, that the value of $y(t_0)$ in trial i_1 and a value of $y(t_0)$ in trial i_2 are equal since the definition of $y(t)$ in Equation (1) includes the error term component.**

The sets of the time points for each trial, $\mathbf{t}^i$, can be specified deterministic or be a realisations of the random variable. Regardless of the assumption on the sampling mechanism, the time points collected in the set $\mathbf{t}^i$ can be missing. Therefore, we may distinguish the complete and incomplete cases for the sampling times $\mathbf{t}^i$ and the deterministic or random sampling framework. In the following section we will consider the simplest case, when the elements of $\mathbf{t}^i$ are obtained deterministically and are not missing. Set up notation and cases for the frameworks which we will consider later for subsampling

2.1. Review of Denoising Approaches

1. median filter
2. spline filter
3. smoothing splines
4. L1 trend filter
5. Exponential moving average (EMA) / a weighted moving average (WMA)

3. Gaussian Process Representation Given Splines Formulation of $S(t)$.

As remarked in Subsection **??**, the EMD procedure required that the underlying signal has a continuous formulation. For the EMD to exist, the underlying signal $S(t)$ needs to be approximated. Such approximation is covered in this work by a natural cubic spline representation of $S(t)$.

The natural cubic spline is characterised over time intervals, where the local cubic is expressed in a local time window. The time intervals are structured by points known as knot points; in this paper, such knot points are placed at the sampling times. This gives us a representation of the original signal, identified by $S(t)$ as follows:

$$S(t) = \sum_{i=1}^{N-1} \left(a_i t^3 + b_i t^2 + c_i t + d_i \right) \mathbb{1}\left[t \in [t_{i-1}, t_i] \right], \tag{9}$$

where the spline coefficients will be estimated from the original sample path, such that the representation exactly matches the sample values at these time points, $a_i = S(t_i) = s(t_i)$. We need to construct an analog continuous signal from the discrete one, since our basis decomposition requires a continuous smooth signal for the basis extraction. The number of total convexity changes (oscillations) of the analog signal $S(t)$ corresponds to $K \in \mathbb{N}$ within the time domain t, over which the signal was observed. Note that $S(t)$ is decomposed according to direct extraction of the energy associated with various intrinsic time scales. This is the property that makes it suitable to non-linear and non-stationary processes. One may now define the EMD defined in Section 4 of the signal $S(t)$ as in Equation (2).

Given the spline representation of $S(t)$, each IMF $\gamma_{m,j}^{(i)}$ can be obtained as a natural cubic spline, defined as $\gamma_{m,j}^{(i)}(t)$ with the following formulation:

$$\gamma_{m,j}^{(i)}(t) = \begin{cases} s_1(t) = a_1 t^3 + b_1 t^2 + c_1 t + d_1 & \text{for} \quad t \in (t_1^i, t_2^i) \\ s_2(t) = a_2 t^3 + b_2 t^2 + c_2 t + d_2 & \text{for} \quad t \in (t_2^i, t_3^i) \\ \dots \quad \dots \quad \dots \quad \dots \quad \dots & \\ s_{N_i}(t) = a_{N_i} t^3 + b_{N_i} t^2 + c_{N_i} t + d_{N_i} & \text{for} \quad t \in (t_{N_i-1}^i, t_{N_i}^i) \end{cases}$$

A shorter version of the above system of equations can be given by:

$$\gamma_{m,j}^{(i)}(t) = \sum_{j=1}^{N_i} \left(a_j t^3 + b_j t^2 + c_j t + d_j \right) \mathbb{1} \left(t \in (t_{j-1}^i, t_j^i) \right) = \sum_{j=1}^{N_i} s_j(t) \mathbb{1} \left(t \in (t_{j-1}^i, t_j^i) \right) \tag{10}$$

where $\mathbb{1}$ represents the indicator function.
Note that, in the above representation, $\gamma_m(t)$ is not explicitly expressed in a functional form, as opposed to classical stationary methods where a cosine basis or a wavelet basis function is specified. Here, the basis can take any functional form so long as it satisfies the decomposition relationship and the properties stated on the IMF. A natural way to proceed to represent an IMF is to utilise a smooth, flexible characterisation that can adapt to local non-stationary time structures; we have, again, selected the cubic spline in this work to represent $\gamma_m(t)$.
Given a mathematical representation for the IMFs, we must now proceed to outline the process applied to extract recursively the IMF spline representations. This procedure is known as *sifting*. The first step consists of computing extrema of $S(t)$. By taking the first derivative $S'(t)$ and set it equal to zero, maxima and minima within each interval are calculated, producing the sequence of time points at which maxima and minima of $S(t)$ are located being given by:

$$\left\{ t_j^* \right\} = \left\{ \left[-\frac{b_j}{3a_j} \pm \sqrt{\frac{{b_j}^2 - 3a_j c_j}{9{a_j}^2}} \right] : t \in (t_1, t_N) \quad \& \quad \frac{dS(t)}{dt} = 0 \quad j = 1, \dots, M \right\} \tag{11}$$

where $\{t_j^*\}_{j=1:M}$ represents the sequence of extrema and $M << N$. Since maxima and minima always alternate, in 11 the plus refers to the maxima, while the minus to the minima. Without loss of generality, the first detected extremum is a maximum and the second one is a minimum; then maxima occur at odd intervals, i.e. t_{2j+1}^*, and minima occur at even intervals, i.e. t_{2j}^*. The second step of sifting builds an upper and lower envelope of $S(t)$ as two natural cubic splines through the sequence of maxima and the sequence of minima respectively. We therefore provide the semi-parametric forms for the conditions of the envelopes functions defined in 18 and 19 respectively. Note they should respect such conditions in principle, although guaranteeing them is a challenging task due to numerical undershoot or overshoot of the cubic spline. The two envelopes are then defined as:

$$S^{U_m}(t) = \sum_{j=1}^{M-1} \left(a_{2j+1} t^3 + b_{2j+1} t^2 + c_{2j+1} t + d_{2j+1} \right) \mathbb{1} \left(t \in \left[t_{2j}^*, t_{2j+1}^* \right] \right), \tag{12}$$

such that $S^{U_m}(t_{2j+1}^*) = S(t_{2j+1}^*)$ for all odd t_j^*. . Equivalently, the lower envelope corresponds to:

$$S^{L_m}(t) = \sum_{j=1}^{M-1} \left(a_{2j} t^3 + b_{2j} t^2 + c_{2j} t + d_{2j} \right) \mathbb{1} \left(t \in \left[t_{2j-1}^*, t_{2j}^* \right] \right), \tag{13}$$

such that $S^{L_m}(t_{2j}^*) = S(t_{2j}^*)$ for all even t_j^*.Next, one utilises these envelopes to construct the mean signal denoted by $m_m(t)$ given in equation 17, which will then be used to compensate the original speech signal $S(t)$ in a recursive fashion, until an IMF is obtained. The procedure is detailed in the following algorithm.

It is often the case that such an algorithm does not reach a mean equal to 0; therefore, multiple solutions in the literature have been proposed as stopping criteria of the sifting procedure. For further details, see **?**. From the sifting process, it is clear that these bases are recursively extracted; this means that, once the k-1 IMF is obtained, it is subtracted by the main signal and the sifting procedure is applied to the residual signal. Hence, it is highly essential to understand the linking relationship between the coefficients of two successive extracted IMFs. By exploiting the definition of cubic spline used in the representation of the analog speech signal $S(t)$ and the IMF basis functions, one can obtain a mathematical connection between the coefficients of $S(t)$ and the coefficients of $\gamma_m(t)$ detailed as follows:

Proposition 1. *The m-th extracted IMF denoted as $\gamma_m(t)$ can be expressed as a cubic spline whose coefficients are a linear combination of the spline coefficients of $S(t)$ and the coefficients of the $m-1$ IMFs extracted until such point of the sifting procedure and the coefficients of its mean envelopes, i.e.*

$$\gamma_m(t) = S(t) - \sum_{j=1}^{m-1} \gamma_j(t) - m_m(t) = \sum_{i=1}^{N-1} \left(a_i^m t^3 + b_i^m t^2 + c_i^m t + d_i^m \right) \mathbb{1}\left(t \in [t_{i-1}, t_i]\right) \tag{14}$$

where the spline coefficients are given as follows:

- $a_i^m = a_i - \sum_{j=1}^{m-1} a_i^j - \frac{1}{2}(a_i^{U_m} + a_i^{L_m})$
- $b_i^m = b_i - \sum_{j=1}^{m-1} b_i^j - \frac{1}{2}(b_i^{U_m} + b_i^{L_m})$
- $c_i^m = c_i - \sum_{j=1}^{m-1} c_i^j - \frac{1}{2}(c_i^{U_m} + c_i^{L_m})$
- $d_i^m = d_i - \sum_{j=1}^{m-1} d_i^j - \frac{1}{2}(d_i^{U_m} + d_i^{L_m})$

Such a proposition expresses the EMD construction of an IMF by considering the outer loop steps of the described algorithm. This means that, by looking at Algorithm **??**, the proposition considers 1), 2) and 3) to prove the statement. Note that in our notation $\gamma_m(t)$ in the case study, we suppressed the m upper script for the coefficients to avoid redundancy. The proof is provided in the appendix **??**.

4. Background: Empirical Mode Decomposition

4.1. Empirical Mode Decomposition

Definition 1. *The Empirical Mode Decomposition of signal $S(t)$ is represented by the Intrinsic Mode Functions finite basis expansion given by*

$$S(t) = \sum_{m=1}^{M} \gamma_m(t) + r(t) \tag{15}$$

here the collection of $\{\gamma_m(t)\}$ basis functions are known as the Intrinsic Mode Functions (IMFs) and $r(t)$ represents the final residual (or final tendency) extracted, which has only a single convexity. In general the γ_m basis will have k-convexity changes throughout the domain (t_1, t_N) and furthermore, each IMF satisfies the following mathematical properties:

- ***Oscillation*** *The number of extrema and zero-crossing must either equal or differ at most by one;*

$$abs\left(\left| \left\{ \frac{d\gamma_m(t)}{dt} = 0 : \quad t \in (t_1, t_N) \right\} \right| - \left| \{\gamma_m(t) = 0 : \quad t \in (t_1, t_N)\} \right| \right) \in [0, 1] \tag{16}$$

- ***Local Symmetry*** *The local mean value of the envelope defined by the local maxima and the envelope of the local minima is equal to zero pointwise i.e.*

$$m_m(t) = \left(\frac{S^{U_m}(t) + S^{L_m}(t)}{2} \right) \mathbb{1}\left(t \in [t_1, t_N]\right) = 0 \tag{17}$$

where the lower script m refers to the interested IMF. The minimum requirements of the upper and lower envelopes are:

$$\begin{aligned} &S^{U_m}(t) = \gamma_m(t), \;\; if \;\; \frac{d\gamma_m(t)}{dt} = 0 \quad \& \quad \frac{d^2\gamma_k(t)}{dt^2} < 0, \\ &S^{U_m}(t) > \gamma_m(t) \quad \forall t, \quad (t_1, t_N) \end{aligned} \tag{18}$$

$$S^{L_m}(t) = \gamma_m(t), \;\; if \;\; \frac{d\gamma_m(t)}{dt} = 0 \quad \& \quad \frac{d^2\gamma_m(t)}{dt^2} > 0, \\ S^{L_m}(t) < \gamma_m(t) \quad \forall t, \quad (t_1, t_N) \tag{19}$$

4.2. Instantaneous Frequency

Need to shrink it more - Dorota and Gareth

The goal of this section is understanding the concept of instantaneous frequency strictly related to the EMD above introduced. Classical Fourier methods require stationarity, where the frequency component is static over time, see **?**, **?**. Nevertheless, signals are often non-stationary and non-linear and, therefore, they carry time-varying frequency component in their IMFs basis. Though it is possible to have time-varying coefficients Fourier methods **?**, **?**, which tend to capture non-stationarity with fix basis, the EMD provides more flexibility. By being a data-driven, a posteriori method, its basis, i.e. the IMFs, are indeed more general. To find the instantaneous frequency, some steps have to be performed. The first one is to find the Hilbert Transform of each $\gamma_m(t)$, so that we can construct an analytic extension of the given IMF. The Hilbert Transform can be computed in close form only if $\gamma_m(t)$ respects the restrictions defined in 18 and 19.
Define $z(t) = \gamma_m(t) + \jmath\tilde{\gamma}_m(t)$ or $z(t) = a(t)e^{\jmath\theta(t)}$ the analytic extension of $\gamma_m(t)$, where $\tilde{\gamma}_m(t)$ represents the imaginary part and forms the conjugate pairs with $\gamma_m(t)$; also, $a(t) = \sqrt{\gamma_m^2(t) + \tilde{\gamma}_m^2(t)}$ corresponds to the amplitude of $z(t)$ and $\theta(t) = \arctan \frac{\tilde{\gamma}_m(t)}{\gamma_m(t)}$ to the instantaneous phase. The Hilbert Transform of an IMF $\gamma_m(t)$ is then given by:

$$\tilde{\gamma}_m(t) = -\frac{1}{\pi} \lim_{\epsilon \to \infty} \int_{-\epsilon}^{+\epsilon} \frac{\gamma_m(t+\tau) - \gamma_m(t-\tau)}{t} d\tau \tag{20}$$

Once that $\tilde{\gamma}_m(t)$ is computed, the second step corresponds to find the phase of the analytical signal defined as $\theta(t)$; afterwards, by differentiating such quantity with respect to t, the instantaneous frequency $f(t)$ is obtained as follows:

$$f(t) = \frac{1}{2\pi}\frac{d\theta(t)}{dt} = \frac{1}{2\pi}\frac{\tilde{\gamma}'_m(t)\gamma_k(t) - \tilde{\gamma}_m(t)\gamma'_m(t)}{\gamma_m^2(t) + \tilde{\gamma}_m^2(t)} \tag{21}$$

The instantaneous frequency is performed per IMF so that we can understand the local frequency and how it varies over time with each basis. To provide such concept in the context of non-stationary signals, **?** needed to detect local structures of the data by assuming equations 18 and 19. If such conditions of the IMFs are not satisfied, the instantaneous frequency often assumes negative values which lack physical meaning. Just like in the Fourier methods where there is a natural ordering of the static frequency (phase) for each basis, in this case, although the frequencies are time-varying, the extraction of the IMFs and property of the IMFs will still preserve the ordering in time of the instantaneous frequency.
The instantaneous frequencies derived from each IMF through the Hilbert Transform offer a full energy-frequency-time distribution. Such representation is known as the Hilbert Spectrum. The computation of the Hilbert Transform along with the instantaneous frequency in a closed form of a given IMF is provided below.
Assume that the interpolated signal $\tilde{s}(t)$ can be decomposed into components respecting 18 and 19. After the EMD and the Hilbert Transform of the IMFs are computed, $\tilde{s}(t)$ can be expressed in a Fourier-like expansion as:

$$\tilde{s}(t) = Re\left\{\sum_{k=1}^{K+1} a_m(t) \exp\{\jmath \int_{t_1}^{t_N} 2\pi f_m(t) dt\}\right\}$$

in which the residual $r(t)$ is included $(K+1)$. The index k refers to each IMF and $Re\{.\}$ denotes the real part of a complex quantity. This expansion, proposed in **?**, is known as the Hilbert- Huang transform (HHT). Note that the differences with the classical Fourier expansion are the amplitude a_k and the frequency f_k which are time-varying. The classical Fourier expansion is given by:

$$\tilde{s}(t) = Re\left\{\sum_{k=1}^{K+1} a_m \exp\{\jmath \int_{t_1}^{t_N} 2\pi f_m dt\}\right\}$$

To compute the instantaneous frequency of a given IMF, we firstly compute its Hilbert Transform. This is provided within the next proposition.

Proposition 2. *Consider the m-th IMF $\gamma_m(t)$ defined in 1 and remark that the form of its analytic signal is $z(t) = \gamma_m(t) + \jmath\tilde{\gamma}_m(t)$. The Hilbert Transform $\tilde{\gamma}_m(t)$ of $\gamma_m(t)$, defined at time points $S = \tau_1, \ldots, \tau_m = T$, is given in the the following equation:*

$$\tilde{\gamma}_m(t) = -\frac{1}{\pi} \lim_{\epsilon \to \infty} \int_{S-\epsilon}^{T+\epsilon} \frac{\gamma_m(t+\tau) - \gamma_m(t-\tau)}{t} d\tau \tag{22}$$

Such equation corresponds to:

$$\begin{aligned} \tilde{\gamma}_m(t) = \sum_{i=1}^{m} \left[a_i \frac{(\tau_{i-1} - \tau_i)^4}{4t} + b_i \frac{(\tau_{i-1} - \tau_i)^3}{3t} + c_i \frac{(\tau_{i-1} - \tau_i)^2}{2t} + d_i (\tau_{i-1} - \tau_i) \frac{1}{t} \right] - \\ \sum_{i=1}^{m} \left[a_i \frac{(\tau_{i-1} - \tau_i)^4}{4t} + b_i \frac{(\tau_{i-1} - \tau_i)^3}{3t} + c_i \frac{(\tau_{i-1} - \tau_i)^2}{2t} + d_i (\tau_{i-1} - \tau_i) \frac{1}{t} \right] \end{aligned} \tag{23}$$

The above results shows that the Hilbert transform of an IMF (which is represented by a cubic spline) is finite and exists. It is now possible to compute the instantaneous frequency of each IMF as the derivative with respect to the time as defined within 21.

5. EMD Feature Extraction

Dorota and Gareth - discussion how - TODO: I want a toy example for next week to show you I have to check notation here - wrong for spline To take into account both time and frequency domains, IMFs and related instantaneous frequencies are considered as features. By following our discussion in section 21, each IMF is a linear combination of cubic splines, therefore the coefficients of the IMFs splines are also considered as features. By remarking that non-stationarity is proper of sensor data, classical statistics are also a feature extracted on a window. The structure of the extraction process in the case study is given by: (1) discrete realizations of models are considered providing a segmentation indexed by t_i with $t_i \in \{0 = t_1, \ldots, t_n = n\}$; (2) a natural cubic spline is fitted through t_i and evaluated at $t_i^{'}$ such that $t_i^{'} \in \{0 = t_1^{'}, \ldots, t_N^{'} = N\}$, with $N = n$; (3) the EMD is performed over each interpolated signal and IMFs are stored; (4) instantaneous frequencies of IMFs are computed; (6) coefficients of the cubic spline of each IMF are calculated; (7) classical statistics are extracted by sliding a window of fixed length over an IMF such that $W\left[\tau_1, \tau_{j+1}\right] = W\left[\tau_{j+1}, \tau_{j+2}\right] = \cdots = W\left[\tau_{j+N-1}, \tau_{j+N}\right]$, where $\tau_j \in \{0 = \tau_1, \ldots, \tau_V = N\}$. It is worth noting that a method identifying an appropriate length of the window is beyond the final purpose of this work; therefore, windows are selected according to a trade-off between accuracy of the estimation (there should be enough number of samples to compute certain statistics) and stationarity. For each sensor data sample, 5 IMFs are considered: the first three with highest frequency, the lowest and the residual; then, from these 5 functions, the remaining features are extracted. We call them EMD features to underline their EMD derivation. The next table summaries such extraction:

Table 1. Extracted features. Description: $\tilde{c}_i = \min\left[\tau_i, \tau_{i+1}\right)$, $c_i^* = \max\left[\tau_i, \tau_{i+1}\right)$

EMD Feature	Label	Window
IMFs	$\gamma_1(t_i'), \gamma_2(t_i'), \gamma_3(t_i'), \gamma_k(t_i'), r(t_i')$	NA
Instantaneous Frequencies	$f_1(t_i'), f_2(t_i'), f_3(t_i'), f_k(t_i'), f_r(t_i')$	NA
Cubic Spline Coefficients	$\vec{b^1}(t_i'), \vec{b^2}(t_i'), \vec{b^3}(t_i'), \vec{b^k}(t_i'), \vec{b^r}(t_i')$	NA
	$\vec{c^1}(t_i'), \vec{c^2}(t_i'), \vec{c^3}(t_i'), \vec{c^k}(t_i'), \vec{c^r}(t_i')$	NA
	$\vec{d^1}(t_i'), \vec{d^2}(t_i'), \vec{d^3}(t_i'), \vec{d^k}(t_i'), \vec{d^r}(t_i')$	NA
Classical Statistics	$\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_k, \hat{\mu}_r$	$W\left[\tau_j, \tau_{j+1}\right]$
	$\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2, \hat{\sigma}_k^2, \hat{\sigma}_r^2$	$W\left[\tau_j, \tau_{j+1}\right]$
	$\tilde{c}_1, \tilde{c}_2, \tilde{c}_3, \tilde{c}_k, \tilde{c}_r$	
	$c_1^*, c_2^*, c_3^*, c_k^*, c_r^*$	$W\left[\tau_j, \tau_{j+1}\right]$
	$\hat{\beta}_{2_1}, \hat{\beta}_{2_2}, \hat{\beta}_{2_3}, \hat{\beta}_{2_k}, \hat{\beta}_{2_r}$	$W\left[\tau_j, \tau_{j+1}\right]$
	$\hat{k}_1, \hat{k}_2, \hat{k}_3, \hat{k}_k, \hat{k}_r$	$W\left[\tau_j, \tau_{j+1}\right]$
	$RMS_1, RMS_2, RMS_3, RMS_k, RMS_r$	$W\left[\tau_j, \tau_{j+1}\right]$

The considered classical statistics are (in order from the top to the bottom): mean, variance, minimum, maximum, kurtosis, skewness and root mean square (RMS). The length of the window will be later specified.

6. Review of Stationary and Non-stationary Kernels

Based on Bochner's theorem, the Fourier transform of a continuous shift-invariant positive definite kernel $K(x, x')$ is a proper probability distribution function $\pi(\omega)$, assuming that $K(x, x')$ is properly scaled, that is

$$K(x, x') = \int \pi(\omega) e^{i\omega^T (x - x')} \, d\omega = \mathbb{E}_\omega \left[\phi_\omega(x) \phi_\omega(x')^*\right] \tag{24}$$

for $\phi_\omega(x) = e^{j\omega^T x} = r\big(\cos(\omega x) + i \sin(\omega x)\big)$. The density of ω is denoted by spectral density.

TODO: produce some plots about the kernel choice for IMFS, plots like from the Turners presentation - elipsoids, a priori generated sample, a posteriori distribution given a few points

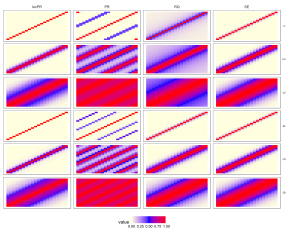


Figure 1. Stationary kernels under 6 different sets of hyper-parameters.

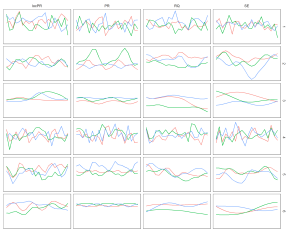


Figure 2. The 3 path of the signal c_k simulated from the a priori distribution in Equation (3) under different stationary kernel assumptions (columns wise) and for 6 different sets of hyper-parameters.

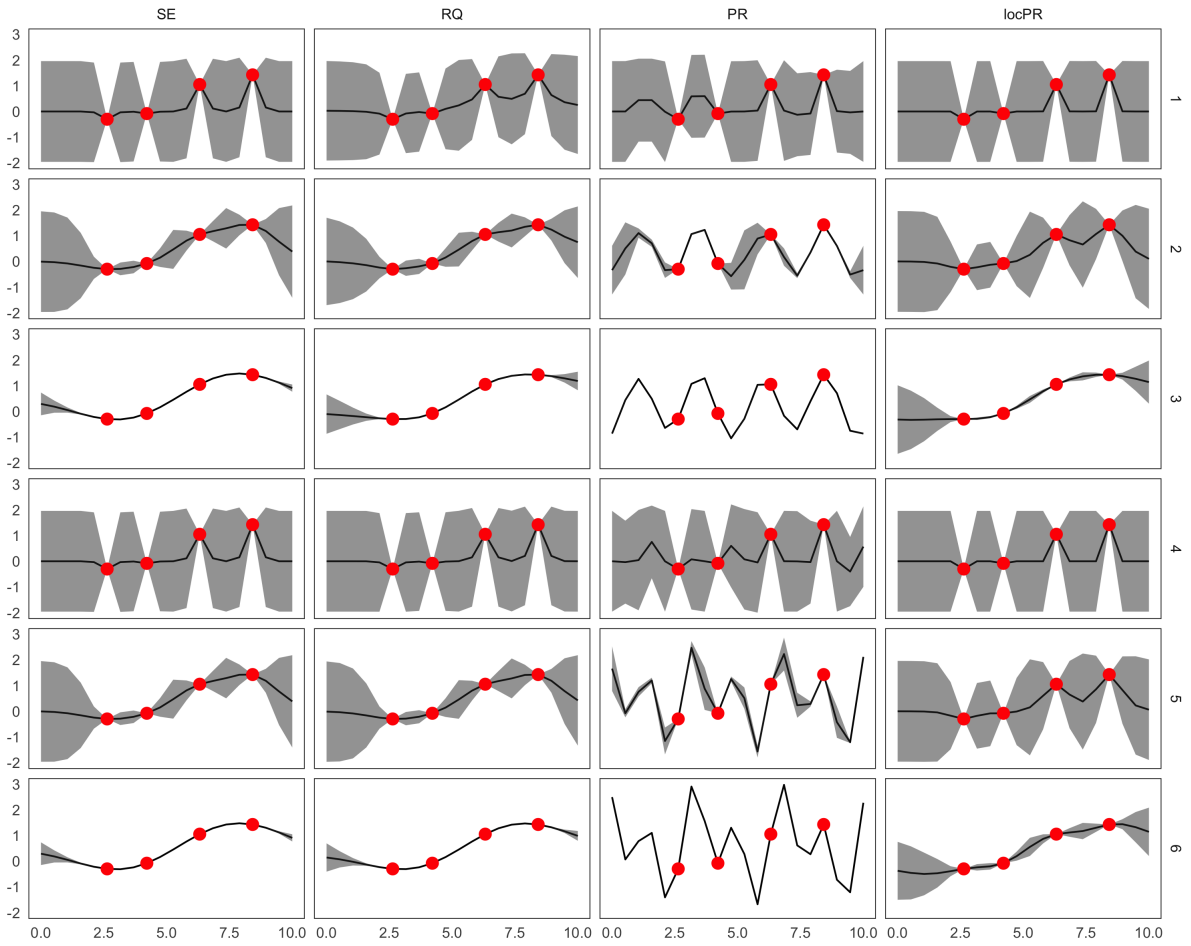


Figure 3. The predictive conditional mean of the IMF with the confidence intervals under noise-free assumption for different stationary kernel assumptions (columns wise) given 6 different sets of hyper-parameters. The red dots correspond to the observed values of the signal

Non-stationary Kernels

The non-stationary kernel can be characterised by a spectral density $\pi(\omega, \omega')$ such that

$$k(t,t') = \int\int \pi(\omega,\omega') e^{2\pi i \omega t - \omega' t'} \, d\omega d\omega' = \mathbb{E}_{\omega,\omega'}\left[\phi_{\omega,\omega'}(x)\phi_{\omega,\omega'}(x')^*\right] \tag{25}$$

where $\pi(\omega)$ is a spectral density of the kernel k over frequencies ω and ω'. For a scalar inputs x and x' the density $\pi(\omega, \omega')$ corresponds to bivariate distribution.

1. Non-stationary generalisation of the Squared Exponential kernel (GSE)

$$k_{GSE}(t,t') = \sigma(t)\sigma(t')\left(\frac{2\mu(t)\mu(t')}{\mu(t)^2+\mu(t')^2}\right)^{-\frac{1}{2}} \exp\left\{-\frac{(t-t')^2}{\mu(t)^2+\mu(t')^2}\right\}$$

for

$$\log\sigma(t) \sim \mathcal{GP}\left(\mu_\sigma, k_\sigma(t,t')\right)$$
$$\log\mu(t) \sim \mathcal{GP}\left(\mu_\mu, k_\mu(t,t')\right)$$

2. Spectral Mixture Kernel by (**?**) which is formulated as follows

$$k_{SM}(t,t') = \sum_{q=1}^{Q} w_q^2 \exp\Big\{ -2\pi^2 \tilde{\mathbf{t}}\boldsymbol{\Sigma}_q \tilde{\mathbf{t}}^T \Big\} \boldsymbol{\phi}_{\mu_q,\mu_q'}(t) \boldsymbol{\phi}_{\mu_q,\mu_q'}(t') \tag{26}$$

where

$$\tilde{\mathbf{t}} = \begin{bmatrix} t \\ -t' \end{bmatrix} \text{ and } \boldsymbol{\phi}_{\mu_q,\mu_q'}(t) = \begin{bmatrix} \cos(2\pi\mu_q t) + \cos(2\pi\mu_q' t) \\ \sin(2\pi\mu_q t) + \sin(2\pi\mu_q' t) \end{bmatrix}.$$

The spectral density of the kernel k_{SM} is defined by a weighted mixture

$$\pi(\omega,\omega') = \sum_{q=1}^{Q} w_q^2 \pi_q(\omega,\omega) \tag{27}$$

where $\pi_q(\omega,\omega)$ is a sum of bivariate normal densities with two dimension mean vectors are equal to the the eight combinations of the two element permutations of the set $\{\mu_q, \mu_q'\}$ and $\{-\mu_q, -\mu_q',\}$, and the covariance matrix $\boldsymbol{\Sigma}_q$.

3. Generalized Spectral Mixture Kernel by (**?**) with Gibbs kernel formulations which is formulated as follows

$$k_{GSM}(t,t') = \sum_{q=1}^{Q} w_q(t) w_q(t') \Big(\frac{2 l_g(t) l_q(t')}{l_g(t)^2 + l_q(t')^2} \Big)^{-\frac{1}{2}} \exp\Big\{ -\frac{(t-t')^2}{l_q(t)^2 + l_q(t')^2} \Big\} \cos\Big(2\pi(\mu_q(t)t - \mu_q(t')t') \Big) \tag{28}$$

for

$$\begin{aligned} \log l_q(t) &\sim \mathcal{GP}\Big(\mu_l, k_l(t,t')\Big) \\ \log it\mu_q(t) &\sim \mathcal{GP}\Big(\mu_\mu, k_\mu(t,t')\Big) \\ \log w_q(t) &\sim \mathcal{GP}\Big(\mu_w, k_w(t,t')\Big) \end{aligned}$$

The spectral density of the kernel k_{GSM} is ...

4. Sparse Spectrum Kernel by (**?**)

7. Brownian Bridge Analogue to construct IMFs

GP representation does not ensures itself that the predicted function from a given Gaussian process is IMF , that is, it satisfies (I1)-(I2). Therefore, we explire the following approches

Weiner process is a zero mean non-stationary Gaussian Process with the kernel $K(t,t') = min(t,t')$, that is

$$W(t) \sim \text{GP}(0, K(t,t')) \tag{29}$$

The Brownian Bridge for $t \in [0,T]$ is defined as

$$B(t) = W(t) - \frac{t}{T} W(T) \tag{30}$$

Therefore, it is also the Gaussian Process which is zero mean and has the covariance kernel equals to

$$\begin{aligned} \mathbf{Cov}(B(t),B(s)) &= \mathbf{Cov}(W(t),W(s)) - \frac{s}{T}\mathbf{Cov}(W(t),W(T)) - \frac{t}{T}\mathbf{Cov}(W(s),W(T)) + \frac{st}{T^2}\mathbf{Cov}(W(T),W(T)) \\ &= K(t,s) - \frac{s}{T}K(t,T) - \frac{t}{T}K(T,s) + \frac{ts}{T^2}K(T,T) \\ &= min(t,t') - \frac{ts}{T} \end{aligned}$$

The described process $B(t)$ satisfies that $B(0) = B(T) = 0$. The Brownian bridge which statisfied $B(t_0) = a$ and $B(_t1) = b$ is a solution to the following SDE system of equations

$$\begin{cases} dB(t) = dW(t) \\ B(t_0) = a \\ B(t_1) = b \end{cases} \text{, for } t_0 \leq t \leq t_1 \tag{31}$$

Therefore, the Brownian Bridge

and after calculations results in the form

$$B(t) = a + (b-a)\frac{t}{T} + W(t) - \frac{t}{T}W(T) \tag{32}$$

and therefore it is a Gaussian Process

$$B(t) \sim \mathrm{GP}\Big(a + (b-a)\frac{t}{T}, K(t,t') - \frac{tt'}{T}\Big) \tag{33}$$

for $0 \leq t \leq T$

7.0.1. Brownian Bridge Movement Model

Let $W(t)$ denote the Brownian Motion such that

$$W(t) \sim \mathcal{GP}, \big(0, k(t,t')\big) \tag{34}$$

with the kernel function $k(t,t') = \min\{t,t'\}$ and $t \in [0,T]$. Let us define the sequence of N points $0 \leq t_1 < t_2 < \ldots < t_N \leq < T$ such that we require that the process $W(t)$ had the fixed values at that points $W(t_i) = a_i$ for $i \in \{1,\ldots,N\}$. Theretofore, we are looking for a Gaussian process for $t \in [0,T]$ model of a conditional variable defined as

$$B(t) := W(t)|W(t_1) = a_1, W(t_2) = a_2, \ldots, \ldots W(t_N) = a_N \tag{35}$$

Let $\mathbf{W} = \big[W(t_1), W(t_2), \ldots, \ldots W(t_N)\big]$ and $\mathbf{t} = \big[t_1, t_2, \ldots, t_N\big]$ be N-dimensional vectors. Since $W(t)$ is a Gaussian Process, the random variable $W(t)|W(t_1), W(t_2), \ldots, \ldots W(t_N)$ is also a Gaussian Process with the conditional mean

$$\mu(t) := \mathbb{E}_{W(t)|W(t_1),W(t_2),\ldots,\ldots W(t_N)}\big[W(t)\big] = \mathbf{k}\big(t,\mathbf{t}\big)^T\mathbf{K}\big(\mathbf{t},\mathbf{t}\big)^{-1}\mathbf{W} \tag{36}$$

and the covariance function

$$\big(t,t') = \mathbb{E}_{W(t)|W(t_1),W(t_2),\ldots,\ldots W(t_N)}\Big[\big[W(t)-\mu(t)\big]\big[W(t')-\mu(t')\big]\big]\Big] = k(t,t') - \mathbf{k}\big(t,\mathbf{t}\big)^T\mathbf{K}\big(\mathbf{t},\mathbf{t}\big)^{-1}\mathbf{k}\big(t',\mathbf{t}\big) \tag{37}$$

where

$$\mathbf{K}(\mathbf{t},\mathbf{t}) := \begin{bmatrix} t_1 & t_1 & t_1 & \ldots & t_1 \\ t_1 & t_2 & t_2 & \ldots & t_2 \\ t_1 & t_2 & t_3 & \ldots & t_3 \\ \vdots & \vdots & \ddots & \vdots & \\ t_1 & t_2 & t_3 & \vdots & t_N \end{bmatrix}_{N\times N} \tag{38}$$

and

$$\mathbf{k}(t,\mathbf{t}) := \begin{bmatrix} \min(t,t_1) \\ \min(t,t_2) \\ \vdots \\ \min(t,t_N) \end{bmatrix}_{N\times 1} \tag{39}$$

7.1. Symmetric Local Extremas of IMFs

On every time internal there is a Brownian bridge or constrained Brownian bridge which starts and end from local extrema which are $x^{min}(t) = -x^{max}(t)$ for $t \in [\tau_i, \tau_{i+1}$

7.2. Nonsymmetric

7.3. Bayesian EMD

1. Construct a set of functions in Bayesian setting to have a IMF representation with restricted posterior (what needs to be satisfied on maxima and minima and how to ensure it) 2. Analogous of Brownian Bridge IMFs in Bayesian setting

Berger's optimal theory. Books on smoothing

7.4. Kernel Target Alignment

In this section we highlight an alternative method which estimates the Kernel matrix of the Gaussian Process of each IMF $\gamma_m(t)$, called Kernel Target Alignment (KTA) (Cristianini et al., 2002). There are two main encountered motivations behind such fact. The first one corresponding to a problem of identification of the Gaussian Process. It is well known (reference - need to check) that with Gaussian Processes, when it comes to several kernels (many of them - gives examples), the likelihood (or profile likelihood) of certain hyperparameters results to be a flat function (or very close to that) at the maximum. From a statistical perspective, this means that the model is not identifiable. Therefore, an identification challenge is found and there is an intuition for that: TODO: examples to generate: take a finite amount of data from a Gaussian Process - just radial basis function, which has two parameters: length scale (variance term) and shape scale. The variance term or length scale, in the profile likelihood (for certain amount of data) is not different if you move from t_1 to t_100 or t_1000, and so on. It can be really really flat - and the packages that you download - they do MLE - they just hit the barrier - whatever is the truncation barrier of the parameter space is - and report that barrier. Hence, the solution could be either the identification is forced, but this may be difficult if many kernels and, especially, complex kernels are employed, or an alternative procedure which does not suffer from identification issues is exploited. An example could be moment matching and the matrix version of it precisely corresponds to the Kernel Target Alignment (KTA). An important advantage of such method is that it is exact or unbiased since if you match exactly the second moment of a sample data and a Gaussian Process, you obtain a sufficient statistic (TODO: we are not just capturing partial information - it is equivalent in the sense of sufficiency or unbiasedness - we don't know about consistency of minimum variance - we need to work on it). The second motivation which justifies the employment of the KTA is the fact that it does not require the computation of the inverse of the Gram matrix; this operation corresponds to the order of $O(n^3)$ and can be highly expensive when dealing with big datasets. Moreover, the inversion is often numerically very unstable due to ill conditioning of the matrix.

In our context, we will consider realisations of a single IMF $\gamma_m(t)$ which is constructed through a smooth function; Gaussian Processes are realisations of functions which are smooth and therefore our population covariance matrix will be given by $\gamma_m(t)\gamma_m(t)'$ which is usually denoted in KTA frameworks as yy'. Our final objective will be identifying the set of hyperparameters for which the chosen kernel provides the moments which are as close as possible to the moments of the population model. This can be normally done through gradient based method (see Gareth notes for reeferences - Nystrom method); however, such techniques would require additional computational cost which are not needed in this settings since we will have few hyperparameters for each IMF to be estimated. The procedure will therefore as follows: for the Gaussian Process of each IMF we will select only one kernel (based on the in-sample learning check with Gareth). We then constraint maximum and minimum values of the hyperparameters of that kernel in a sensible way by building a grid. Afterwards, for each grid point, we evaluate the KTA and we choose the grid point which minimises the KTA. This will be the least difference between the population moments from the model and the sample model and the one of interest in our study.

Let us remark that the population covariance matrix in this case is given by $K_m(\boldsymbol{t}, \boldsymbol{t})$. In classification tasks, KTA consists of measuring the similarity between the "ideal" or target kernel and the one computed on the training set of the considered features. The alignment provides a measure for the degree of fitness of the given kernel. As given in Cristianini et al.,, the (empirical) alignment of a kernel k_1 with a kernel k_2 with respect to an (unlabelled) sample $S = \{x_1, \ldots, x_m\}$ is given by:

$$\hat{A}\left(S, k_1, k_2\right) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \tag{40}$$

where $\langle K_1, K_2 \rangle_F = \sum_{i,j=1}^{m} K_1(x_i, x_j) K_2(x_i, x_j)$ represents the inner product between Gram matrices. If $K_2 = yy'$, where y in our case corresponds to $\gamma_m(t)$, then the above equation becomes:

$$\hat{A}\left(S, K, \gamma_m(t)\gamma_m(t)'\right) = \frac{\langle K, \gamma_m(t)\gamma_m(t)' \rangle_F}{\sqrt{\langle K, K \rangle_F \langle \gamma_m(t)\gamma_m(t)', \gamma_m(t)\gamma_m(t)' \rangle_F}} = \frac{\langle K, \gamma_m(t)\gamma_m(t)' \rangle_F}{m\sqrt{\langle K, K' \rangle_F}}, \quad \text{since} \quad \langle \gamma_m(t)\gamma_m(t)', \gamma_m(t)\gamma_m(t)' \rangle_F = m^2 \tag{41}$$

Cristianini et al. proved that $\hat{A}$ gives a reliable estimate of its expected value by being concentrated around its mean.

TODO: add the part on CA which is the centered alignment which is required so that the moments of population model and sample covariance are actually true - we need to preprocess our GP (detrending it) or applying CA and not KTA. - I would prefer first option.

TODO: for next week - some example of this on R.