Quantifying Statistical Interdependence PART III: $n > 2$ Multi-Dimensional Point Processes

J. Dauwels [a,b,*,1] T. Weber [c] F. Vialatte [d] T. Musha [e] A. Cichocki [d]

[a] *Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA.*

[b] *Amari Research Unit, RIKEN Brain Science Institute, Saitama, Japan.*

[c] *Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.*

[d] *Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Saitama, Japan.*

[e] *Brain Functions Laboratory, Inc., Yokohama, Japan.*

Abstract

Stochastic event synchrony (SES) is a technique to quantify the similarity of pairs of signals. In this paper (Part III), SES is extended from pairs of signals to collections of signals. As in Part I and II, first "events" are extracted from the given time series. Next, one tries to align events from one time series with events from the other. The better the alignment, the more similar the collection of time series is considered to be. As in Part II, this paper deals with multi-dimensional events. Although the basic idea is similar to the pairwise case, the extension to collection of point processes involves an NP-hard combinatorial problem, and therefore, it is far from trivial.

17 October 2009

The problem of jointly computing the alignment and SES parameters is again cast as a statistical inference problem. This problem is solved by coordinate descent, more specifically, by alternating the following two steps: (i) one estimates the SES parameters from a given alignment; (ii) with the resulting estimates, one refines the alignment. The SES parameters are computed by maximum a posteriori (MAP) estimation (Step 1), in analogy to the pairwise case. The alignment (Step 2) is solved as an integer program.

In order to test the robustness and reliability of the proposed multivariate SES method, it is first applied to surrogate data. Next it is applied to detect anomalies in EEG synchrony of Mild Cognitive Impairment (MCI) patients. The multivariate approach helps to further improve the diagnosis and enables a more detailed analysis.

Key words: coincident event, maximum a posteriori estimation, Morris-Lecar neuron model, EEG, Alzheimer's disease, Mild Cognitive Impairment (MCI)

1 Introduction

Quantifying the interdependence between signals or time series is an important but challenging problem. Although it is straightforward to quantify linear de-

* Corresponding author.
Email addresses: justin@dauwels.com (J. Dauwels), theo_w@mit.edu (T. Weber), fvialatte@brain.riken.jp (F. Vialatte), musha@bfl.co.jp (T. Musha), cia@brain.riken.jp (A. Cichocki).

[1] J.D. was in part supported by post-doctoral fellowships from the Japanese Society for the Promotion of Science (JSPS), the King Baudouin Foundation, and the Belgian American Educational Foundation (BAEF). Part of this work was carried out while J.D. and T.W. were at the RIKEN Brain Science Institute, Saitama, Japan.

pendencies, the extension to non-linear correlations is far from trivial. In this paper, we introduce a new exemplar-based measure of statistical interdependence between an arbitrary number of spatial point processes. It can also be applied to multidimensional signals, after they have been converted into point processes which capture "bursts" of activity of the signal in some appropriate domain. As such, we attempt to measure the synchrony of the main patterns in the data, while ignoring background activity (which can be intrinsic to the system studied, or which can be noise).

This paper is organized as follows. In the following section, we outline the exemplar-based statistical model for synchrony; in Section 4 we describe how to perform inference in that model, and we characterize the underlying combinatorial problem. Lastly, we apply our method to detect MCI induced perturbations in EEG synchrony (Section 6.1). At the end of the paper, we make some concluding remarks.

2 Principle

Suppose that we are given a pair of continuous-time signals, e.g., EEG signals recorded from two different channels, and we wish to determine the similarity of those two signals. As a first step, we extract point processes from those signals, which may be achieved in various ways. As an example, we generate point processes in time-frequency domain: first the time-frequency ("wavelet") transform of each signal is computed in a frequency band $f \in [f_{\min}, f_{\max}]$. Next those maps are approximated as a sum of half-ellipsoid basis functions, referred to as "bumps" (see Fig. 3; we will provide more details on bump modeling in Section 6.2.3). Each bump is described by five param-

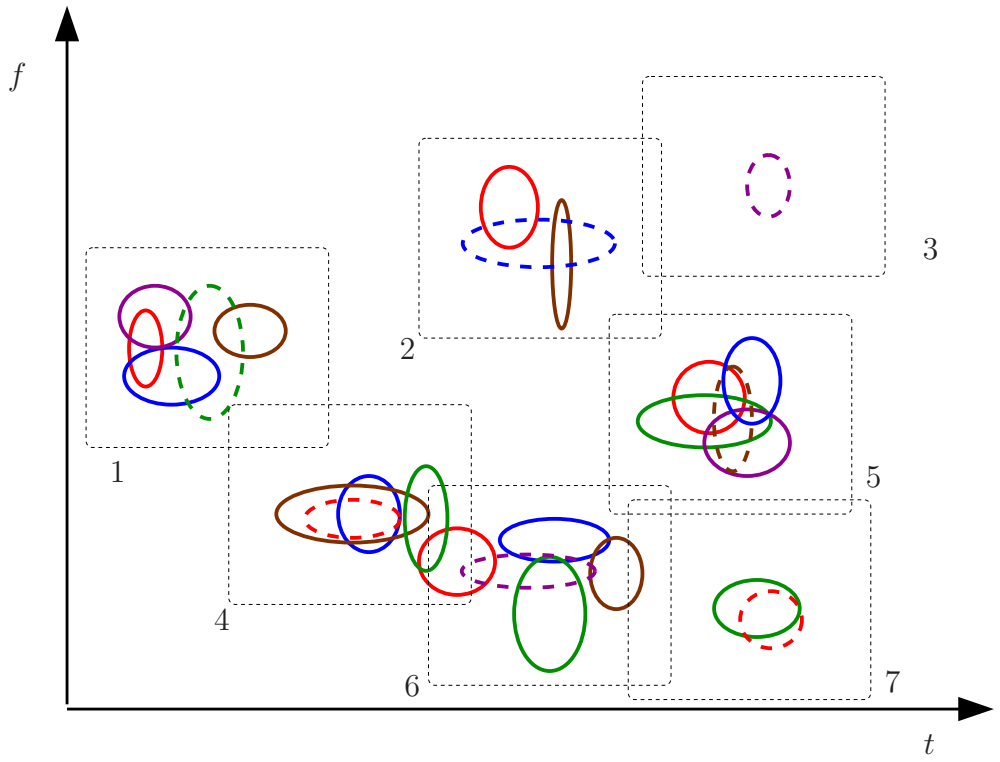


Fig. 1. Five bump models on top of each other ($N = 5$); the dashed boxes indicate clusters, the dashed ellipses correspond to exemplars.

eters: time t, frequency f, width Δt, height Δf, and amplitude w. The resulting bump models $e = ((t_1, f_1, \Delta t_1, \Delta f_1, w_1), \ldots, (t_n, f_n, \Delta t_n, \Delta f_n, w_n))$ and $e' = ((t'_1, f'_1, \Delta t'_1, \Delta f'_1, w'_1), \ldots, (t'_{n'}, f'_{n'}, \Delta t'_{n'}, \Delta f'_{n'}, w'_{n'}))$ represent the most prominent oscillatory activity in the signals at hand. This activity may correspond to various physical or biological phenomena, for example:

- oscillatory events in EEG and other brain signals are believed to occur when assemblies of neurons are spiking in synchrony (**?** 18),
- oscillatory events in calcium imaging data are due to oscillations of intracellular calcium, which are believed to play an important role in signal transduction between cells (see, e.g., (**?**)),
- oscillations and waves are of central interest in several fields beyond neuroscience, such as oceanography (e.g., oceanic "normal modes" caused by

convection (**?**)) and seismography (e.g., free earth oscillations and earth oscillations induced by earthquakes, hurricanes, and human activity (**?**)).

In the following, we will develop SES for bump models. In this setting, SES quantifies the synchronous interplay between oscillatory patterns in two given signals, while it ignores the other components in those signals ("background activity"). In contrast, classical synchrony measures such as amplitude or phase synchrony are computed from the entire signal, they make no distinction between oscillatory components and the background activity. As a consequence, SES captures alternative aspects of similarity, and hence, it provides complementary information about synchrony.

Besides bump models, SES may be applied to other sparse representations of signals, for example:

- matching pursuit (**?**) and refinements such as orthogonal matching pursuit (**?**), stage-wise orthogonal matching pursuit (**?**), tree matching pursuit (**?**) and chaining pursuit (**?**),
- chirplets (see, e.g., (**? ? ?**)),
- wave atoms (**?**),
- curvelets (**?**),
- sparsification by loopy belief propagation (**?**),
- the Hilbert-Huang transform (**?**),
- compressed sensing (**? ?**).

Moreover, the point processes may be defined in other spaces than the time-frequency plane, for example, they may occur in two-dimensional space (e.g., images), space-frequency (e.g., wavelet image coding) or space-time (e.g., movies);

they may also be defined on more complicated manifolds, such as curves, surfaces, etc. Such extensions may straightforwardly be derived from the example of bump models. We consider several extensions in Section **??**.

Our extension of stochastic event synchrony to multi-dimensional point processes (and bump models in particular) is derived from the following observation (see Fig. **??**): bumps in one time-frequency map may not be present in the other map ("non-coincident" bumps); other bumps are present in both maps ("coincident bumps"), but appear at slightly different positions on the maps. The black lines in Fig. **??** connect the centers of coincident bumps, and hence, visualize the offsets between pairs of coincident bumps.

Such offsets jeopardize the suitability of classical similarity measures for time-frequency maps. For example, let us consider the Pearson correlation coefficient r between two time-frequency maps $x_1(t,f)$ and $x_2(t,f)$:

$$r = \frac{\sum_{t,f}(x_1(t,f)-\bar{x}_1)(x_2(t,f)-\bar{x}_2)}{\sqrt{\sum_{t,f}(x_1(t,f)-\bar{x}_1)^2}\sqrt{\sum_{t,f}(x_2(t,f)-\bar{x}_2)^2}}, \tag{1}$$

where $\bar{x}_i = \sum_{t,f} x_i(t,f)$ $(i = 1,2)$. Note that r, like many other classical similarity measures, is based on pointwise comparisons, in other words, it compares the activity at instance (t,f) in map x_1 to the activity in x_2 at the *same* instance (t,f). Therefore, if the correlated activity in the maps $x_1(t,f)$ and $x_2(t,f)$ is slightly delayed or a little shifted in frequency, the correlation coefficient r will be small, and as a result, it may not be able to capture the correlated activity. Our approach alleviates this shortcoming, since it explicitly handles delays and frequency offsets.

We quantify the interdependence between two bump models by five parameters, i.e., the parameters ρ, δ_t, and s_t introduced in Part I:

Fig. 2. Generative model for e and e'. One first generates a hidden process v, next one makes two identical copies of v and shifts those over $(-\delta_t/2, -\delta_f/2)$ and $(\delta_t/2, \delta_f/2)$ respectively; the events of the resulting point process are slightly shifted (with variance (s_t,s_f)), and some of those events are deleted (with probability p_d), resulting in e and e'.

- ρ: fraction of non-coincident bumps,
- δ_t: the average timing offset (delay) between coincident bumps,
- s_t: the variance of the timing offset between coincident bumps,

in addition to:

- δ_f: the average frequency offset between coincident bumps,
- s_f: the variance of the frequency offset between coincident bumps.

We determine those 5 parameters and the pairwise alignment of e and e' by *statistical inference*, as in the one-dimensional case (cf. Section 3 and 4 in Part I). We start by constructing a statistical model that captures the relation between the two bump models e and e'; that model contains the 5 SES parameters, besides variables related to the pairwise alignment of the bumps of e and e'. Next we perform inference in that model, resulting in estimates for the SES parameters and the pairwise alignment. More concretely, we apply coordinate descent, as in the case of one-dimensional point processes. In the following section, we outline our statistical model. In Section **??**, we describe the factor graph of that model. From that factor graph, we derive the inference algorithm for multi-dimensional SES; in Section **??**, we outline that inference algorithm. We refer to Appendix **??** for the detailed derivations. In Section **??**, we suggest various extensions of our statistical model.

3 Statistical Model

Consider N signals $S_1, \ldots, S_N$ from which we extract point processes $X_1, \ldots, X_N$ by some appropriate method. Each point process X_i is a list of n_i points (later referred to as "events" or "bumps" of activity) in a given multi-dimensional set $\mathcal{S} \subseteq \mathbb{R}^M$, i.e., $X_i = \{X_{i,1}, X_{i,2}, \ldots, X_{i,n_i}\}$ with $X_{i,k} \in \mathcal{S}$ for $k = 1, \ldots, n_i$ and $i = 1 \ldots N$. As an example, consider the bump model (9) extracted from the time-frequency maps of EEG signals (see Fig. 3). The time-frequency ("wavelet") transform of each EEG signal is approximated as a sum of half-ellipsoid basis functions, referred to as "bumps" (9); each bump is described by five parameters: time T, frequency F, width ΔT, height ΔF, and amplitude W (as such, a bump is also a point in $\mathbb{R}^5$). We wish to quantify to which

extent the N resulting bump models $X_i = ((T_{i,1}, F_{i,1}, \Delta T_{i,1}, \Delta F_{i,1}, W_{i,1}), \ldots, (T_{i,n_i}, F_{i,n_i}, \Delta T_{i,n_i}, \Delta F_{i,n_i}, W_{i,n_i}))$ are similar.

Intuitively speaking, N signals X_i may be considered well-synchronized if bumps appear in all models (or almost all) simultaneously, potentially with some small offset in time and frequency. In other words, if one overlays N partially synchronous bump models (cf. Fig. 1 with $N = 5$), bumps naturally appear in clusters that contain precisely one bump from all (or almost all) bump models. In the example of Fig. 3, cluster 1, 5 and 6 contain bumps from all 5 models X_i, cluster 2, 4 and 7 contains bumps from 3, 4, and 2 models respectively, and cluster 3 consists of a single bump.

This intuitive concept of similarity may readily be translated into a generative stochastic model. In that model, the N point processes X_i are treated as independent noisy observations of a hidden "mother" process $\tilde{X}$. An observed sequence $(X_i)_{i=1,\ldots,N}$ is obtained from $\tilde{X}$ by the following three-step procedure:

(1) COPY: generate a copy of the mother bump model $\tilde{X}$,
(2) DELETION: delete some of the copied mother bumps,
(3) PERTURBATION: slightly alter the position and shape of the remaining mother bump copies, amounting to the bump model X_i.

As a result, each sequence X_i consists of "noisy" copies of a non-empty subset of mother bumps. The point processes X_i may be considered well-synchronized if there only few deletions (cf. Step 2) and if the bumps of X_i are "close" to the corresponding mother bumps (cf. Step 3). One way to determine the synchrony of given point processes X_i is to first reconstruct the hidden mother process $\tilde{X}$, and to next determine the number of deletions and the average distance

between the point processes X_i and the mother process $\tilde{X}$. Inferring the mother process is a high-dimensional estimation problem, the underlying probability distribution typically has a large number of local extrema. Therefore, we will use an alternative procedure: we will assume that each cluster contains one *identical* copy of a mother bump, the other bumps in that cluster are *noisy* copies of that mother bump. The identical copy, referred to as "exemplar", plays the role of "center" or "representative" of each cluster (see Fig. 1). We will assume, without loss of generality, that there is one exemplar for each mother bump. Note that under this assumption, the mother process $\tilde{X}$ is equal to the list of all exemplars. Some point processes X_i may additionally contain noisy copies of that mother bump, but this does not need to be the case, in other words, there might be clusters of size one, solely consisting of an exemplar (cf. cluster 3 in Fig. 1).

The exemplar-based formulation amounts to the following inference problem: given the point processes X_i, we need to identify the bumps that are exemplars and the ones that are noisy copies of some exemplar, with the constraints that an exemplar and its noisy copies all stem from different point processes. Obviously, this inference problem also has potentially many locally optimal solutions, however, in contrast to the original (continuous) inference problem, we can in practice find the global optimum by integer programming (see Section 4).

We now proceed from the example of bump models to general point processes X_i, and describe the underlying stochastic model in more detail. The mother process $\tilde{X} = \left\{\tilde{X}_1, \ldots, \tilde{X}_M\right\}$, which is the source of all points ("events") in $X_1, X_2, \ldots X_N$, is modeled as follows:

- The number M of points in $\tilde{X}$ is geometrically distributed with parameter $\lambda \operatorname{vol}(S)$:

$$p(M) = (1 - \lambda \operatorname{vol}(S))\Big(\lambda \operatorname{vol}(S)\Big)^M, \tag{2}$$

  where $\operatorname{vol}(S)$ is the multi-dimensional volume of set S.

- Each point $\tilde{x}_m$ for $m = 1, \ldots, M$ is uniformly distributed in S:

$$p(\tilde{x}|M) = \operatorname{vol}(S)^{-M}. \tag{3}$$

With those two choices, the prior of the mother process $\tilde{X}$ equals:

$$p(\tilde{x}, M) = p(M)p(\tilde{x}|M) = (1 - \lambda \operatorname{vol}(S))\lambda^M. \tag{4}$$

For convenience we will in the following use the short-hand notation $p(\tilde{x})$ for $p(\tilde{x}, M)$, i.e., we will not explicitly mention the dependency on M.

From the mother process $\tilde{X}$, the point processes X_i for $i = 1, \ldots, N$ are generated according to the following steps:

- For each event $\tilde{X}_m$ in the mother process $\tilde{X}$, one of the point process X_i with $i \in \{1, \ldots, N\}$ is chosen at random, denoted by $X_{i(m)}$, and a copy of mother event $\tilde{X}_m$ is created in $X_{i(m)}$; this identical copy is referred to as "exemplar". For convenience, we will adopt a uniform prior $p\Big(i(m) = i\Big) = 1/N$ for $i = 1, \ldots, N$ (the model can be easily generalized to any prior).

  Next for each event $\tilde{X}_m$ in the mother process $\tilde{X}$ (with $m = 1, \ldots, M$), a "noisy" copy may be created in the point processes X_j with $j \neq i(m)$ (at most one copy per point process X_j). More precisely, the noisy copies are modeled as follows:

- The number C_m of copies is modeled by a prior $p(c_m|\theta^c)$, parameterized by θ^c, which in turn has a prior $p(\theta^c)$. In this paper, we consider as prior for C_m: (i) a binomial distribution $\operatorname{Bi}(p_s)$ with $N - 1$ trials and probabil-

ity of success p_s; (ii) a multinomial distribution Mult(γ) with parameter γ. We adopt conjugate priors for the parameters p_s and γ, i.e., the beta distribution $\mathrm{B}(\kappa, \lambda)$ and Dirichlet distribution $\mathrm{Di}(\zeta)$ respectively.

Note that a binomial prior $\mathrm{Bi}(p_s)$ for C_m is equivalent to deleting copies of the mother events *independently* with probability $1 - p_s$ (cf. DELETION step). On the other hand, if the prior $p(c_m|\theta^c)$ is a multinomial distribution, the copies are in general no longer deleted independently.

- Conditional on the number C_m of copies, the copies are attributed uniformly at random to other signals X_j, with the constraints of at most one copy per signal and $j \neq i(m)$; since there are $\binom{N-1}{c_m}$ possible attributions $\mathcal{A}_m \subseteq \{1, \ldots, i(m)-1, i(m)+1, \ldots, N\}$ with $|\mathcal{A}_m| = c_m$, the probability mass of an attribution $\mathcal{A}_m$ is $p(\mathcal{A}_m|c_m) = \binom{N-1}{c_m}^{-1}$.
- The process of generating a noisy copy $X_{i,r}$ from a mother bump $\tilde{X}_m$ is described by a conditional distribution $p_x(x_{i,r}|\tilde{x}_m; \theta_i^x)$, parameterized by some vector θ_i^x that may differ for each point process X_i. For the sake of simplicity, the conditional distribution p_x is assumed to be identical for all mother bumps $\tilde{X}_m$ and noisy copies $X_{i,r}$. The vectors θ_i^x may be treated as (mutually independent) random vectors with non-trivial priors $p(\theta_i^x)$.

  In the case of bump models (cf. Fig. 1), a simple mechanism to generate copies is to slightly shift the mother bump center while the other mother bump parameters (width, height, and amplitude) are drawn from some prior distribution, independently for each copy; the latter four bump parameters could be taken into account in a less trivial way, but due to space constraints we omit such extensions here. The center offset may be modeled as a bivariate Gaussian random variable with mean vector $(\delta_{t,i}, \delta_{f,i})$ and diagonal non-isotropic covariance matrix $V_i = \mathrm{diag}(s_{t,i}, s_{f,i})$, and hence, $\theta_i^x = (\delta_{t,i}, \delta_{f,i}, s_{t,i}, s_{f,i})$. For simplicity, we will assume that $s_{t,i} = s_t$ and

$s_{f,i} = s_f$ for all i. We adopt the improper priors $p(\delta_{t,i}) = 1 = p(\delta_{f,i})$ for $\delta_{t,i}$ and $\delta_{f,i}$ respectively, and conjugate priors for s_t and s_f, i.e. scaled inverse chi-square distributions:

$$p(s_t) = \frac{(s_{0,t}\nu_t/2)^{\nu_t/2}}{\Gamma(\nu_t/2)} \frac{e^{-\nu_t s_{0,t}/2s_t}}{s_t^{1+\nu_t/2}} \tag{5}$$

$$p(s_f) = \frac{(s_{0,f}\nu_f/2)^{\nu_f/2}}{\Gamma(\nu_f/2)} \frac{e^{-\nu_f s_{0,f}/2s_f}}{s_f^{1+\nu_f/2}}, \tag{6}$$

where ν_t and ν_f are the degrees of freedom, and $s_{0,t}$ and $s_{0,f}$ are the width of the scaled inverse chi-square distributions, and $\Gamma(x)$ is the Gamma function. It is noteworthy that there might be a non-trivial timing and frequency offset between the bump models. The parameters $(\delta_{t,i}, \delta_{f,i})$ are introduced in the model to account for such offsets.

For later convenience, we will introduce some more notation. The exemplar associated to mother event $\tilde{X}_m$ is denoted by $X_{i(m),k(m)}$, it is the event $k(m)$ in point process $X_{i(m)}$. We denote the set of pairs $(i(m), k(m))$ by $\mathcal{I}^{\text{ex}}$. A noisy copy of $\tilde{X}_m$ is denoted by $X_{j(m),\ell(m)}$, it is the event $\ell(m)$ in point process $X_{j(m)}$ with $j(m) \in \mathcal{A}_m$. We denote the set of all pairs $(j(m), \ell(m))$ associated to $\tilde{X}_m$ by $\mathcal{I}_m^{\text{copy}}$, and furthermore define $\mathcal{I}^{\text{copy}} \triangleq \mathcal{I}_1^{\text{copy}} \cup \cdots \cup \mathcal{I}_M^{\text{copy}}$ and $\mathcal{I} = \mathcal{I}^{\text{ex}} \cup \mathcal{I}^{\text{copy}}$. In this notation, the overall probabilistic model may be written as:

$$\begin{aligned} p(\tilde{X}, X, \mathcal{I}, \theta) = {} & p(\theta^c)p(\theta^x)(1 - \lambda \operatorname{vol}(S))\lambda^M N^{-M} \\ & \prod_{m=1}^{M} \delta\Big(x_{i(m),k(m)} - \tilde{x}_m\Big) p(c_m|\theta^c) \binom{N-1}{c_m}^{-1} \\ & \prod_{(i,j)\in\mathcal{I}_m^{\text{copy}}} p_x(x_{i,j}|\tilde{x}_m, \theta^x). \end{aligned} \tag{7}$$

If the point processes $X = (X_1, \ldots, X_N)$ are well-synchronized, almost all processes X_i contain a copy of each mother bump $\tilde{X}_m$, and therefore, the sets $\mathcal{I}_m^{\text{copy}}$ are either of size $N-1$ or are slightly smaller. Moreover, in the case of

bump models, the variances s_t and s_f are then small. Therefore, given point processes $X = (X_1, \ldots, X_N)$, we wish to infer $\mathcal{I}$ and θ, since those variables contain information about similarity.

We gain additional insight into this inference problem by considering the logarithm of the above stochastic model:

$$\begin{aligned}
&-\log p(\tilde{X}, X, \mathcal{I}, \theta) = \\
&-\log p(\theta^c) - \log p(\theta^x) - \log(1 - \lambda \operatorname{vol}(S)) - M \log \frac{\lambda}{N} \\
&- \sum_{m=1}^{M} \log \delta\Big(x_{i(m),k(m)} - \tilde{x}_m\Big) - \log \left(p(c_m|\theta^c) \binom{N-1}{c_m}^{-1}\right) \\
&- \sum_{(i,j) \in \log \mathcal{I}_m^{\text{copy}}} \log p_x(x_{i,j}|\tilde{x}_m, \theta^x). \qquad (8)
\end{aligned}$$

The term $-\log p_x(x_{i,j}|\tilde{x}_m, \theta^x)$ may be interpreted as a measure for the distance between $x_{i,j}$ and $\tilde{x}_m$; note that this measure is not necessarily symmetric or non-negative. If p_x is a Gaussian distribution (as in the case of bump models), this measure is an Euclidean distance. In other applications, non-Euclidean distances may be more appropriate. The proposed algorithm can straightforwardly handle arbitrary distance measures.

Let us now consider specific choices for $p(c_m|\theta^c)$; if the latter is a binomial distribution with $N-1$ trials and probability of success p_s, and the prior for

p_s is a beta distribution $\mathrm{B}(\kappa, \lambda)$, we have:

$$
\begin{aligned}
&-\log p(\tilde{X}, X, \mathcal{I}, \theta) = \\
&-\log \mathrm{B}(p_s; \kappa, \lambda) - \log p(\theta^x) - \log(1 - \lambda \operatorname{vol}(S)) \\
&- M \log \frac{\lambda}{N} - \sum_{m=1}^{M} \log \delta\Big(x_{i(m),k(m)} - \tilde{x}_m\Big) \\
&- M(N-1) \log \delta - \sum_{m=1}^{M} (N - 1 - c_m) \log \frac{1 - p_s}{p_s} \\
&- \sum_{(i,j) \in \mathcal{I}_m^{\mathrm{copy}}} \log p_x(x_{i,j} | \tilde{x}_m, \theta^x),
\end{aligned} \tag{9}
$$

which can be rewritten as:

$$
\begin{aligned}
&-\log p(\tilde{X}, X, \mathcal{I}, \theta) = \\
&-\log \mathrm{B}(p_s; \kappa, \lambda) - \log p(\theta^x) - \log(1 - \lambda \operatorname{vol}(S)) + \alpha M \\
&- \sum_{m=1}^{M} \log \delta\Big(x_{i(m),k(m)} - \tilde{x}_m\Big) + \beta \sum_{m=1}^{M} (N - 1 - c_m) \\
&- \sum_{(i,j) \in \mathcal{I}_m^{\mathrm{copy}}} \log p_x(x_{i,j} | \tilde{x}_m, \theta^x),
\end{aligned} \tag{10}
$$

where

$$
\begin{aligned}
\alpha &= -\log \frac{\lambda}{N} - (N-1) \log p_s \\
&\quad \text{and} \\
\beta &= \log\left(\frac{p_s}{1 - p_s}\right)
\end{aligned} \tag{11}
$$

If $p(c_m | \theta^c)$ is a multinomial distribution $\mathrm{Mult}(\gamma)$ with parameter γ, and the

prior for γ is a Dirichlet distribution $\mathrm{Di}(\zeta)$, the expression (8) becomes:

$$
\begin{aligned}
&-\log p(\tilde{X}, X, \mathcal{I}, \theta) = \\
&-\log \mathrm{Di}(\gamma; \zeta) - \log p(\theta^x) - \log(1 - \lambda \,\mathrm{vol}(S)) + \phi M \\
&-\sum_{m=1}^{M} \log \delta\Big(x_{i(m),k(m)} - \tilde{x}_m\Big) + g(c_m) \\
&-\sum_{(i,j)\in \log \mathcal{I}_m^{\mathrm{copy}}} \log p_x(x_{i,j}|\tilde{x}_m, \theta^x). \qquad (12)
\end{aligned}
$$

where $\phi = -\log \frac{\lambda}{N}$, and the non-linear function g is defined as:

$$g(c_m) = -\log \gamma_m + \log \binom{N-1}{c_m}. \qquad (13)$$

4 Statistical Inference

A reasonable approach to infer $(\mathcal{I}, \theta)$ is maximum a posteriori (MAP) estimation:

$$(\hat{\mathcal{I}}, \hat{\theta}) = \underset{(\mathcal{I},\theta)}{\mathrm{argmax}} \log p(\tilde{X}, X, \mathcal{I}, \theta). \qquad (14)$$

There is no closed form expression for (14), therefore, we need to resort to numerical methods. A simple technique to try to find (14) is cyclic maximization: We first choose initial values $\hat{\theta}^{(0)}$, and then perform the following updates for $r \geq 1$ until convergence:

$$\hat{\mathcal{I}}^{(r)} = \underset{\mathcal{I}}{\mathrm{argmax}} \log p(\tilde{X}, X, \mathcal{I}, \hat{\theta}^{(r-1)}) \qquad (15)$$

$$\hat{\theta}^{(r)} = \underset{\theta}{\mathrm{argmax}} \log p(\tilde{X}, X, \hat{\mathcal{I}}^{(r)}, \theta). \qquad (16)$$

First we consider the update (15), which we will carry out by integer programming. Next we treat the update (16) of the parameters θ.

4.1 Integer Program

We write the update (15) as an integer program, i.e., a discrete optimization problem with linear objective function and linear (equality and inequality) constraints. To this end, we introduce the following variables:

- $B_{i,k}$ is a binary variable equal to one iff the k-th event of X_i is an exemplar.
- $B_{i,k,i',k'}$ is a binary variable equal to one iff the k-th event of X_i is copy of exemplar $X_{i',k'}$.
- $B_{i,i',k'}$ is a binary variable equal to one iff no event of X_i is a copy of exemplar $X_{i',k'}$.

Note that $b_{i,k,i,k'} = 0$ for all k and k' and $b_{i,i,k'} = 1$ for all i and k', since X_i must not contain a noisy copy of a mother event $\tilde{X}_m$ if it already contains the exemplar associated to $\tilde{X}_m$.

We will first consider a binomial prior for the number of copies C_m, which directly leads to an integer program. Next we consider a multinomial prior for C_m, which results in a non-linear objective function. By introducing auxiliary variables, this objective function can be written as a linear function in the resulting augmented parameter space, and the associated combinatorial optimization problem can be formulated as an integer program, as we will briefly outline in Section 4.1.2.

4.1.1 Binomial prior

We first assume that the parameters θ^x and p_s of the binomial prior are constant. By substituting (10) in (15), it can be easily shown that with the above

choice of variables B, the conditional maximization (15) may be cast as the following integer program in B:

$$\min_{b} \quad C + \hat{\alpha}^{(r-1)} \sum_{i,1\leq k\leq n_i} b_{i,k} + \hat{\beta}^{(r-1)} \sum_{i,i'\neq i,1\leq k'\leq n_{i'}} b_{i,i',k'} - \sum_{i,i',1\leq k\leq n_i,1\leq k'\leq n_{i'}} b_{i,k,i',k'} \log p_x(x_{i,k}|x_{i',k'};\hat{\theta}^{(r-1)}) \tag{17}$$

subject to

$$\forall i,k, \qquad \sum_{i',k'} b_{i,k,i',k'} + b_{i,k} = 1 \tag{18}$$

$$\forall i,i'\neq i,k', b_{i,i',k'} = b_{i',k'} - \sum_{1\leq k\leq n_i} b_{i,k,i',k'}, \tag{19}$$

where C is a constant, and

$$\begin{aligned} \hat{\alpha}^{(r-1)} &= -\log\frac{\lambda}{N} - (N-1)\log\hat{p}_s^{(r-1)} \\ &\text{and} \\ \hat{\beta}^{(r-1)} &= \log\Big(\frac{\hat{p}_s^{(r-1)}}{1-\hat{p}_s^{(r-1)}}\Big). \end{aligned} \tag{20}$$

The sum $\sum_{i,k} b_{i,k}$ in (17) is equal to the number of exemplars M; therefore, the first term in (17) assigns a cost α to each exemplar. The second term in (17) associates a cost β to every deletion. Indeed, if (i',k') is not an exemplar, $\sum_i b_{i,i',k'}$ is equal to zero; if (i',k') is the exemplar associated to the m-th mother event, $\sum_i b_{i,i',k'} = (N-1-c_m)$, which is the number of deletions in the m-th cluster. The third term assigns a cost to each copy (i,k) of exemplar (i',k'), proportional to the "distance" $-\log p_x$ between both events.

The constraint (18) ensures that each event is either an exemplar or a copy of an exemplar. The constraint (19), combined with the fact that $B_{i,i',k'}$ is a binary variable, encodes the following:

- $B_{i,k,i',k'}$ can only be equal to one if $B_{i',k'}$ is equal to one, i.e. (i,k) can be a

copy of (i', k') iff (i', k') is an exemplar,

- at most one event in X_i can be a copy of (i', k'),
- $B_{i,i',k'}$ is one iff (i', k') is an exemplar but has no copy in X_i.

The discrete optimization problem (17)-(19) is an integer program in B, since the objective function (17) and constraints (18) (19) are linear in the variables B.

4.1.2 Multinomial prior

First we assume that the parameters θ^x and γ of the multinomial prior are constant. By substituting (12) in (15), the conditional maximization (15) results in the following combinatorial optimization problem:

$$\begin{aligned}\min_b \ & \tilde{C} + \phi \sum_{i,1\le k\le n_i} b_{i,k} \\ & + \sum_{i',1\le k'\le n_{i'}} b_{i',k'} \ \hat{g}^{(r-1)}\Big(N-1-\sum_{i\neq i'} b_{i,i',k'}\Big) \\ & - \sum_{i,i',1\le k\le n_i,1\le k'\le n_{i'}} b_{i,k,i',k'} \ \log p_x(x_{i,k}|x_{i',k'};\hat{\theta}^{(r-1)}),\end{aligned} \tag{21}$$

subject to the constraints (18) (19), where $\tilde{C}$ is an arbitrary constant and and the non-linear function $g^{(r-1)}$ is defined as:

$$\hat{g}^{(r-1)}(c) = -\log \hat{\gamma}_c^{(r-1)} + \log \binom{N-1}{c}, \tag{22}$$

for $c = 0, 1, \ldots, N-1$. Note that the objective function (21) is non-linear in B since it involves the non-linear function g. We will now introduce auxiliary variables such that the objective function (21) is linear in those variables; we will then reformulate (21) as an integer program in the augmented space of variables.

Let us first point out that for an arbitrary function f we can always write:

$$f(x) = \sum_{x' \in \mathcal{X}} f(x')\delta[x - x'], \tag{23}$$

with discrete (finite or infinite) set $\mathcal{X}$. By introducing variables $D_{x'}$, we can rewrite (23) as:

$$f(x) = \sum_{x' \in \mathcal{X}} f(x') d_{x'}, \tag{24}$$

with the constraint $d_{x'} = \delta[x - x']$. The key observation here is that (24) is linear in $D_{x'}$.

In this vein, we introduce the binary variables $D_{v,i',k'}$ and rewrite the objective function (21) as:

$$\begin{aligned} \min_{b} \quad & \phi \sum_{i, 1 \le k \le n_i} b_{i,k} + \sum_{v, i', 1 \le k' \le n_{i'}} g_v^{(r-1)}\, d_{v,i',k'} \\ & - \sum_{i, i', 1 \le k \le n_i, 1 \le k' \le n_{i'}} b_{i,k,i',k'} \log p_x(x_{i,k} | x_{i',k'}; \theta) + \tilde{C}, \end{aligned} \tag{25}$$

where $g_v^{(r-1)} = g^{(r-1)}\Big(N - 1 - v\Big)$. This alternative formulation is equivalent to the original expression (21) iff $D_{v,i',k'}$ equals one if both $v = \sum_{i \neq i'} b_{i,i',k'}$ and $b_{i',k'} = 1$, and is zero otherwise. We express those constraints on $D_{v,i',k'}$ as follows:

$$v - \sum_{i \neq i'} b_{i,i',k'} \le a_{v,i',k'}, \tag{26}$$

$$\sum_{i \neq i'} b_{i,i',k'} - v \le a_{v,i',k'}, \tag{27}$$

$$a_{v,i',k'} \le N\,(1 - d_{v,i',k'}), \tag{28}$$

$$\sum_{v} d_{v,i',k'} = b_{i',k'}, \tag{29}$$

where $A_{v,i',k'}$ are additional auxiliary binary variables. The first two constraints encode that $a_{v,i',k'} \ge |v - \sum_{i \neq i'} b_{i,i',k'}|$. If $v \neq \sum_{i \neq i'} b_{i,i',k'}$, the variable $A_{v,i',k'}$ is strictly positive, and from the third inequality it follows that $D_{v,i',k'}$ equals

zero. On the other hand, if $v = \sum_{i \neq i'} b_{i,i',k'}$, the first two constraints no longer force $A_{v,i',k'}$ to be non-zero, and they do not impose any constraint on $D_{v,i',k'}$. However, from the fourth constraint it follows that if $b_{i',k'} = 1$ and hence if (i', k') is an exemplar, one of the $D_{v,i',k'}$ (with fixed i' and k') is equal to one. By setting $D_{v,i',k'}$ equal to one if $v = \sum_{i \neq i'} b_{i,i',k'}$ and zero otherwise, one fulfills then all four constraints. If $b_{i',k'} = 0$ and hence if (i', k') is not an exemplar, all $D_{v,i',k'}$ (with fixed i' and k') are equal to zero. By setting all $D_{v,i',k'}$ equal to zero, one in that case fulfills all four constraints.

In summary: the non-linear combinatorial optimization problem with objective (21) and constraints (18) (19) is equivalent to the integer program with objective (25) and constraints (18) (19) combined with (26)–(29).

4.1.3 Complexity

In the case $N = 2$, it can be seen that the combinatorial optimization problem (15) can be reduced to a *bipartite maximum weighted matching* optimization problem, which can be solved in polynomial time through several methods: linear programming relaxation, Edmond-Karp algorithm, or max-product message-passing algorithm detailed in (5). For $N > 2$, (15) is very similar to solving a *maximum weighted N-dimensional matching*. For the purpose of understanding the combinatorial hardness of the problem, we show that for $N \geq 5$, the *maximum 3-dimensional matching* problem can be reduced to (15) when forgoing the euclidean costs assumptions. Since *maximum 3-dimensional matching* is NP-hard, it results that (15) (with general costs) is also NP-hard. The problem is actually NP-hard for $N \geq 3$, but the proof is more involved and beyond the scope of this paper. Therefore, the exten-

sion from 2 time series to more than 2 is far from trivial. Nevertheless, as we will discuss later, we were able to solve the problem (15) for our purpose in reasonable time using integer programming techniques.

Proposition 1 *The combinatorial problem (15) is NP-hard if* $N \geq 5$.

We include a sketch of the proof ; it is based on a reduction from *maximum weighted 3-dimensional matching* optimization, which is known to be NP-hard and APX-hard (20) (21).

Let $T \subset X \times Y \times Z$, where X, Y, Z are disjoint sets. Consider the following "time-series" X', Y', Z', T', U'. For every $x \in X$ (resp. $y \in Y$, $z \in Z$), create two corresponding bumps $x \in X'$ and $\tilde{x} \in U'$ (resp. $y \in Y'$, $z \in Z'$, $\tilde{y} \in U$, $\tilde{z} \in U'$) and for every $t = (x, y, z) \in T$, create two bumps $t \in T'$ and $\tilde{t} \in U'$. Set the cost function as follows:

- $p_s = 1 - \epsilon$, where ϵ is an extremely small positive constant (practically 0)
- $\lambda = N \exp(1)$
- For any $t = (x, y, z) \in T$, let $s_{x,t} = s_{y,t} = s_{z,t} = 0$. For any bump $b \in X' \cup Y' \cup Z' \cup T'$, let $s_{b,\tilde{b}} = \beta$. For any other two bumps b_1, b_2, let s_{b_1,b_2} be equal to M, where M is a very large positive constant (pratically, $+\infty$).

The first two assumptions effectively set α to -1 and β to a very large number. Note the following: all bumps in U' have to be exemplars (because for any bump $u \in U'$ and any other bump b, $s_{u,b}$ is infinite). The total cost of bumps in U' being exemplars is therefore an additive constant which does not change the solution. Moreover, for any bump $u \in U'$, there exists a unique bump $b \in X' \cup Y' \cup Z' \cup T'$ that can be assigned to it (i.e. for any $b' \neq b$, $s_{b',u} = +\infty$). If this bump b is assigned u, the assignment cost $s_{b,u}$ is β, and since 4 bumps

are missing in the cluster, the cost of missing bumps is 4β. The total cluster cost would therefore be 5β. If the bump b is not assigned to u, 5 bumps are missing in the cluster and the total cluster cost is again 5β. Therefore, bumps $b \in X' \cup Y' \cup Z' \cup T'$ can be assigned to their corresponding exemplar in U' without changing the total cost. For this reason, exemplars in U' can be considered as "fake exemplars" (they serve as bins for unmatched bumps in X', Y', Z' and T'). The next step consists in observing all other exemplars have to be in T'. Indeed, for any bump $b_1 \in X' \cup Y' \cup Z'$, and any other bump b_2, s_{b_2,b_1} is infinite. Moreover, since $\alpha = -1$, the optimization effectively aims at maximizing the number of exemplars in T'. Since the cost of missing bumps β is very large, all "real" clusters (with exemplars in T') have to contain a bump from each time serie X', Y', Z'. Let $t = (x, y, z) \in T$. Then the only possible cluster for exemplar $t \in T'$ consists of the corresponding bumps x, y, z in X', Y', Z' (all other assignements bring the cost up to infinity). Finally, since each bump in X', Y', Z' can only be assigned to one exemplar in T', the clusters differ in each coordinate. It finally follows that the set of real clusters is the maximum 3-dimensional matching of $T \subset X \times Y \times Z$.

In practice (see the applications of Section 6.1), we were often able to solve the corresponding integer program very efficiently: for integer programs with more than 10'000 variables and 5'000 constraints, the solution of a given was obtained in less than 1 second on a fast processor (3GHz). The total running time of the algorithm (iterations of equations (14) and (15) until convergence) was under 7 seconds on average. This is perhaps surprising, especially in the light of the fact that message-passing algorithms did relatively poorly on the problem (slow convergence, relatively weak solutions). We believe that the good performance of the IP stems from the relatively good performance of the

LP relaxation. Typically, in our instances, the LP relaxation approximated the optimal cost within 3%; and a third of the positive components of the optimal LP solution were integer.

4.2 Parameter Estimation

We now consider the update (16), i.e., estimation of the parameters $\theta = (\theta^x, \theta^c)$. The estimate $\hat{\theta}^{(r+1)} = (\hat{\theta}^{x^{(r+1)}}, \hat{\theta}^{c^{(r+1)}})$ (16) is often available in closed-form. This is in particular the case for the parametrization $\theta_i^x = (\delta_{t,i}, \delta_{f,i}, s_t, s_f)$. The point estimates $\hat{\delta}_{t,i}^{(r+1)}$ and $\hat{\delta}_{f,i}^{(r+1)}$ are the (sample) mean of the timing and frequency offset respectively, computed between all noisy copies in X_i and their associated exemplars:

$$\hat{\delta}_{t,i}^{(r)} = \frac{1}{n_i^{(r)}} \sum_{k,i',k'} \hat{b}_{i,k,i',k'}^{(r)} \left(T_{i,k} - T_{i',k'}\right) \tag{30}$$

$$\hat{\delta}_{f,i}^{(r)} = \frac{1}{n_i^{(r)}} \sum_{k,i',k'} \hat{b}_{i,k,i',k'}^{(r)} \left(F_{i,k} - F_{i',k'}\right), \tag{31}$$

where $n_i^{(r)}$ is the number of noisy copies in X_i:

$$n_i^{(r)} = \sum_{k,i',k'} \hat{b}_{i,k,i',k'}^{(r)} = n_i - \sum_k \hat{b}_{i,k}^{(r)}. \tag{32}$$

The estimates $\hat{s}_t^{(r)}$ and $\hat{s}_f^{(r)}$ are obtained as:

$$\hat{s}_t^{(r)} = \frac{\nu_t s_{0,t} + n^{(r)} \hat{s}_{t,\text{sample}}^{(r)}}{\nu_t + n^{(r)} + 2} \tag{33}$$

$$\hat{s}_f^{(r)} = \frac{\nu_f s_{0,f} + n^{(r)} \hat{s}_{f,\text{sample}}^{(r)}}{\nu_f + n^{(r)} + 2}, \tag{34}$$

where $s^{(r)}_{t,\text{sample}}$ and $s^{(r)}_{f,\text{sample}}$ are computed over all exemplars and their noisy copies:

$$\hat{s}^{(r)}_{t,\text{sample}} = \frac{1}{n^{(r)}} \sum_{i,k,i',k'} \hat{b}^{(r)}_{i,k,i',k'} \, (T_{i,k} - T_{i',k'})^2, \tag{35}$$

$$\hat{s}^{(r)}_{f,\text{sample}} = \frac{1}{n^{(r)}} \sum_{i,k,i',k'} \hat{b}^{(r)}_{i,k,i',k'} \, (F_{i,k} - F_{i',k'})^2, \tag{36}$$

and $n^{(r)}$ is the total number of noisy copies:

$$n^{(r)} = \sum_{i,k,i',k'} \hat{b}^{(r)}_{i,k,i',k'} = \sum_i n_i - \sum_{i,k} \hat{b}^{(r)}_{i,k} = \sum_i n_i - \hat{M}^{(r)}. \tag{37}$$

The parameter p_s of the binomial prior for the number of copies C_m is estimated as:

$$\hat{p}^{(r)}_s = \frac{\kappa + \sum_i n_i - \hat{M}^{(r)} - 1}{\kappa + \lambda + \hat{M}^{(r)} - 2}. \tag{38}$$

The parameter γ of the multinomial prior for number of copies C_m is estimated as:

$$\hat{\gamma}^{(r)}_j = \frac{\zeta_i - 1 + \sum_{i',k'} \hat{b}^{(r)}_{i',k'} \; \delta\Big[\sum_{i,k} \hat{b}^{(r)}_{i,k,i',k'} - j\Big]}{\sum_i \zeta_i - N + \sum_{i,k} \hat{b}^{(r)}_{i,k}}, \tag{39}$$

for $j = 0, 1, \ldots, N-1$.

5 Analysis of Surrogate Data

As in the one-dimensional case (Part I, Section 6), we investigate the robustness and reliability of multi-dimensional SES by means of surrogate data. We randomly generated 1'000 pairs of two-dimensional point processes (e, e') according to the symmetric procedure depicted in Fig. 2.

We considered several values of the parameters ℓ, p_d, δ_t, δ_f, s_t (σ_t) and s_f (σ_f). More specifically, the length ℓ was chosen as $\ell = \ell_0/(1-p_d)$, where $\ell_0 \in \mathbb{N}_0$ is a

constant. With this choice, the expected length of e and e' is ℓ_0, independently of p_d. We considered the values $\ell_0 = 40$ and 100, $p_d = 0$, 0.1, ..., 0.4, $\delta_t =$ 0ms, 25ms, 50ms, $\sigma_t =$ 10ms, 30ms, and 50ms, $\delta_f =$ 0Hz, 2.5Hz, 5Hz, $\sigma_f =$ 1Hz, 2.5Hz, and 5Hz, $t_{\min} =$ 0s, $f_{\min} =$ 0Hz, $t_{\max} = \ell_0 \cdot 100$ms and $f_{\max} = \ell_0 \cdot 1$Hz. With this choice, the average event occurrence rate is about 10Hz, for all ℓ_0 and p_d. The width Δt_k and height Δf_k of all bumps is set equal to 0.5, so that $(\Delta t_k + \Delta t'_{k'}) = 1 = (\Delta f_k + \Delta f'_{k'})$ for all k and k', and hence $\bar{\delta}_t = \delta_t$, $\bar{\delta}_f = \delta_f$, $\bar{s}_t = s_t$, and $\bar{s}_f = s_f$ (cf. (**??**), (**??**), (**??**), (**??**), and Table **??**).

We used the initial values $\hat{\delta}_t^{(0)} = 0$, 30, and 70ms, $\hat{\delta}_f^{(0)} =$ 0Hz, $\hat{s}_t^{(0)} = (30\text{ms})^2$, and $\hat{s}_f^{(0)} = (3\text{Hz})^2$. The parameter β was identical for all parameter settings, i.e., $\beta = 0.005$; it was optimized to yield the best overall results. We used an uninformative prior for δ_t, δ_f, s_t, and s_f,, i.e., $p(\delta_t) = p(\delta_f) = p(s_t) = p(s_f)$ $= 1$.

In order to assess the SES measures $S = s_t, \rho$, we compute for each above mentioned parameter setting the expectation $\mathrm{E}[S]$ and normalized standard deviation $\overline{\sigma}[S] = \sigma[S]/\mathrm{E}[S]$. Those statistics are computed by averaging over 1'000 pairs of point processes (e,e'), randomly generated according to the symmetric procedure depicted in Fig. 2.

The results are summarized in Fig. **??** to **??**. From those figures we can make the following observations:

- The estimates of s_t and p_d are slightly biased, especially for small ℓ_0, i.e., $\ell_0 = 40$, $s_t \geq (30\text{ms})^2$ and $p_d > 0.2$; more specifically, the expected value of those estimates is slightly smaller than the true value, which is due to ambiguity inherent in event synchrony (cf. Fig. **??**). However, the bias is significantly smaller than in the one-dimensional case (cf. Part I, Section 6);

the bias increases with s_f, which is in agreement with our expectations: the more frequency jitter, the more likely that some events are reversed in frequency, and hence are aligned incorrectly.

- As in the one-dimensional case, the estimates of δ_t are unbiased for all considered values of δ_t, δ_f, s_t, s_f, and p_d, likewise the estimates of δ_f (not shown here).
- The estimates of s_t do only weakly depend on p_d, and vice versa.
- The estimates of s_t and p_d do not depend on δ_t and δ_f, i.e., they are robust to lags δ_t and frequency offsets δ_f, since the latter can be estimated reliably.
- The normalized standard deviation of the estimates of δ_t, s_t and p_d grows with s_t and p_d, but it remains below 30%. Those estimates are therefore reliable.
- The expected value of s_t and p_d does hardly depend on the length ℓ_0. On the other hand, the estimates of s_t and p_d are less biased for larger ℓ_0. The normalized standard deviation of the SES parameters decreases as the length ℓ_0 increases, as expected.

In summary, by means of the SES inference method, one may reliably and robustly determine the timing dispersion s_t and event reliability ρ of pairs of multi-dimensional point processes. We wish to reiterate, however, that it slightly underestimates the timing dispersion and the number of event deletions due to the ambiguity inherent in event synchrony (cf. Fig. **??**). Moreover, similarly as in the one-dimensional case, it is critical to choose an appropriate set of initial values $\hat{\delta}_t^{(0)}$, $\hat{\delta}_f^{(0)}$, $\hat{s}_t^{(0)}$, and $\hat{s}_f^{(0)}$.

6 Application: Diagnosis of MCI from EEG

Several clinical studies have shown that the EEG of Alzheimer's disease (AD) patients is generally less coherent than of age-matched control subjects; this is also the case for patients suffering from mild cognitive impairment (see (**?**) for a review). In this section, we apply SES to detect subtle perturbations in EEG synchrony of MCI patients.

First we describe the EEG data at hand (Section 6.1), then we describe how we preprocess the EEG, extract bump models, and apply SES (Section 6.2); at last, we present our results (Section **??**).

6.1 EEG Data

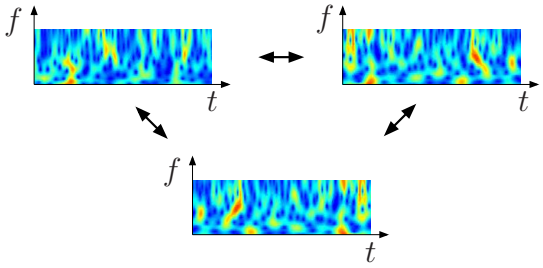
The EEG data used here have been analyzed in previous studies concerning early diagnosis of Alzheimer's disease (AD) (**? ? ? ? ?**).

Ag/AgCl electrodes (disks of diameter 8mm) were placed on 21 sites according to 10-20 international system, with the reference electrode on the right earlobe. EEG was recorded with Biotop 6R12 (NEC San-ei, Tokyo, Japan) using analog bandpass filtering in the frequency range 0.5-250Hz at a sampling rate of 200Hz. As in (**? ? ? ? ?**), the signals were then digitally band pass filtered between 4 and 30Hz using a third-order Butterworth filter.

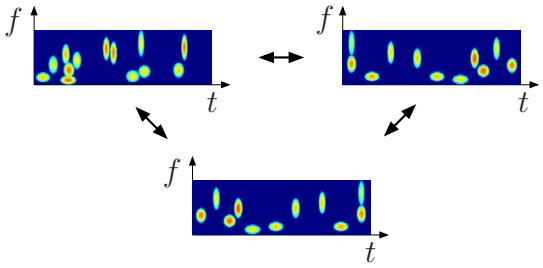
The subjects comprised two study groups. The first consisted of a group of 25 patients who had complained of memory problems. These subjects were then diagnosed as suffering from mild cognitive impairment (MCI) and subsequently developed mild AD. The criteria for inclusion into the MCI group

were a mini mental state exam (MMSE) score = 24 (max score = 30), though the average score in the MCI group was 26 (SD of 1.8). The other group was a control set consisting of 56 age-matched, healthy subjects who had no memory or other cognitive impairments. The average MMSE of this control group was 28.5 (SD of 1.6). The ages of the two groups were 71.9 ± 10.2 and 71.7 ± 8.3, respectively. Finally, it should be noted that the MMSE scores of the MCI subjects studied here are quite high compared to a number of other studies. For example, in (**?**) the inclusion criterion was MMSE = 20, with a mean value of 23.7, while in (**?**), the criterion was MMSE = 22; the mean value was not provided. The disparity in cognitive ability between the MCI and control subjects was thus comparatively small, making the present classification task relatively difficult.

All recording sessions were conducted with the subjects in an awake but resting state with eyes closed; the EEG technicians prevented the subjects from falling asleep (vigilance control). After recording, the EEG data has been carefully inspected. Indeed, EEG recordings are prone to a variety of artifacts, for example due to electronic smog, head movements, and muscular activity. The EEG data has been investigated by an EEG expert, blinded from the results of this analysis. In particular, only those subjects were retained in the analysis whose EEG recordings contained at least 20s of artifact-free data. Based on this requirement, the number of subjects in the two groups described above was further reduced to 22 and 38, respectively. From each subject, one EEG segment of 20s was analyzed (for each of the 21 channels).



(a) Time-frequency maps.



(b) Bump models.

Fig. 3. Similarity of three EEG signals ($N = 3$); from their time-frequency transforms (top), one extracts two-dimensional point processes ("bump models"; bottom), which are then aligned by the proposed algorithm.

6.2 Methods

We successively apply the following transformations to the EEG signals:

(1) wavelet transform,
(2) normalization of the wavelet coefficients,
(3) bump modeling of the normalized wavelet representation,
(4) aggregation of the resulting bump models in several regions.

Eventually, we compute the SES parameters for each pair of aggregated bump models. In the following, we detail each of those five operations.

6.2.1 Wavelet Transform

In order to extract the oscillatory patterns in the EEG, we apply a wavelet transform. More specifically, we use the complex Morlet wavelets (**? ?**):

$$\psi(t) = A \exp\left(-t^2/2\sigma_0^2\right) \exp(2i\pi f_0 t), \tag{40}$$

where t is time, f_0 is frequency, σ_0 is a (positive) real parameter, and A is a (positive) normalization factor. The Morlet wavelet (40) has proven to be well suited for the time-frequency analysis of EEG (see (**? ?**)). The product $w_0 = 2\pi f_0 \cdot \sigma_0$ determines the number of periods in the wavelet ("wavenumber"). This number should be sufficiently large (≥ 5), otherwise the wavelet $\psi(t)$ does not fulfill the admissibility condition:

$$\int \frac{|\psi(t)|^2}{t} dt < \infty, \tag{41}$$

and as a result, the temporal localization of the wavelet becomes unsatisfactory (**? ?**). In the present study, we choose a wavenumber $w_0 = 7$, as in the earlier studies (**?** 9); this choice yields good temporal resolution in the frequency range we consider in this study.

The wavelet transform $x(t, s)$ of an EEG signal $x(t)$ is obtained as:

$$x(t, s) \triangleq \sum_{t'=1}^{K} x(t')\, \psi^*\left(\frac{t' - t}{s}\right), \tag{42}$$

where $\psi(t)$ is the Morlet "mother" wavelet (40), s is a scaling factor, and $K = f_s T$, with f_s the sampling frequency and T the length of the signal. For the EEG data at hand, we have $T = 20s$ and $f_s = 200$Hz and hence $K = 4000$. The scaled and shifted "daughter" wavelet in (42) has center frequency $f \triangleq f_0/s$. In the following, we will use the notation $x(t, f)$ instead of $x(t, s)$.

Next we compute the squared magnitude $s(t, f)$ of the coefficients $x(t, f)$:

$$s(t, f) \triangleq |x(t, f)|^2. \tag{43}$$

Intuitively speaking, the time-frequency coefficients $s(t, f)$ represents the energy of oscillatory components with frequency f at time instances t. It is noteworthy that $s(t, f)$ contains no information about the phase of that component.

It is well known that EEG signals have very non-flat spectrum with an overall 1/f shape, besides state-dependent peaks at specific frequencies. Therefore, the map $s(t, f)$ contains most energy at low frequencies f. If we directly apply bump modeling to the map $s(t, f)$, most bumps would be located in the low-frequency range, in other words, the high-frequency range would be under-represented. Since relevant information might be contained at high frequency, we normalize the map $s(t, f)$ before extracting the bump models.

We wish to point out that the time-frequency map $s(t, f)$ may be determined by alternative methods. For example, one may compute $s(t, f)$ by the multi-taper method (**?**) or by filterbanks (**?**). We decided to use the Morlet wavelet transformation for two reasons:

- Morlet wavelets have the optimal joint time-frequency resolution. We remind the reader of the fact that the joint time-frequency resolution is fundamentally limited by the uncertainty principle: the resolution in both time and frequency cannot be arbitrarily high *simultaneously*. It is well known that the Morlet wavelets achieve the uncertainty relation with *equality* (**? ? ?**).
- EEG signals are typically highly non-stationary; the wavelet transform is

ideally suited for non-stationary signals (**?**), in contrast to approaches based on multitapers and filterbanks.

However, it may also be meaningful to use multitaper method or filterbanks. For example, the multitaper method too is optimal in some sense: it minimizes out-of-band "leakage", and all "voxels" of the time-frequency domain have the same size and shape. In addition, in the multitaper method all estimates are independent due to orthogonality, a property not shared by wavelets (**?**).

6.2.2 Normalization

The coefficients $s(t, f)$ are centered and normalized, resulting in the coefficients $\tilde{z}(t, f)$:

$$\tilde{z}(t, f) \triangleq \frac{s(t, f) - m_s(f)}{\sigma_s(f)}, \tag{44}$$

where $m_s(f)$ is obtained by averaging $s(t, f)$ over the whole length of the EEG signal:

$$m_s(f) = \frac{1}{K} \sum_{t=1}^{K} s(t, f). \tag{45}$$

Likewise, $\sigma_s^2(f)$ is the variance of $s(t, f)$:

$$\sigma_s^2(f) = \frac{1}{K} \sum_{t=1}^{K} \Big(s(t, f) - m_s(f)\Big)^2. \tag{46}$$

In words: the coefficients $\tilde{z}(t, f)$ encode fluctuations from the baseline EEG power at time t and frequency f. The normalization (44) is known as z-score (see, e.g., (**?**)), and is commonly applied (**? ? ? 9?**). The coefficients $\tilde{z}(t, f)$ are positive when the activity at t and f is stronger than the baseline $m_s(f)$ and negative otherwise.

There are various approaches to apply bump modeling to the z-score $\tilde{z}(t, f)$. One may first set the negative coefficients to zero, and next apply bump mod-

eling. The bump models in that case represent peak activity. Alternatively, one may first set the positive coefficients equal to zero, reverse the sign of the negative coefficients, and then apply bump modeling. In that case, the bump models represent dips in the energy maps $s(t, f)$.

In the application of diagnosing AD (see Section 6.1), we will follow yet another approach. In order to extract bump models, we wish to exploit as much information as possible from the $\tilde{z}$ maps. Therefore we will set only a small fraction of the coefficients $\tilde{z}(t, f)$ equal to zero, i.e., the 1% smallest coefficients. This approach was also followed in (9), and is equivalent to the following transformation: we shift the coefficients (44) in the positive direction by adding a constant α, the remaining negative coefficients are set to zero:

$$z(t, f) \triangleq \left\lceil \tilde{z}(t, f) + \alpha \right\rceil^+ = \left\lceil \frac{s(t, f) - m_s(f)}{\sigma_s(f)} + \alpha \right\rceil^+, \tag{47}$$

where $\lceil x \rceil^+ = x$ if $x \geq 0$ and $\lceil x \rceil^+ = 0$ otherwise. The constant α is chosen such that only 1% of the coefficients remains negative after addition with α; this corresponds to $\alpha = 3.5$ in the application of diagnosing AD (see Section 6.1). (In the study of (9), it corresponds to $\alpha = 2$.) The top row of Fig. 3 shows the normalized wavelet map z (47) of two EEG signals.

6.2.3 Bump Modeling

Next, bump models are extracted from the coefficient maps z (see Fig. 3 and (9)). We approximate the map $z(t, f)$ as a sum $z_{\text{bump}}(t, f, \theta)$ of a "small" number of smooth basis functions or "bumps" (denoted by f_{bump}):

$$z(t, f) \approx z_{\text{bump}}(t, f, \theta) \triangleq \sum_{k=1}^{N_b} f_{\text{bump}}(t, f, \theta_k), \tag{48}$$

where θ_k are vectors of bump parameters and $\theta \triangleq (\theta_1, \theta_2, \ldots, \theta_{N_b})$. The sparse bump approximation $z_{\text{bump}}(t, f, \theta)$ represents regions in the time-frequency plane where the EEG contains more power than the baseline; in other words, it captures the most significant oscillatory activities in the EEG signal.

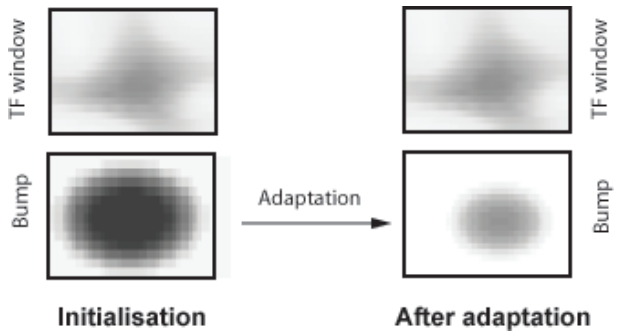


Fig. 4. Learning the bump parameters by minimizing the quadratic cost function (49); Top (left and right): a given patch of the time-frequency map. Bottom left: initial bump; Bottom right: bump obtained after adaptation.

We choose half-ellipsoid bumps since they are well suited for our purposes (**?** 9) (see Fig. 4). Since we wish to keep the number of bump parameters as low as possible, the principal axes of the half ellipsoid bumps are restricted to be parallel to the time-frequency axes. As a result, each bump is described by five parameters: the coordinates of its center (i.e., time t_k and frequency f_k), its amplitude $w_k > 0$, and the extension Δt_k and Δf_k in time and frequency respectively, in other words, $\theta_k = (t_k, f_k, w_k, \Delta t_k, \Delta f_k)$. More precisely, the ellipsoid bump function $f_{\text{bump}}(t, f, \theta_k)$ is defined as:

$$f_{\text{bump}}(t, f, \theta_k) = \begin{cases} w_k \sqrt{1 - \kappa(t, f, \theta_k)} & \text{for } 0 \leq \kappa(t, f, \theta_k) \leq 1 \\ 0 & \text{for } \kappa(t, f, \theta_k) > 1, \end{cases} \tag{49}$$

where

$$\kappa(t, f, \theta_k) = \frac{(t - t_k)^2}{(\Delta t_k)^2} + \frac{(f - f_k)^2}{(\Delta f_k)^2}. \tag{50}$$

For the EEG data described in Section 6.1, the number of bumps N_b (cf. (48)) is typically between 50 and 100, and therefore, $z_{\text{bump}}(t, f, \theta)$ is fully specified by a few hundred parameters. On the other hand, the time-frequency map $z(t, f)$ consists of between 10^4 and 10^5 coefficients; the bump model $z_{\text{bump}}(t, f, \theta)$ is thus a sparse (but approximate) representation of $z(t, f)$.

The bump model $z_{\text{bump}}(t, f, \theta)$ is extracted from $z(t, f)$ by the following algorithm (**?** 9):

(1) Define appropriate boundaries for the map $z(t, f)$ in order to avoid finite-size effects.

(2) Partition the map $z(t, f)$ into small zones. The size of these zones depends on the time-frequency ratio of the wavelets, and are optimized to model oscillatory activities lasting 4 to 5 oscillation periods. Larger oscillatory patterns are modeled by multiple bumps.

(3) Find the zone $\mathcal{Z}$ that contains the most energy.

(4) Adapt a bump to that zone; the bump parameters are determined by minimizing the quadratic cost function (see Fig. 4):

$$\mathcal{E}(\theta_k) \triangleq \sum_{t,f \in \mathcal{Z}} \Big(z(t, f) - f_{\text{bump}}(t, f, \theta_k) \Big)^2. \tag{51}$$

Next withdraw the bump from the original map.

(5) The fraction of total intensity contained in that bump is computed:

$$F = \frac{\sum_{t,f \in \mathcal{Z}} f_{\text{bump}}(t, f, \theta_k)}{\sum_{t,f \in \mathcal{Z}} z(t, f)}. \tag{52}$$

If $F < G$ for three consecutive bumps (and hence those bumps contain only a small fraction of the energy of map $z(t, f)$), stop modeling and proceed to (6), otherwise iterate (3).

(6) After all signals have been modeled, define a threshold $T \geq G$, and remove

the bumps for which $F < T$. This allows us to trade off the information loss and modeling of background noise: when too few bumps are generated, information about the oscillatory activity of the brain is lost. On the other hand, if too many bumps are generated, the bump model also contains low-amplitude oscillatory components; since the measurement process typically introduces a substantial amount of noise, it is likely that the low-amplitude oscillatory components do not stem from organized brain oscillations but are instead due measurement noise. By adjusting the threshold T, we try to find an appropriate number of bumps.

In the present application, we used a threshold $G = 0.05$. With this threshold, each bump model contains many bumps. Some of those bumps may actually model background noise. Therefore, we further pruned the bump models (cf. Step 6). We tested various values of the threshold $T \in [0.2, 0.25]$; as we will show, the results depend on the specific choice of T: the optimal separation between MCI and age-matched control subjects is obtained for $T = 0.22$, the separation gradually diminishes for increasing and decreasing values of T. We refer to (**?** 9) for more information on bump modeling. In particular, we used the same choice of boundaries (Step 1) and partitions (Step 2) as in those references.

Eventually, we obtain 21 bump models, i.e., one per EEG channel. In the following, we describe how those models are further processed.

6.2.4 Aggregation

As a next step, we group the 21 electrodes into a small number N_R of regions, as illustrated in Fig. 5 for $N_R = 5$; we will report results for $N_R = 3$, 5, and

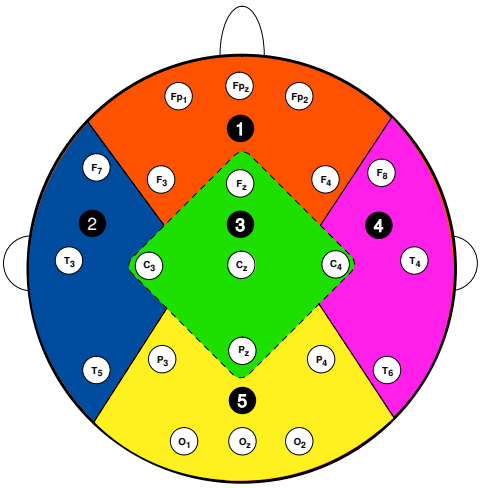


Fig. 5. The 21 electrodes used for EEG recording, distributed according to the 10–20 international placement system (18). The clustering into $N_R = 5$ zones is indicated by the colors and dashed lines (1 = frontal, 2 = left temporal, 3 = central, 4 = right temporal and 5 = occipital).

7. From the 21 bump models obtained by sparsification (cf. Section 6.2.3), we extract a single bump model for each of the zones by means of the aggregation algorithm described in (9).

6.2.5 Stochastic Event Synchrony

The results from our exemplar-based approach are summarized in Table 1; the results are for the binomial prior, the results for the multinomial prior closely agree. In other words, the choice of prior does again not seem to be crucial. We adopted constant parameters, because time-varying parameters are less suitable since we consider spontaneous EEG. We studied the following statistics:

- Posterior distribution $p(c_m = i|X) = p_i^c$ of the number of copies of each

exemplar c_m, parameterized by $(p_0^c, p_1^c, \ldots, p_4^c)$,

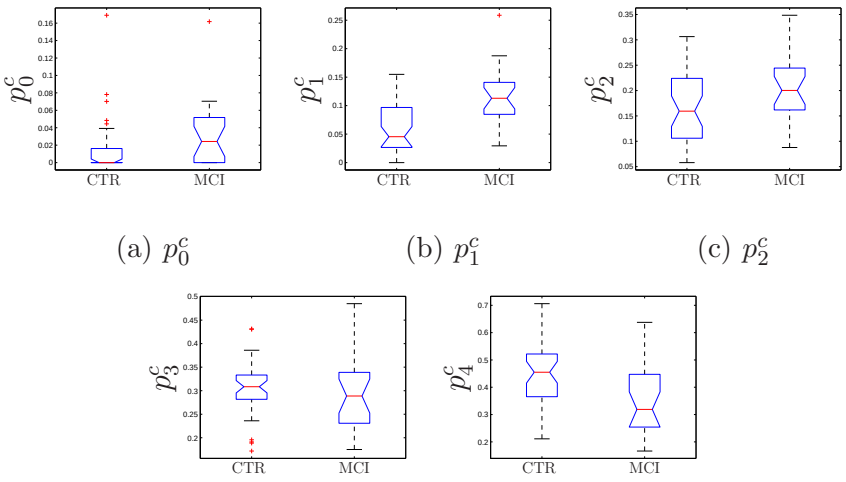
- $\bar{c}_m$: average number of copies per cluster,
- s_t: variance in time domain ("time jitter"),
- s_f: variance in frequency domain ("frequency jitter"),
- $\Delta\bar{T}$: average width of bumps,
- $\Delta\bar{F}$: average height of bumps,
- $\bar{F}$: average frequency of bumps.

We also consider the linear combination h^c of all parameters p_i^c that optimally separates both subject groups. Interestingly, the latter statistic amounts to about the same p-value as the index ρ of SES (5). The posterior $p(c_m|X)$ mostly differs in p_1^c, p_2^c and p_4^c (see Fig. 6): in MCI patients, the number of clusters of size five (p_4^c) significantly decreases; on the other hand, the number of clusters of size one (p_1^c) and two (p_2^c) significantly increases. This explains and confirms the observed increase of ρ in MCI patients (5). Combining h^c with ffDTF and $\Delta\bar{T}$ (or $\bar{c}_m$ with ffDTF and $\Delta\bar{T}$) allows to separate the two groups quite well (more than 90% correctly classified), as shown in Fig. 8; this is far better than what can be achieved by means of classical similarity measures (about 75% correctly classified). Classification rates between 80 and 85% can be obtained by combining two features (see Fig. 7(a)).

Stat.	p_0^c	p_1^c	p_2^c	p_3^c	p_4^c	$\bar{c}_m$	h^c	σ_t	σ_f	$\Delta\bar{T}$	$\Delta\bar{F}$	$\bar{F}$
p-value	0.016	2.9^{-4**}	0.089	0.59	0.0054*	1.10^{-3**}	1.10^{-4**}	0.46	0.28	$2.3.10^{-4**}$	0.023*	2.10^{-3**}

Table 1

Sensitivity of multivariate SES for diagnosing MCI (p-values for Mann-Whitney test; * and ** indicate $p < 0.05$ and $p < 0.005$ respectively).



(a) p_0^c (b) p_1^c (c) p_2^c

(d) p_3^c (e) p_4^c

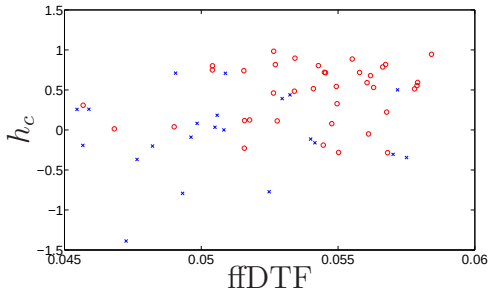
Fig. 6. Box plots for posterior distribution $p(c_m = i|X) = p_i^c$.

7 Conclusion

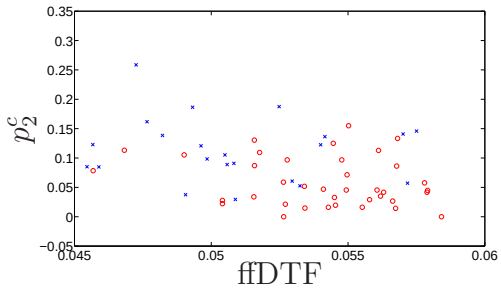
We proposed an approach to determine the similarity of multiple (one- and multi-dimensional) point processes; it is based on an exemplar-based statistical model that describes how the point processes are related through a common hidden "mother" process. The similarity of the point processes is determined by performing inference in that model by means of integer programming techniques in conjunction with point estimation of the parameters. The proposed technique may be used for various applications in neuroscience (e.g., in brain-computer interfaces, analysis of spike data), biomedical signal processing, and beyond.

References

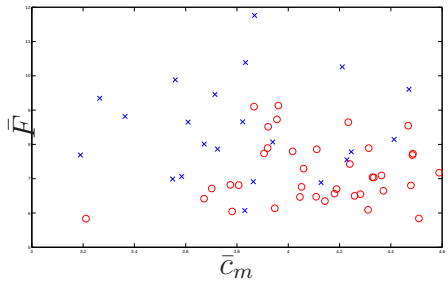
[1] E. Pereda, R. Q. Quiroga, and J. Bhattacharya, "Nonlinear multivariate analysis of neurophsyiological signals," *Progress in Neurobiology,* 77 (2005) 1–37.



(a) h_c vs. ffDTF



(b) p_2^c vs. ffDTF



(c) $\bar{F}$ vs. $\bar{c}_m$

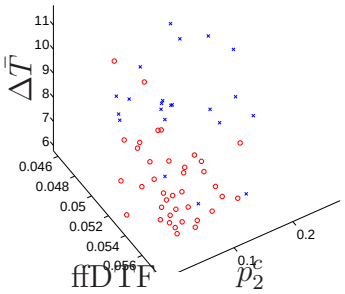
Fig. 7. Combination of two features.

[2] G. Buzsáki, *Rhythms of the Brain,* Oxford University Press, 2006.

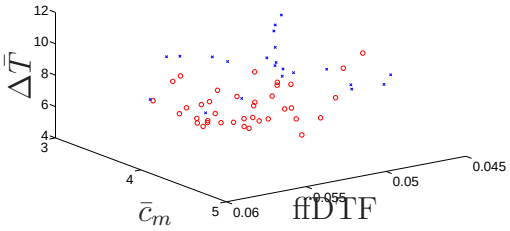
[3] T. Womelsdorf, J.M. Schoffelen, R. Oostenveld, W. Singer, R. Desimone, A.K. Engel, P. Fries, "Modulation of neuronal interactions through neuronal synchronization," *Science,* 316:1609–1612.

[4] P. Uhlhaas and W. Singer, "Neural synchrony in brain disorders: relevance for cognitive dysfunctions and pathophysiology," *Neuron,* 52:155–168, 2006.

[5] J. Dauwels, F. Vialatte, T. Rutkowski, and A. Cichocki, "Measuring neural synchrony by message passing," *Advances in Neural Information Processing Sys-*



(a) $\Delta\bar{T}$ vs. ffDTF vs. p_2^c



(b) $\Delta\bar{T}$ vs. $\bar{c}_m$ vs. ffDTF

Fig. 8. Combination of three features.

tems 20, MIT Press.

[6] B. Frey and D. Dueck, "Clustering by passing messages between data points," *Science,* vol. 315, No. 5814, pp. 972–976, 2007.

[7] B. Frey and D. Dueck, "Mixture modelling by affinity propagation," In *Advances in Neural Information Processing Systems 18,* MIT Press.

[8] D. Lashkari and P. Golland, "Convex clustering with exemplar-based models," In *Advances in Neural Information Processing Systems 20,* MIT Press.

[9] F. Vialatte, C. Martin, R. Dubois, J. Haddad, B. Quenet, R. Gervais, and G. Dreyfus, "A Machine learning approach to the analysis of time-frequency maps, and its application to neural dynamics," *Neural Networks,* 2007, 20:194–209.

[10] J. Dauwels, F. Vialatte, and A. Cichocki, "A comparative study of synchrony measures for the early detection of AD," *Proc. ICONIP 2007,* Kitakyushu, September 2007.

[11] J.-P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela, "Measuring Phase Synchrony in Brain Signals," *Human Brain Mapping* 8:194208 (1999).

[12] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating Mutual Information," *Phys. Rev. E* 69 (6) 066138, 2004.

[13] S. Aviyente, "A Measure of Mutual Information on the Time-Frequency Plane," *Proc. of ICASSP 2005,* vol. 4, pp. 481–484, March 18–23, 2005, Philadelphia, PA, USA.

[14] M. Kamiński and Hualou Liang, "Causal Influence: Advances in Neurosignal Analysis," *Critical Review in Biomedical Engineering,* 33(4):347–430 (2005).

[15] I. J. Schoenberg, "Spline functions and the problem of graduation," *Mathematics* 52: 974–50.

[16] D. MacKay, *Information Theory, Inference, and Learning Algorithms,* Cambridge University Press, 2003.

[17] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Processing Magazine*, Jan. 2004, pp. 28–41.

[18] P. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG,* Oxford University Press, 2006.

[19] `http://www.ilog.com/products/optimization/`

[20] V. Kann, "Maximum bounded 3-dimensional matching is MAX SNP-complete", *Inform. Process. Lett.* 37, 27–35, 1991.

[21] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness,* W. H. Freeman & Co., 1979.