A Hierarchical Spatio-Temporal Model for Time-Frequency Data: An application in bioacoustic analysis

Book of the Short Papers

Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni

SIS Società Italiana di Statistica

UNIVERSITÀ POLITECNICA DELLE MARCHE

Dipartimento di Scienze Economiche e Sociali DISES

LIUC Università Cattaneo

LIUC Università Cattaneo | BUSINESS ANALYTICS AND DATA SCIENCE HUB

Pearson

CHAIRS

Salvatore Ingrassia (Chair of the Program Committee) - *Università degli Studi di Catania*
Maria Cristina Recchioni (Chair of the Local Organizing Committee) - *Università Politecnica delle Marche*

PROGRAM COMMITTEE

Salvatore Ingrassia (Chair), Elena Ambrosetti, Antonio Balzanella, Matilde Bini, Annalisa Busetta, Fabio Centofanti, Francesco M. Chelli, Simone Di Zio, Sabrina Giordano, Rosaria Ignaccolo, Filomena Maggino, Stefania Mignani, Lucia Paci, Monica Palma, Emilia Rocco.

LOCAL ORGANIZING COMMITTEE

Maria Cristina Recchioni (Chair), Chiara Capogrossi, Mariateresa Ciommi, Barbara Ermini, Chiara Gigliarano, Riccardo Lucchetti, Francesca Mariani, Gloria Polinesi, Giuseppe Ricciardo Lamonica, Barbara Zagaglia.

ORGANIZERS OF INVITED SESSIONS

Pierfrancesco Alaimo Di Loro, Laura Anderlucci, Luigi Augugliaro, Ilaria Benedetti, Rossella Berni, Mario Bolzan, Silvia Cagnone, Michela Cameletti, Federico Camerlenghi, Gabriella Campolo, Christian Capezza, Carlo Cavicchia, Mariateresa Ciommi, Guido Consonni, Giuseppe Ricciardo Lamonica, Regina Liu, Daniela Marella, Francesca Mariani, Matteo Mazziotta, Stefano Mazzuco, Raya Muttarak, Livia Elisa Ortensi, Edoardo Otranto, Ilaria Prosdocimi, Pasquale Sarnacchiaro, Manuela Stranges, Claudia Tarantola, Isabella Sulis, Roberta Varriale, Rosanna Verde.

FURTHER PEPOPLE OF LOCAL ORGANIZING COMMITTEE

Elisa D'Adamo, Christian Ferretti, Giada Gabbianelli, Elvina Merkaj, Luca Pedini, Alessandro Pionati, Marco Tedeschi, Francesco Valentini, Rostand Arland Yebetchou Tchounkeu

Technical support: Matteo Mercuri, Maila Ragni, Daniele Ripanti

PUBLISHED BY PEARSON
WWW.PEARSON.COM
ISBN 9788891935618AAVV

A Hierarchical Spatio-Temporal Model for Time-Frequency Data: An application in bioacoustic analysis

Hiu Ching Yip[a], Gianluca Mastrantonio[a], Enrico Bibbona[a], Daria Valente[b], and Marco Gamba[b]

[a]Politecnico di Torino, Italy; hiu.yip@polito.it, gianluca.mastrantonio@polito.it, enrico.bibbona@polito.it

[b]Università di Torino, Italy; daria.valente@unito.it, marco.gamba@unito.it

Abstract

A hierarchical spatio-temporal model that infers the latent spectral shape from a set of bio-acoustic signals by means of the Nearest neighbour Gaussian process is proposed. The model aims to account for the effects of the relative relationship between time and the spectral shape of the recorded vocalizations and that of time discretization. The goal is to obtain a representative model of the inherent acoustic structure of the species.

Keywords: Bio-acoustic, time-frequency, spatio-temporal model, nearest neighbour Gaussian process, spectral shape

1. Motivation & Data

In comparative bio-acoustic studies, one area of interest is to understand the vocalizations of non-human primates in order to provide insights into the evolutionary mechanism of the communication systems of our closest relatives. Since bioacoustic data are almost always represented in the form of a spectrogram, bioacoustic analysis is therefore a form of time-frequency analysis that requires computational methods to process and learn from the bio-acoustic signals in large quantities. The most commonplace practice is to apply feature engineering methods in order to select and compare a set of basis-features. Such methods often treat the time-frequency bins of spectrograms as independent features and are known to entail perceptual bias due to the reliance on biologists to manually select relevant features. The identification and interpretation of meaningful features are usually costly to acquire and difficult to generalize for cross-species comparison. Furthermore, feature engineering methods in bioacoustic analysis, whether supervised or unsupervised, almost always ignore the effects of time by assuming that all the time-frequency bins from various recorded bio-acoustic signals are independent of each other. The aim of this project is to propose a spatio-temporal model that accounts for the effects of time in bio-acoustic analysis.

The available dataset for this work is a set of vocal signals of lemurs that were recorded in Madagascar. The format of the dataset is similar to those in (2). Each recorded analogue digital signals is

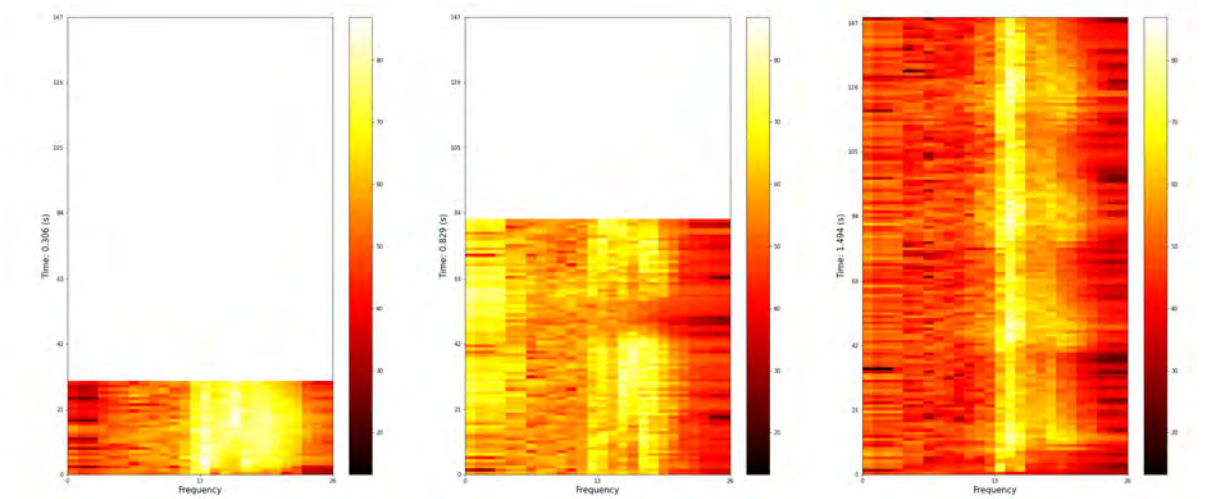

Figure 1: The spectrograms of three recorded signals labelled by the same species label and the same call-type from the dataset. The x-axis and y-axis represent the discretized frequency and time domain, respectively. Each discrete sound point of the discretized recording is measured in decibel scale on the third dimension of the spectrogram. Each axis is labelled by the number of discrete time and frequency coordinates. The number of frequency coordinates on the x-axis is the same for all recorded signals, while the number of time coordinates on the y-axis differ. This discernible disparity between the numbers of time coordinates on the y-axes is a result of the difference between the durations of the three signals. The unique duration of each signal is indicated on the label of the y-axis, which are 0.306, 0.829 and 1.494 seconds respectively (from left to right). The corresponding numbers of time coordinates are 30, 82, and 149 respectively, while all have 26 frequency coordinates.

discretized using a constant time-step of 0.01 seconds during the initial stage of signal pre-processing. Figure 1 is a spectrogram representation of 3 discretized recorded signals of different durations from the dataset. Each signal is categorized by one species label and one call-type label. The species label is the acronym of the scientific name of the lemur that emitted the recorded signal, whereas the call-type label of each recording is assigned according to the behaviour of the animal during the emission of the signal.

2. The Model

Assume that the recordings that are characterized by the same species but different behavioural call types are independent from each other. The model specification is then restricted to the recorded signals that are labelled by a single call type from a single species. Let N be the total number of signals that are classified by the same combination of species and call type with $i = 1, \cdots, N$. As per Figure 1, each recorded signal is represented by a spectrogram with one axis representing the time domain and another representing the frequency domain. Assume that each i-th recorded signal is a realization of a two-dimensional process $\mathcal{Y}_i(t, h) \in \mathbb{R}$ where $t \in \mathbb{R}_{\geq 0}$, $h \in \mathbb{R}$ over an observed regular grid. Let $n_{t,i}$ denote the number of time coordinates on the discrete time axis $\mathcal{T}_i^* \subset [0, l_i]$ where l_i is the duration of the signal and let n_h denote the number of log-frequency bins on the frequency axis $\mathcal{H}$, respectively. The regular grid of the i-th recorded signal is then composed of $n_i = n_{t,i} \times n_h$ time-frequency coordinates in total. In order to compare the recorded signals of different durations, the time domain of each recorded

signal $\mathcal{T}_i^* \subset [0, l_i]$ is rescaled into a new time domain, denoted by $\mathcal{T}_i \subset [0, 1]$, such that it is always one in duration. By contrast, the number of frequency coordinates n_h is constant for all signals and so, it follows that the log-frequency bands that are denoted by the frequency coordinates on $\mathcal{H}$ are also the same for all recorded signals. Hence, let the realization of the process $\mathcal{Y}_i(t, h)$ on a regular grid be denoted by $\boldsymbol{y}_i = \{y_{i,t,h}\}_{t \in \mathcal{T}_i, h \in \mathcal{H}}$ where

$$\mathcal{T}_i = \left\{ \frac{k-1}{n_{t,i}-1} | k = 1, \ldots, n_{t,i} \right\}, \quad \mathcal{H} = \{0.23k + \log 63 | k = 1, \ldots, n_h\} \tag{1}$$

Clearly, each observed spectrogram represented by the $n_{t,i} \times n_h$ regular grid $\mathcal{T}_i \times \mathcal{H}$ is unique in its own right as a consequence of the varying numbers of time coordinates $n_{t,i}$. As explained earlier on, the aim of the model is to infer the latent spectral shape of vocalizations from a dataset of N recorded signals that share the same species and call type labels. It is reasonable to assume that all N recorded acoustic signals have the same inherent spectral shape which can be described by the same latent Gaussian process. This ideal representation of the inherent spectral shape of the vocalizations from a single species of a single call type is henceforth termed the "mother call". Let $\mathcal{W}_i(t, h)$ be the latent process that describes the mother call. Assume that if $(i, t, h) \neq (i', t', h')$, then $\mathcal{Y}_i(t, h)$ and $\mathcal{Y}_{i'}(t', h')$ are conditionally independent given the mother call. The model is:

$$\begin{aligned} \mathcal{Y}_i(t, h) &= \mu_i + \mathcal{W}(t, h) + \epsilon_i(t, h) \\ \mathcal{W}(t, h) &\sim \text{GP}(0, C(\cdot, \cdot | \boldsymbol{\theta})) \\ \epsilon_i(t, h) &\sim \text{GP}(0, \tau_i^2) \end{aligned} \tag{2}$$

where $\mu_i \in \mathbb{R}$ is the mean sound intensity of the i-th recorded signal, $\mathcal{W}(t, h)$ is the latent process over the domain of the mother call, $\epsilon_i(t, h) \sim \text{GP}(0, \tau_i^2)$ is the $i.i.d.$ random noise, and

$$C((t, h), (t', h') | \boldsymbol{\theta}) = Cov(\mathcal{W}(t, h), \mathcal{W}(t', h')) : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^+ \tag{3}$$

is the cross-covariance function for the latent process $\mathcal{W}(\cdot, \cdot)$ that is dependent on the vector of parameters $\boldsymbol{\theta}$. The covariance function $C(\cdot, \cdot | \boldsymbol{\theta})$ for the latent process $\mathcal{W}(\cdot, \cdot)$ is defined by

$$C((t, h), (t', h') | \boldsymbol{\theta}) = \sigma^2 \text{exp}(-\phi_t |t - t'| + \phi_h |h - h'|) \tag{4}$$

such that $\boldsymbol{\theta} = (\sigma, \phi_t, \phi_h)$. The parameters that need to be inferred are the time-frequency decay ϕ_t, ϕ_h and the variance σ, respectively. Note that the equivalent formulation is the generative model $\mathcal{Y}_i(t, h) \mid \mathcal{W}(t, h) \sim \text{GP}\big(\mu_i + \mathcal{W}(t, h), \tau_i^2\big)$. Dependence between the entire set of recorded signals is thus introduced through the latent process $\mathcal{W}(t, h)$ which describes the mother call. That is, if the model in equation (2) is marginalized over the latent process, then the observed processes are dependent on each other. Subsequently, the acoustic structures of each recorded signal is composed of the natural change in the spectral shape across the observed time-frequency grid with independent error τ_i^2, which need to be accounted for by an additional component on the diagonal of the covariance matrix for the observed processes. This is the nugget effect that arises from the covariance of the variables in each observed process. Let $\mathbb{I}(\cdot)$ be an indicator function, then the covariance function for the observed processes is:

$$Cov\big(\mathcal{Y}_i(t, h), \mathcal{Y}_{i'}(t', h')\big) = C((t, h), (t', h') | \boldsymbol{\theta}) + \tau_i^2 \mathbb{I}((i, t, h) = (i', t', h')) \tag{5}$$

Since direct implementation of the generative model in equation (2) is computationally infeasible due to the fact that it requires the mother call to be sampled for every single recorded sound point in the dataset, the marginalized model is implemented instead. To simplify notations, re-write each single recorded sound point $y_{i,t,h} = y_{i,j}$ where $y_{i,j}$ is the realization of $\mathcal{Y}_i(t, h)$. Define $\boldsymbol{y}_i = \{y_{i,j} | j = 1, \cdots, n_i\}$ as the vector of realizations from the i-th recorded signal such that the elements are sorted in the ascending order of time and the increasing value of log-frequencies within each time. Define $\mathbf{1}_i$ as the $n_i \times 1$ vector of ones. Write $\boldsymbol{\Sigma}_i$ as the exact covariance matrix of $\boldsymbol{y}_i$ and write $\boldsymbol{\Sigma}_{i,i'}$ as the exact

cross-covariance matrix of $\boldsymbol{y}_i$ and $\boldsymbol{y}_{i'}$ that are given by the covariance function in equation (5). The joint density of all realizations can be expressed by:

$$\begin{matrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\\ \vdots\\ \boldsymbol{y}_N\end{matrix} \sim \mathrm{N}\left(\begin{pmatrix}\mu_1\mathbf{1}_1\\ \mu_2\mathbf{1}_2\\ \vdots\\ \mu_3\mathbf{1}_N\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{1,2} & \cdots & \boldsymbol{\Sigma}_{1,N}\\ \boldsymbol{\Sigma}_{2,1} & \boldsymbol{\Sigma}_2 & \cdots & \boldsymbol{\Sigma}_{2,N}\\ \vdots & \cdots & \ddots & \vdots\\ \boldsymbol{\Sigma}_{N,1} & \boldsymbol{\Sigma}_{N,2} & \cdots & \boldsymbol{\Sigma}_N\end{pmatrix}\right) \tag{6}$$

In a more compact form, write $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i=1,\cdots,N}$ as the collection of all realizations and re-write formula (6) into $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{y}})$.

3. Implementation

The available dataset of lemur signals for this paper has a total number of time-frequency coordinates of $n = \sum_i^N n_{t,i} n_h$. With a computational cost of $\mathcal{O}(n^3)$, the inversion of the exact covariance matrix in equation (6) is too computationally expensive; accordingly, one of the methods for efficiently and accurately approximating Gaussian processes, namely the Nearest Neighbours Gaussian Process (NNGP) method, is adopted in this work. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ denote the density of the realizations $\boldsymbol{y}$ that depends on the parameters $\boldsymbol{\theta}$ which can be decomposed by conditioning as follow:

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}) &= f(\boldsymbol{y}_1|\boldsymbol{\theta}) \prod_{i=2}^{N} f(\boldsymbol{y}_i|\boldsymbol{\theta}) \\ &= f(y_{1,1}|\boldsymbol{\theta}) \prod_{j=2}^{n_1} f(y_{1,j}|y_{1,j-1}, y_{1,j-2}, \cdots, y_{1,1}, \boldsymbol{\theta}) \times \\ &\prod_{i=2}^{N} f(y_{i,1}|\boldsymbol{y}_{i-1}, \boldsymbol{y}_{i-2}, \cdots, \boldsymbol{y}_1, \boldsymbol{\theta}) \prod_{j=2}^{n_i} f(y_{i,j}|y_{i,j-1}, y_{i,j-2}, \cdots, y_{i,1}, \boldsymbol{y}_{i-1}, \boldsymbol{y}_{i-2}, \cdots, \boldsymbol{y}_1, \boldsymbol{\theta}) \end{aligned} \tag{7}$$

The idea of the NNGP method is that for a Gaussian process which is stationary, if the covariance function is monotonic with respect to the distances between the spatio-temporal coordinates, then only the immediate neighbourhoods rather than the entire conditional sets are necessary to approximate the likelihoods of the realizations of the process. Define $\mathcal{N}_{i,j}$ as a subset of variables in the conditional set of $y_{i,j}$ which is the immediate neighbourhood called the neighbour set. The elements of the neighbour set are called neighbours. Since the covariance function in equation (4) is monotonically decreasing with respect to the form of distance that it depends on, the neighbours in the neighbour set should have minimal non-zero distances and the formation of the neighbour set should be characterized by a distance function that measures the absolute time-frequency lags on the time-frequency domain. The above density of $\boldsymbol{y}$ can then be approximated into:

$$f(\boldsymbol{y}|\boldsymbol{\theta},) \approx \prod_{i=1}^{N} \prod_{j=1}^{n_i} f(y_{i,j}|\mathcal{N}_{i,j}, \boldsymbol{\theta}) \tag{8}$$

with $\mathcal{N}_{1,1}$ being an non-empty neighbour set. The distance function that gives rise to the neighbour set is defined by :

$$d((t,h),(t',h')) = |t-t'| + |h-h'| \tag{9}$$

It is more reasonable to select neighbours for $y_{i,j}$ from both the same i-th signal and the previous $i-1$-th signal instead of just the same recorded signal alone on the grounds that the realizations $\boldsymbol{y}_i$ and $\boldsymbol{y}_{i-1}$ are not independent in spite of the product form of the approximated density in the above equation (8). Spatio-temporal dependence between the two different recordings has to be re-introduced

into the approximated density through the neighbour sets. For the j-th observed sound variable of the i-th recorded signal, denote the neighbour set with elements selected only from the i-th signal by $\mathcal{N}_{i,j}^{i}$ and similarly, denote the neighbour set with elements selected only from the previous $i-1$-th signal by $\mathcal{N}_{i,j}^{i-1}$. The definition of the neighbour set for the realized variable $y_{i,j}$ is:

$$\mathcal{N}_{i,j} = \mathcal{N}_{i,j}^{i} \cup \mathcal{N}_{i,j}^{i-1} \tag{10}$$

Following (1), setting the size of the neighbour set in between 10 to 20 should enable an accurate approximation of the original process. Finally, the main objective of this work is to obtain the representative acoustic structure of the vocalizations that belong to the same species and call-type, the mother call, which is the finite realizations of $\mathcal{W}(t,h)$ over a specified grid. By the NNGP method, the approximated precision matrix $\mathbf{\Sigma}_{\boldsymbol{y}}^{-1}$ admits a Cholesky decomposition and enables standard posterior sampling. Let $\mathbf{\Sigma}_{w,\boldsymbol{y}}$ be the $1 \times n$ cross-covariance vector of $\mathcal{W}(t,h)$ and $\boldsymbol{y}$. The realization of the latent process $\mathcal{W}(t,h)$ at any location can then be obtained by:

$$\mathcal{W}(t,h)|\boldsymbol{y},\boldsymbol{\theta} \sim \mathrm{N}\big(\mathbf{\Sigma}_{w,\boldsymbol{y}}\mathbf{\Sigma}_{\boldsymbol{y}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}),\ \sigma^2 - \mathbf{\Sigma}_{w,\boldsymbol{y}}\mathbf{\Sigma}_{\boldsymbol{y}}^{-1}\mathbf{\Sigma}_{w,\boldsymbol{y}}^{\top}\big) \tag{11}$$

4. Discussion

As this is an ongoing and multidisciplinary work, the current proposed model is a preliminary version of the complete model and as such, further progress to the final results is being made. Another preliminary and less refined version of this work has been communicated to the 36th International Workshop on Statistical Modelling in the form of a poster presentation and the submitted abstract was published in the proceedings that is available in (3).

The current preliminary model in this paper, although incomplete, starkly contrasts many contemporary methods in bio-acoustic analysis. The model accounts for the effects of the time-varying components in the observed vocalizations by using the entire bio-acoustic dataset as realizations of a spatio-temporal process, rather than reducing the high-dimensional time-frequency data into some independent features, such as absolute pitches, with arbitrarily assigned meanings. The main contribution is that the resulting mother call, i.e. the representative acoustic structure of a particular call-type from a species, can subsequently be used to measure the distances between the inherent acoustic structures of different species in the animal kingdom. Such quantitative measures can hopefully ease cross-species comparison in bio-acoustic analysis and facilitate biological studies on the evolutionary basis for the variations of the vocal repertoires of various species. The next steps being taken in this work are summarized as follow.

A closer inspection of the spectrograms of the observed signals in Figure 1 reveals that there are two major issues with the preliminary model in equation (2): (i.) the noticeable distortions of the observed time domains with respect to each other caused by the unique duration of each recordings. Though each recorded signal can be thought of a marginal realization of the same latent process over the observed regular grid, the unique duration of each recording entails distortions of the observed time-axis with respect to each other, and perhaps misalignments of the time domain for the observed process with respect to the time domain for the latent mother call process. The same earliest recorded time coordinate from the 1st recorded signal might not coincide with the very same earliest recorded time coordinate from another signal, for instance. In fact, each observed time domain may be treated as a somewhat stretched portion of the latent time domain for the mother call. (ii.) the oscillations along the time axis on the lower frequency spectrum that arises from time discretization and signal reconstruction during the initial stage of analogue signal pre-processing. When analogue signals are being discretized in time, if the time-step is less than the period of the true waveform of the low frequency, then the sampled frequency becomes an artifact because it does not capture the original periodicty of the true waveform of the sound. This leads to a periodic artefact that oscillates along the time domain of the reconstructured, discretized signal on the lower frequency spectrum.

Both (i.) and (ii.) render the model specification and inference for the mother call more difficult as the challenges posed by the misalignments of the time domains as well as the presence of the artefacts must be resolved. In light of (i.), the preliminary model and its separable covariance function must be

extended to incorporate a time-distortion function with parameters that can describe and quantify the distortions of the time domains of the observed processes with respect to the latent time domain of the mother call. In view of (ii.), an additional component that can explain the periodicity of the artefacts must be included into the covariance function of the observed processes. Considering that the sampling artefacts appear solely in the recorded discretized signals, care must be taken to ensure that only the covariance for the observed processes must possess this additional component, but not the covariance for the mother call. The parameters of the time-distortion function and this additional component must be part of the inference. Once the final model is formulated, it is tested on simulated data in order to demonstrate its efficacy and to recognize if the model suffers from the problem of non-identifiability or any other technical issue in the implementation stage. After all the implementation details are sorted, the final model is then ready to be implemented on the real dataset.

References

[1] Datta A., Banerjee S., Finley A.O. and Gelfand A.E.: Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Am. Stat. Assoc. (2016) **111**(514), 800–812

[2] Valente D., De Gregorio C., Torti V., Miaretsoa L., Friard O., Randrianarison R.M., Giacoma C. and Gamba M.: Finding meanings in low dimensional structures: stochastic neighbor embedding applied to the analysis of Indri indri vocal repertoire. Animals(Basel). (2019) **9**(5):243. doi:10.3390/ani9050243. PMID: 31096675; PMCID: PMC6562776.

[3] Yip H.C., Mastrantonio G., Bibbona E., Gamba M., Valente D.: Nearest neighbours Gaussian process model for time-frequency data: An application in Bio-acoustic Analysis. Proceedings of the 36th International Workshop on Statistical Modelling. (July 18-22, 2022. Trieste, Italy.) Available via https://www.openstarts.units.it/handle/10077/33740