Nearest Neighbours Gaussian Process Model for Time-Frequency Data: An Application in Bio-acoustic Analysis

Hiu Ching Yip[1], Gianluca Mastrantonio[1], Enrico Bibbona[1], Marco Gamba[2], Daria Valente[2]

[1] Politecnico Di Torino, Italy
[2] Universita degli Studi di Torino, Italy

E-mail for correspondence: hiu.yip@polito.it

Abstract: Bio-acoustic signal analysis often reduces to feature analysis on the frequency structure in a lower dimensional space. This approach usually treats the time-frequency bins of spectrograms as independent features or extracts common statistics from waveforms. It is known to entail human perceptual bias that is induced by the neglect of the relative relationship between the spectral shape of vocalization and time as well as the dependence on domain knowledge of animals' behaviours. In light of this, we propose a Nearest Neighbour Gaussian Process (NNGP) model to account for the time varying components in the latent spectral structure of bio-acoustic data.

Keywords: NNGP; Time-frequency data; Latent spectral structure; Time-varying effects; Bio-acoustics

1 Motivation & Data

In comparative bio-acoustic studies, one area of interest is to understand the acoustic structures of non-human primates in order to provide insights on the evolution of the communication mechanism of our closest relatives. The most common practices are feature engineering methods, which involves selecting a set of basis-features for quantitative comparison. The identification of meaningful features in the vocal repertoire relies on biologists to observe and interpret the behavioural contexts in which the animals emit the signals. These interpretations are costly to acquire, inaccurate due to human subjectivity and difficult to generalize for cross-species comparison. Furthermore, feature selection always ignores the time-varying effect

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

of observed vocalizations on the latent acoustic structure. The aim of this project is to propose a NNGP model that accounts for time in bio-acoustic analysis. The dataset that will be available for model implementation are vocal signals of lemurs that were recorded in Madagascar.

The data format is equivalent to the published data in (Valente, D. et al. 2019). Each recorded signal is represented by a spectrogram and lasts for a unique duration of time that is measured in seconds. We refer to Figure 1 for a time-frequency representation of 3 signals of different durations. Furthermore, each signal is categorized by a call-type label and a species label, which are characterized by the behaviour of the lemur during emission and the species to which the lemur belongs to, respectively. As an example, Table 1 lists the number of recorded signals of 3 different species/call-type groups of signals. The group labels are given by biologists.

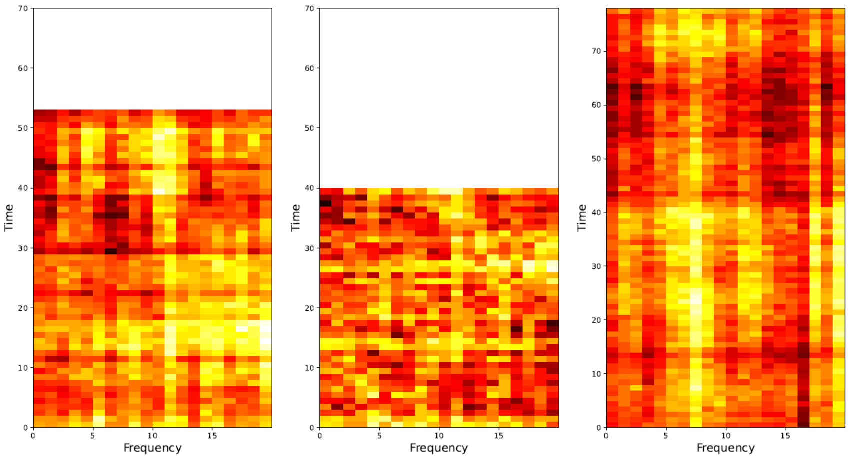

FIGURE 1. Spectrograms of 3 observed signals with different durations

TABLE 1. Number of recorded signals of 3 different species/call-type categories

species	call-type	# signals
Indri indri (II)	Clacson (CL)	622
Indri indri (II)	Grunt (GR)	1145
Indri indri (II)	Hum (HU)	418

2 NNGP Hierarchical Model

Let G denote the total number of observed signals of one species/call-type category. Denote each g-th observed signal by $\boldsymbol{z}_g$ where $g \in \{1, 2, ..., G\}$. Let t_g^* be its respective duration. Each $\boldsymbol{z}_g$ is a Gaussian field : $\boldsymbol{z}_g = \{ \ z_g(t, h) \ : \ (t, h) \in \mathcal{U}_g \ \}$ where (t, h) is a location in the observed spatial domain $\mathcal{U}_g = \mathcal{T}_g \times \mathcal{H}_g$, $\mathcal{T}_g = [0, t_g^*]$ is the time-axis and $\mathcal{H}_g$ is the frequency-axis. Each domain $\mathcal{U}_g$ is unique. Let $\boldsymbol{w}_{\mathcal{R}} = \{ \ w(t, h) \ : \ (t, h) \in \mathcal{R} \ \}$ be a latent Gaussian field of zero mean over $\mathcal{R} = \mathcal{T} \times \mathcal{H} \ : \ \mathcal{T} = [0, 1]$ that needs to be inferred from the data $\boldsymbol{z}_g$. This latent field $\boldsymbol{w}_{\mathcal{R}}$ is the inherent acoustic structure of a given set of signals of the same species/call-type category that has factored in the effects of each unique $\mathcal{U}_g$. The model is :

$$\begin{aligned} z_g(t, h) &= \mu_g + y_g(t, h) + \epsilon_g(t, h) \\ &= \mu_g + w(\alpha_g + \beta_g \ t, h) + \epsilon_g(t, h) \end{aligned}$$

where $y_g(t, h) = w(\alpha_g + \beta_g \ t, h)$ is a point value evaluated at the location $(\alpha_g + \beta_g \ t, h) \in \mathcal{R}$; $\alpha_g, \ \beta_g$ are the time-distortion parameters *i.i.d.* $\forall \ g$: $\alpha_g + \beta_g \ t \in \mathcal{T} = [0, 1] \ \forall \ t \in \mathcal{T}_g = [0, t_g^*]$; μ_g is the scalar mean *i.i.d* $\forall \ g$; and; $\epsilon_g(t, h) \sim N(0, \tau_g^2)$ is the random noise *i.i.d.* $\forall \ g, t, h$.

Let $C(\cdot)$ be the covariance kernel of that is specified by :

$$\begin{aligned} C((t_1, h_1), (t_2, h_2)) = \sigma^2 e^{-(\ \psi_t |t_1 - t_2| \ + \ \psi_h |h_1 - h_2| \ )} \\ + \ \sigma_c^2 e^{-\psi_c d_c(t_1, t_2, \gamma)} \end{aligned}$$

$\forall \ (t_1, h_1), (t_2, h_2) \in \mathcal{R}$. The first component of $C(\cdot)$ describes how the acoustic structure changes across the time-frequency grid $\mathcal{R}$. The second component addresses the circular nature of time-frequency data. The distance function $d_c(t_1, t_2, \gamma)$ is the periodic distance between two time points on $\mathcal{T}$ such that $d_c(t_1, t_2, \gamma) \in [0, \gamma/2] \ \forall \ t_1, t_2 \in \mathcal{T}$. The parameters of $C(\cdot)$ that need to be inferred are the time and frequency decay : $\psi_t, \ \psi_h$; the periodicity and its decay : $\gamma, \ \psi_c$; and ; the variances : $\sigma, \ \sigma_c$. Write $\boldsymbol{\Sigma}$ as the exact covariance matrix given by the kernel $C(\cdot)$ and $\boldsymbol{\theta} = \{\psi_t, \ \psi_h, \ \psi_c, \ \gamma, \ \sigma, \ \sigma_c\}$. The hierarchical model is :

$$\begin{aligned} \boldsymbol{z}_g \mid \mu_g, \ \boldsymbol{y}_g, \ \tau_g^2 &\sim \text{GP}(\ \mu_g + \boldsymbol{y}_g \ , \ \tau_g^2 \ ) \\ \boldsymbol{y}_g \mid \boldsymbol{\theta}, \ \alpha_g, \ \beta_g &\sim \text{GP}(0, \boldsymbol{\Sigma}) \\ \boldsymbol{w}_{\mathcal{R}} \mid \boldsymbol{\theta} &\sim \text{GP}(0, \boldsymbol{\Sigma}) \end{aligned}$$

We refer to Figure 2 for a graphical representation of the relationship between $\boldsymbol{w}_{\mathcal{R}}, \ \boldsymbol{y}_g$ and $\boldsymbol{z}_g$. The relative relationship between the times given by data and the spectral shape of $\boldsymbol{w}_{\mathcal{R}}$ is described by the time-distortion

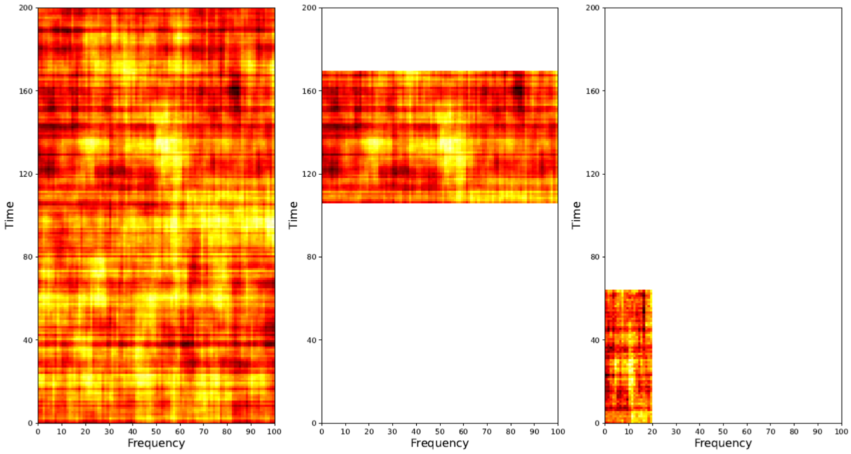

FIGURE 2. (From left to right) Spectrogram of $\boldsymbol{w}_{\mathcal{R}}$, $\boldsymbol{y}_g$ and $\boldsymbol{z}_g$ respectively

parameters : α_g, β_g, which map the data points in $\mathcal{U}_g$ onto the latent domain $\mathcal{R}$. Since $\boldsymbol{y}_g$ evaluates spatial locations in $\mathcal{R}$, it is thus specified by the same distribution of $\boldsymbol{w}_{\mathcal{R}}$. The data $\boldsymbol{z}_g$ can be marginalized over $\boldsymbol{y}_g = \{\ w(\alpha_g + \beta_g\ t, h)\ :\ (t,h) \in \mathcal{U}_g\ \}$. The marginal distribution of $\boldsymbol{z}_g$ over $\boldsymbol{y}_g$ is completely specified by the scalar mean μ_g, the noise τ_g^2, the time-distortion parameters α_g, β_g and the kernel parameters $\boldsymbol{\theta}$. Let k_g be the number of data points $z_g(t,h) \in \boldsymbol{z}_g$, the g-th observation. Define D_g as the diagonal matrix of dimension $k_g \times k_g$ with τ_g^2 as the diagonal entries. The marginal distribution is :

$$\begin{matrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \\ \boldsymbol{z}_G \end{matrix} \sim \mathrm{GP}\left(\begin{matrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_G \end{matrix}\ ,\ \begin{bmatrix} \boldsymbol{\Sigma}_{1,1} + D_1 & \boldsymbol{\Sigma}_{1,2} & \cdots & \boldsymbol{\Sigma}_{1,G} \\ \boldsymbol{\Sigma}_{2,1} & \boldsymbol{\Sigma}_{2,2} + D_2 & \cdots & \boldsymbol{\Sigma}_{2,G} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{G,1} & \boldsymbol{\Sigma}_{G,2} & \cdots & \boldsymbol{\Sigma}_{G,G} + D_G \end{bmatrix} \right)$$

Since inverting the high-dimensional exact covariance matrix $\boldsymbol{\Sigma}$ is too computationally expensive, we resort to the approximated NNGP instead of the exact GP. The idea of NNGP is that for Gaussian processes, if the covariance kernel is monotonic with respect to the distance between two spatial points, then only the data at neighbouring locations is needed for inference. Define the neighbour set $\mathcal{N}(t,h)$ as the set of m points that are "closest" to the point (t,h) such that the points in $\mathcal{N}(t,h)$ have the maximum correlation with point (t,h) given by $C(\cdot)$.

Let n denotes the total number of points in $w_{\mathcal{R}}$ and $(t_i, h_i) \in \mathcal{R}\ \forall\ i =$

$1, 2, ..., n$. The density of $w_{\mathcal{R}}$ that is expressed in terms of the full conditional densities can then be approximated in terms of the neighbour sets $\mathcal{N}(t, h)$. Write $\boldsymbol{z} = \{\ \boldsymbol{z}_1,\ \boldsymbol{z}_2,\ ...,\ \boldsymbol{z}_g\ \}$. The marginal distribution of $\boldsymbol{z}$ specified above can also be approximated similarly.

$$\begin{aligned}
\mathbb{P}(\boldsymbol{w}_{\mathcal{R}}) &= \ w(t_1, h_1) \prod_{i=2}^{n} \mathbb{P}\big(\ w(t_i, h_i) \mid \{w(t_j, h_j)\ :\ t_j \leq t_i\ ,\ h_j \leq h_i\}\ \big) \\
&\approx \ w(t_1, h_1) \prod_{i=2}^{n} \mathbb{P}\big(\ w(t_i, h_i) \mid \boldsymbol{w}_{\mathcal{N}(t_i, h_i)}\ \big) \\
\mathbb{P}(\boldsymbol{z}) &= \ z_1(t_1, h_1) \prod_{g=1}^{G} \prod_{i=1}^{k_g} \mathbb{P}\big(\ z_g(t_i, h_i) \mid \{z_{g'}(t_j, h_j)\ :\ g' \leq g\ ,\ \alpha_{g'} + \beta_{g'}\ t_j \leq \alpha_g + \beta_g\ t_i\ ,\ h_j \leq h_i\}\ \big) \\
&\approx \ z_1(t_1, h_1) \prod_{g=1}^{G} \prod_{i=1}^{k_g} \mathbb{P}\big(\ z_g(t_i, h_i) \mid \boldsymbol{z}_{\mathcal{N}(\alpha_g + \beta_g t_i, h_i)}\ \big)
\end{aligned}$$

References

Datta, A. et al. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *J Am Stat Assoc*, **111(514)**, 800-812.

Sainburg, T., Thielk, M., and Gentner, T.Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PloS Comput Biol*, **16(10)**, e1008228.

Valente, D. et al. (2019). Finding Meanings in Low Dimensional Structures: Stochastic Neighbor Embedding Applied to the Analysis of Indri indri Vocal Repertoire. *Animals : an open access journal from MDPI*, **9(5)**, 243.