Quantifying Statistical Interdependence PART III: $N > 2$ Point Processes

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

[a] School of Electrical and Electronic Engineering (EEE), Nanyang Technological University (NTU), Singapore.

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

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

[d] Laboratoire SIGMA, ESPCI ParisTech, Paris, France.

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

Abstract

Stochastic event synchrony (SES) is a recently proposed family of similarity measures. First, "events" are extracted from the given signals; next, one tries to align events across the different time series. The better the alignment, the more similar the N time series are considered to be. The similarity measures quantify the reliability of the events (fraction of "non-aligned" events) and the timing precision.

So far, SES has been developed for pairs of one-dimensional (Part I) and multi-dimensional (Part II) point processes. In this paper (Part III), SES is extended from pairs of signals to $N > 2$ signals.

The alignment and SES parameters are again determined through statistical inference, 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 by linear integer programming.

In order to test the robustness and reliability of the proposed N-variate SES method, it is first applied to synthetic data. We show that N-variate SES results in more reliable estimates than bivariate SES. Next N-variate SES is applied to two problems in neuroscience: it used to quantify the firing reliability of Morris-Lecar neurons, and to detect anomalies in EEG synchrony of Mild Cognitive Impairment (MCI) patients; those problems were also considered in Part I and II respectively. In both cases, the N-variate SES approach yields a more detailed analysis.

Key words: timing precision, event reliability, stochastic event synchrony, maximum a posteriori estimation, Morris-Lecar neuron, firing reliability, time-frequency map, bump model, EEG, Alzheimer's disease, mild cognitive impairment

1 Introduction

Neural synchrony may play an important role in information processing in the brain. Although the details of this coding mechanism have not been fully revealed, it has been postulated that neural synchrony may be involved in cognition (Varela *et al.*, 2001) and even in consciousness (Singer, 2001). The correlation between different brain signals has been studied intensively in recent years both by experimental neuroscientists (e.g., (Abeles *et al.*, 1993)) and computational neuroscientists (e.g., (Amari *et al.*, 2003)), and also by neurologists; indeed, various medical studies have reported that neurological diseases (such as Alzheimer's disease and epilepsy) are related to perturbations in neural synchrony (Matsuda *et al.*, 2001; Jeong, 2004).

Motivated by the intensified interest in neural synchrony, numerous researchers have in the last years developed and refined methods to quantify the synchrony between signals (see, e.g., (Stam, 2005; Quiroga *et al.*, 2002; Pereda *et al.*, 2005; Toups *et al.*, 2011)). In recent work (Dauwels *et al.*, 2007, 2009a,b), we have proposed a new family of synchrony measures referred to as stochastic event synchrony (SES); this class of synchrony measures is inspired by the Victor-Purpura distance metrics (Victor *et al.*, 1997). The basic idea is the following: First, we extract "events" from the given time series; next, we try to align events from one time series with events from the other. The better the alignment, the more similar the time series are considered to be. We also quantify the timing jitter between matched ("coincident") events; the smaller the timing jitter, the larger the synchrony. SES thus considers two different aspects of synchrony: reliability and timing precision. Those concepts were also recently considered in (Toups *et al.*, 2011), and they can be understood from the following analogy: when you wait for a train in the station, the train may come at the station or it may not come at all; for example, it may be out of service due to some mechanical problem. If the train comes, it may or may

* 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] Part of this work was carried out while J.D. was at the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology (MIT), Cambridge, MA.

not be on time. The former uncertainty is related to reliability, whereas the latter is related to precision.

So far, SES was restricted to pairs of signals. In this paper (Part III), we extend SES from pairs of signals to $N > 2$ signals. The underlying principle is similar, but the inference algorithm to compute the SES parameters is fundamentally different. In bivariate SES (Dauwels *et al.*, 2007, 2009a,b) we applied the max-product algorithm to align pairs of sequences. We have implemented the max-product algorithm (and various refinements) for aligning $N > 2$ sequences as well; however, that approach unfortunately leads to inaccurate alignments, probably because N-wise alignment is substantially harder than pairwise alignment. As an alternative, we solve the N-wise alignment by integer linear programming, which yields optimal or near-optimal alignments at reasonable computational cost.

The extension from pairs of signals to $N > 2$ signals is non-trivial. In the following, we briefly touch upon this issue. For $N = 2$, the problem of time series comparison is essentially that of finding an alignment between the time series. The "alignment" approach of Victor & Purpura (Victor *et al.*, 1997) was the inspiration for bivariate SES; we have shown in (Dauwels *et al.*, 2007, 2009a,b) that the alignment between 2 time series (say x_1 and x_2) is the same as finding an underlying time series v which can then be transformed to x_1 and x_2 by a sequence of steps involving jittering the events, and insertions and deletions. Quantifying the distance between x_1 and x_2 is equivalent to finding a v that minimizes some combination of $d(x_1, v)$ and $d(x_2, v)$. By making the combination rule for these two distances accelerating (such as a root-mean-square), one can ensure that v is in the "middle" of x_1 and x_2. Comparing x_1 with x_2 is equivalent to finding a hidden "consensus" process v that can generate both x_1 with x_2. Finding the v "interprets" the similarity of x_1 and x_2 in terms of a consensus that they both represent, but is not necessary to find v to quantify this similarity, since it can also be expressed as an alignment, without an explicit v.

For $N = 3$, one can attempt to jointly minimize some combination of $d(x_1, v)$, $d(x_2, v)$, and $d(x_3, v)$, without attempting to find an underlying v. Or, one can attempt to find a single v for which some combination of $d(x_1, v)$, $d(x_2, v)$, and $d(x_3, v)$ are smallest.

Finally, for $N > 3$, there is at least one other possibility: the point process might be generated by two or more hidden processes. This can first happen at $N = 4$, with two hidden processes, say v and w, for which the distances $d(x_1, v)$, $d(x_2, v)$, $d(x_3, w)$, $d(x_4, w)$ have a lower total than the distance from any single v to all four of the observations x_1, x_2, x_3, and x_4. In other words, for $N > 3$, the problem of finding multivariate similarity is generally not the same as finding the single underlying "consensus" process.

In summary, there are at least 3 kinds of definitions of synchrony and similarity:

(1) a definition based on the $N(N-1)/2$ distances $d(x_i, x_j)$ between pairs of observed processes,
(2) a definition based on finding a consensus process v, which minimizes the N distances $d(x_i, v)$,
(3) a definition based on finding multiple hidden processes.

For $N = 2$, these definitions are identical. For $N = 3$, the second and third are identical, but differ from the first. For $N > 3$, all three definitions are distinct. In this paper, we consider the second definition: the N point processes are generated from a (hidden) consensus process v. In future work, we will extend N-variate SES to multiple hidden processes.

Stochastic event synchrony (SES) is applicable to any kind of time series (e.g., from finance, oceanography, and seismology). We will consider here neural spike trains and electroencephalograms (EEG). More specifically, we will use the SES method to quantify the reliability of Morris-Lecar neurons, and to predict mild cognitive impairment (MCI) from electroencephalograms; we also considered those problems in Part I (Dauwels *et al.*, 2009a) and Part II (Dauwels *et al.*, 2009b) respectively.

This paper is organized as follows. In the next section, we explain how SES can be extended from pairs of point processes to $N > 2$ point processes. We describe the underlying statistical model (Section 3), and outline our inference method[2] (Section 4). We consider various extensions of our statistical model (Section 5). We investigate the robustness and reliability of the SES inference method by means of synthetic data (Section 6). We use SES to quantify the reliability of Morris-Lecar neurons (Section 7), and to detect abnormalities in the EEG synchrony of MCI disease patients (Section 8). At last, we offer some concluding remarks (Section 9).

Readers who are less interested in the technical details may wish to read Section 2, where the general idea is outlined, and Sections 7 and 8, where two applications are discussed; those sections can be read independently of the more technical Sections 3 and 4.

[2] An implementation of N-variate SES for one-dimensional and multi-dimensional point processes is available from `http://www.dauwels.com/SESToolbox/SES.html`.

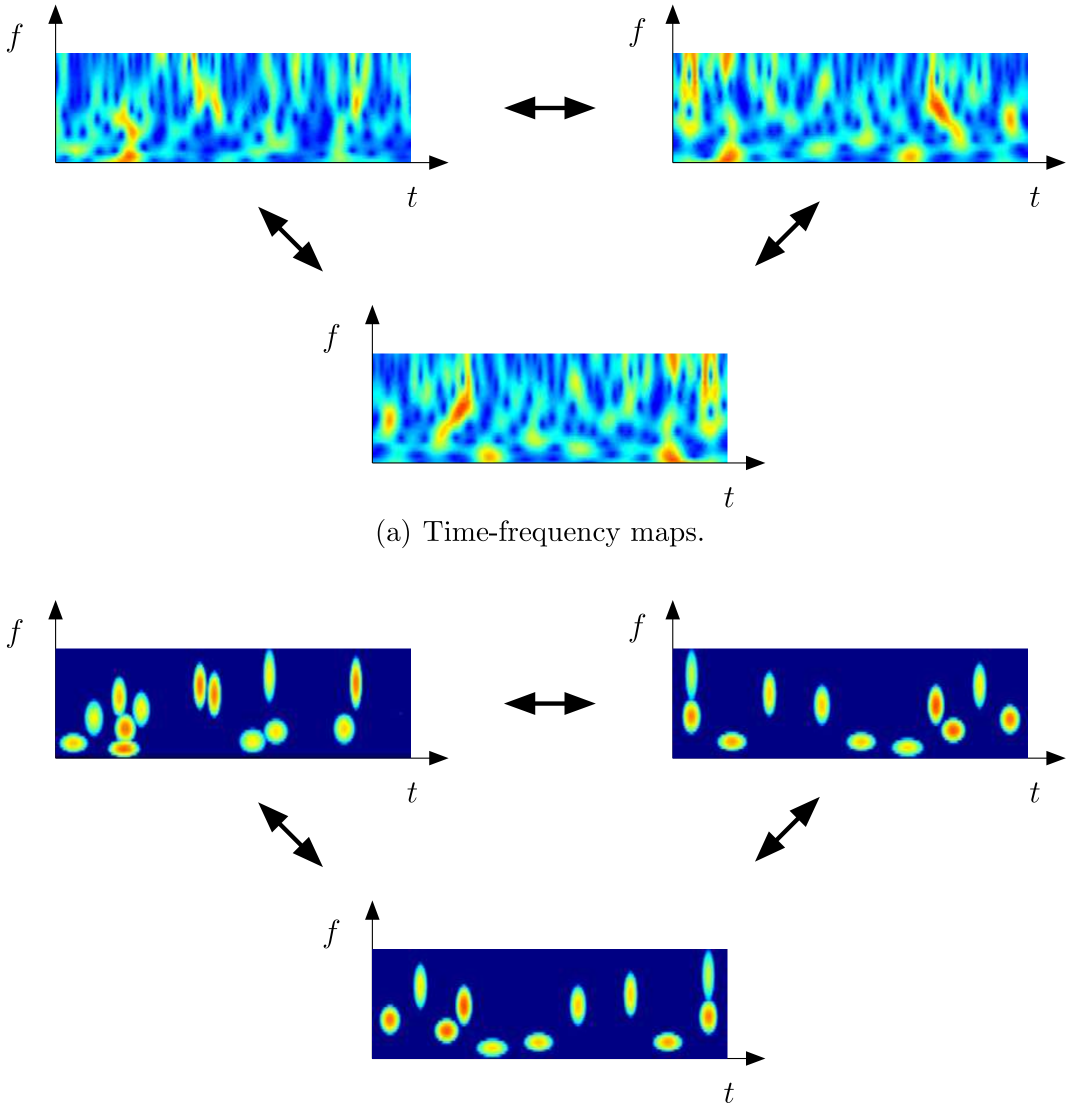
2 Principle

Suppose that we are given $N > 2$ continuous-time signals (e.g., electroencephalogram (EEG) signals recorded from different channels), and we wish to determine the similarity of those signals. In (Dauwels *et al.*, 2009b) we considered pairs of signals ($N = 2$); although the extension to $N > 2$ may seem straightforward, it leads to a combinatorial problem that is much harder. In the following, we will closely follow the setting and notation of (Dauwels *et al.*, 2009b).

As a first step, we extract point processes from those $N > 2$ 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. 1 and (Vialatte *et al.*, 2007)). Each bump is described by five parameters: time t, frequency f, width Δt, height Δf, and amplitude w. The resulting bump models 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 (Buzsáki, 2006; Nunez *et al.*, 2006). In the following, we will develop N-variate SES for bump models. In this setting, SES quantifies the synchronous interplay between oscillatory patterns in $N > 2$ 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. 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). Such extensions may straightforwardly be derived from the example of bump models; we refer to Section 5 and (Dauwels *et al.*, 2009b) for more details.

It is also noteworthy that the events in the point processes may be labeled by discrete tags. For example, in multineuronal recordings, the labels would correspond to the neurons of origin. More generally, the labels may correspond to the sites or processes of origin. SES may then be applied to the point processes associated with each label, and it would quantify the coupling between



(a) Time-frequency maps.

(b) Bump models.

Fig. 1. 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.

the different sites (e.g., neurons).

We now consider the central question: how can we quantify the similarity of $N > 2$ point processes defined on a space S? Let us consider the example of bump models (see Fig. 1 and Fig. 2). Intuitively speaking, N bump models $(x_i)_{i=1,2,\ldots,N}$ may be considered well-synchronized if bumps appear in (almost) all bump models *simultaneously*, apart from a constant offset in time and frequency, and a "small" amount of jitter in time and frequency. If one overlays N well-synchronized bump models, and removes the potential average offsets in time and frequency (denoted by δ_{ti} and δ_{fi} respectively, for $i = 1, 2, \ldots, N$), bumps naturally appear in clusters that contain precisely one

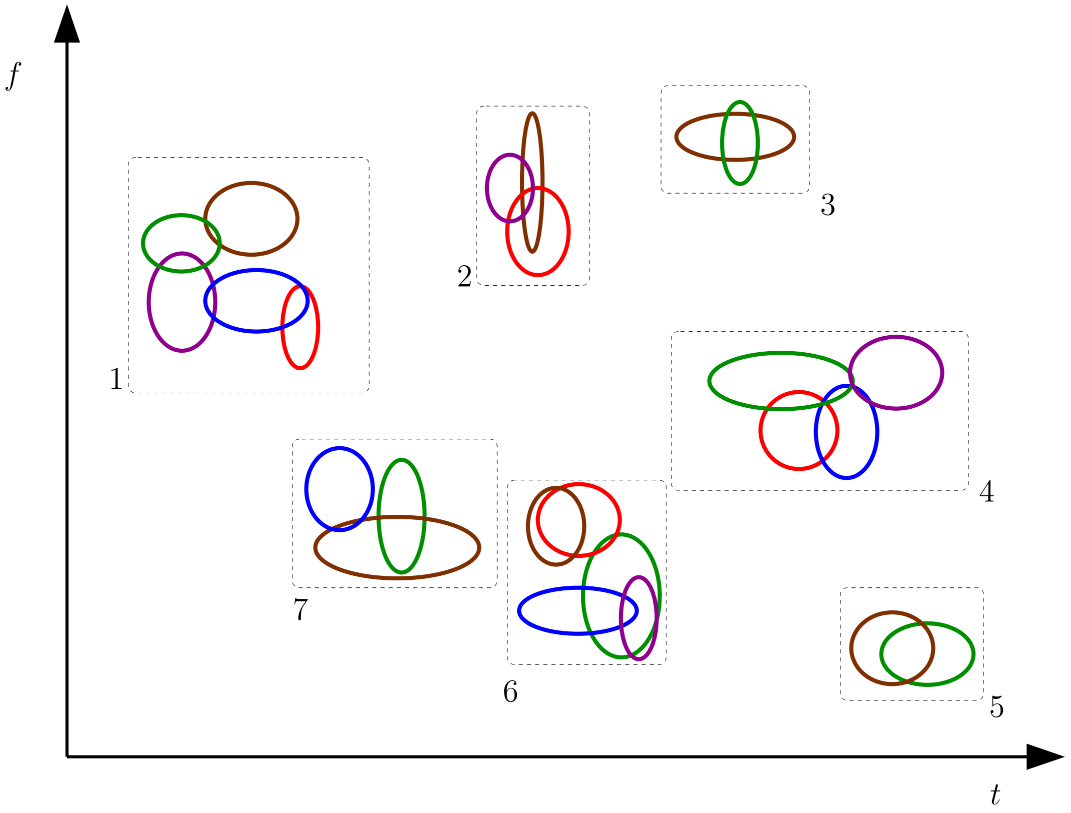


Fig. 2. Five bump models overlayed on top of each other ($N = 5$); the dashed boxes indicate clusters. The average offset between the bumps is close to zero.

bump from all (or almost all) bump models, as illustrated in Fig. 2 for $N = 5$. In that example, cluster 1 and 6 contain bumps from all 5 bump models x_i, cluster 2 and 7 contain bumps from 3 bump models, cluster 3 and 5 contain bumps from 2 bump models, and cluster 4 contains bumps from 4 bump models. In general, the average offset in time and frequency between the bumps is not necessarily zero (as illustrated in Fig. 3), and it may be harder to recognize the different clusters. The algorithm developed in this paper is able to extract such clusters, even in the general case of non-zero average offsets in time and frequency (as in Fig. 3).

If the point processes are well-synchronized, almost all clusters contain (close to) N bumps, specifically, one bump from each (or almost each) of the N bump models. Therefore, an important similarity statistics is the average number of events per cluster, or more generally, the statistical distribution of the number of events per cluster. Moreover, as in the pairwise case (Dauwels *et al.*, 2009a,b), one can quantify how well the bumps are aligned within each cluster, by computing the jitter s_{ti} and s_{fi} in time and frequency respectively, for $i = 1, 2, \ldots, N$.

More generally, N point processes on a space S may be considered similar if

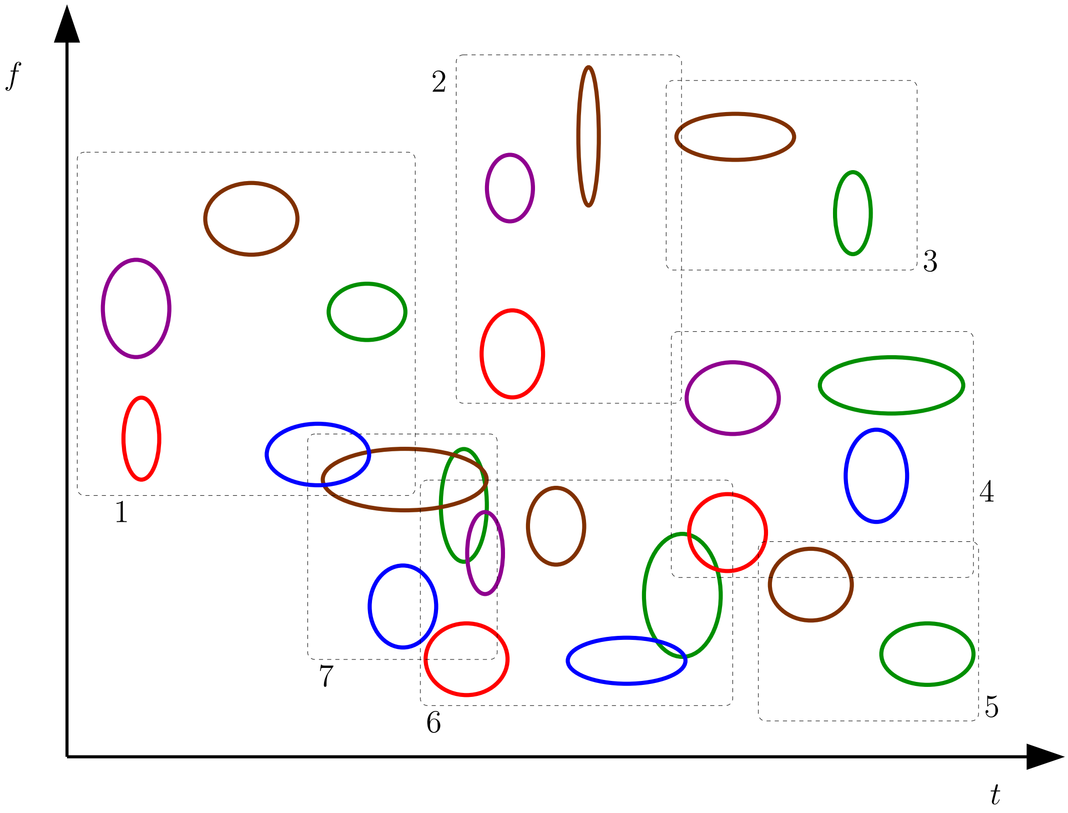


Fig. 3. Five bump models overlayed on top of each other ($N = 5$); the dashed boxes indicate clusters. The average offset between the bumps is non-zero. For example, the red and brown bumps tend to be located at the bottom and top respectively of the clusters, whereas the green bumps tend to lie at the right side of the clusters. Such tendencies cannot be observed in Fig. 2.

events appear in clusters with (close to) N events and with "small" dispersion (computed by the distance measure on S); those clusters may only appear after certain transformations have been applied, e.g., translation (to eliminate offsets as in Fig. 3), rotation, and scaling. In other words, N point processes are considered similar, if they can be transformed into each other by a few operations, including deletions and insertions, "small" random perturbations, and transformations such as translation, rotation, and scaling.

Let us now return to bump models. We determine the SES parameters δ_{ti}, δ_{fi}, s_{ti} and s_{fi} ($i = 1, 2, \ldots, N$), and the event clusters by *statistical inference*, along the lines of the pairwise case (Dauwels *et al.*, 2009a,b). We start by constructing a statistical model that captures the relation between the N bump models; that statistical model contains the SES parameters, besides variables related to the alignment of the different bumps. Next we perform inference in that statistical model, resulting in estimates for the SES parameters and clusters. More concretely, we apply cyclic maximization, as in the pairwise case.

In the following section, we outline our statistical model; in Section 4 we describe how we conduct inference in that statistical model.

3 Statistical Model

3.1 Casual Description

The intuitive concept of similarity outlined in Section 2 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 process v. The observed sequences $(x_i)_{i=1,\ldots,N}$ are obtained from v by the following four-step procedure:

(1) COPY: generate N copies of the hidden point process v,
(2) DELETION: delete some of the copied events,
(3) PERTURBATION: shift the remaining copies over $(\delta_{ti}, \delta_{fi})_{i=1,\ldots,N}$, and randomly perturb the positions, with variance $(s_i)_{i=1,\ldots,N} = (s_{ti}, s_{fi})_{i=1,\ldots,N}$, amounting to the N point processes $(x_i)_{i=1,\ldots,N}$.
(4) INSERTION: Additional events are inserted ("background events"), which are unrelated to the hidden point process v, and are modeled as mutually independent.

As a result, each sequence x_i consists of "noisy" copies of hidden events (generated by Step 1 to 3), besides background events (generated by Step 4). The noisy copies are related to each other through the hidden events v, whereas the background events are all independent of each other. The point processes x_i may be considered well-synchronized if there are only few deletions (cf. Step 2) and insertions (cf. Step 4), and if the events of x_i are "close" to the corresponding hidden events (cf. Step 3), apart from offsets $(\delta_{ti}, \delta_{fi})_{i=1,\ldots,N}$. Fig. 4 illustrates a generative process that results in the bump models of Fig. 3. More generally, as we pointed out in the previous section, one may include other transformations in the perturbation step besides translation over $(\delta_{ti}, \delta_{fi})_{i=1,\ldots,N}$, such as rotation and scaling.

Some readers may wonder why insertions need to be modeled explicitly. Indeed, inserting an event is equivalent to adding a hidden event with N "noisy" copies, followed by $N-1$ deletions. In this way, the SES models for *pairs* of point processes (Part I and II) are able to capture insertions, even though they are not modeled explicitly (Dauwels *et al.*, 2009a,b). However, for large N (e.g., $N > 10$), the cost of an insertion becomes prohibitively large (due to the $N-1$ deletions), and as a consequence, the statistical model does no longer capture insertions. The inserted event will be grouped with other events, lead-

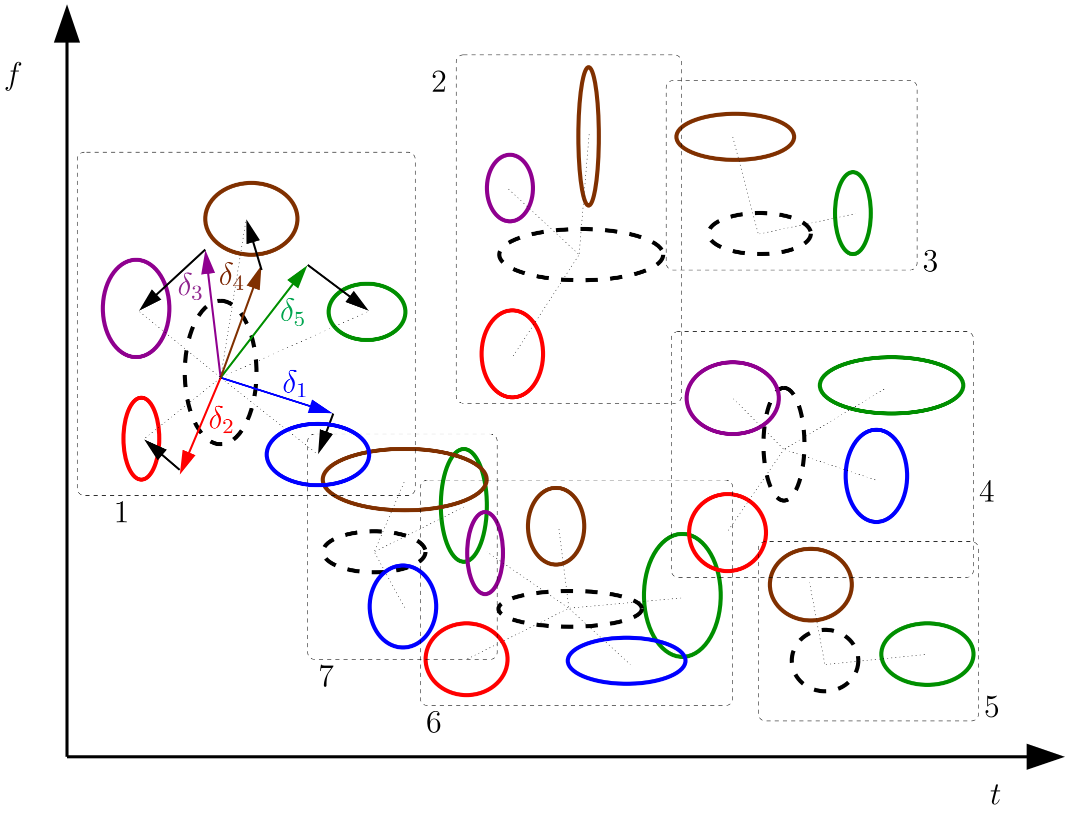


Fig. 4. Generative model for the $N = 5$ bump models $(x_i)_{i=1,\ldots,N}$ of Fig. 3. One first generates a hidden bump model v, indicated in dashed lines. Next one makes $N = 5$ identical copies of v and shifts those over $(\delta_i)_{i=1,\ldots,N} = (\delta_{ti}, \delta_{fi})_{i=1,\ldots,N}$, as indicated by the colored arrows in cluster 1. The resulting events are then slightly shifted, with variance $(s_i)_{i=1,\ldots,N} = (s_{ti}, s_{fi})_{i=1,\ldots,N}$, as indicated by the black arrows in cluster 1. Finally, some of those events are deleted (with probability p_d), resulting in the bump models $(x_i)_{i=1,\ldots,N}$. For example, 2 events are deleted in cluster 2.

ing to incorrect clusters. Therefore, it becomes necessary to model insertions explicitly. As an illustration, Fig. 11(a) shows several event clusters in addition to background events (indicated by red hexagons). More information on this application can be found in Section 7. Note that insertions are also explicitly modeled in the Victor-Purpura distance metrics (Victor *et al.*, 1997), which were the source of inspiration for SES.

3.2 Formal Description

We now describe the underlying stochastic model in more detail. We refer to Table 1 for a summary of all relevant variables and parameters. For convenience, we introduce the following notation. The length of the point process x_i is denoted by L_i (with $i = 1, 2, \ldots, N$). The individual events of point process x_i are denoted by x_{ij} (with $i = 1, 2, \ldots, N$ and $j = 1, \ldots, L_i$). The occurrence

Symbol	Explanation
$(x_i)_{i=1,2,\ldots,N}$	N given bump models
L_i	length of x_i
L^{tot}	total number of bumps in the N models $(x_i)_{i=1,2,\ldots,N}$
x_{ij}	j-th bump in bump model x_i
t_{ij}	occurrence time of x_{ij}
f_{ij}	frequency of x_{ij}
δ_{ti} and δ_{fi}	average offset in time and frequency for bump model x_i
s_{ti} and s_{fi}	jitter in time and frequency for bump model x_i
p_d	deletion probability
v	hidden bump model from which the observed bump models $(x_i)_{i=1,2,\ldots,N}$ are generated
ℓ	length of v
c_{ij}	$c_{ij} = 0$ if x_{ij} is a background event, otherwise index of event in v that generated x_{ij}
$\mathcal{C}_k$	set of n_k copies of v_k (cf. (13))
$\mathcal{I}_k$	index set of $\mathcal{C}_k$ (cf. (13))
$\mathcal{K}$	index set of bump clusters of size $n_k > 1$ (cf. (12))
L	number of bump clusters, i.e., number of hidden events v_k with at least one copy $(n_k > 0)$
ρ	fraction of missing bumps in the clusters
$\tilde{v}$	background events in $(x_i)_{i=1,2,\ldots,N}$
$\tilde{\ell}$	length of $\tilde{v}$
χ	fraction of background events in $(x_i)_{i=1,2,\ldots,N}$
d_0	cost associated with a cluster/exemplar (cf. (25))
$\tilde{d}_0$	cost associated with a background event (cf. (26))
$d(t_{ij}, f_{ij})$	cost associated with an event x_{ij} that belongs to a non-trivial cluster (cf. (27))
$d(t_{ij}, f_{ij}, t_{i'j'}, f_{i'j'})$	cost associated with an event x_{ij} that belongs to exemplar $x_{i'j'}$ (cf. (48))
b_k	equal to one iff cluster k is non-empty ($n_k > 0$ and $k = 1, 2, \ldots, L^{\text{tot}}$), otherwise zero
b_{ijk}	equal to one iff the x_{ij} belongs to cluster k, i.e., $c_{ij} = k$, otherwise zero
b_{ij}	equal to one iff x_{ij} is an exemplar, otherwise zero
$b_{iji'j'}$	equal to one iff x_{ij} is associated with exemplar $x_{i'j'}$, otherwise zero
e_{ij}	equal to one iff x_{ij} is a background event, otherwise zero

Table 1
List of variables and parameters associated with models $p(x, c, v, \theta, \ell)$ (16), $p(x, c, \theta)$ (23) (31), and $p(x, b, e, \theta)$ (43) (49).

time and frequency of those events are referred to as t_{ij} and f_{ij} respectively. Moreover, we will use the notation $\delta_i = (\delta_{ti}, \delta_{fi})$, $s_i = (s_{ti}, s_{fi})$, $\theta_i = (\delta_i, s_i)$ (with $i = 1, 2, \ldots, N$), and $\theta = (\theta_1, \theta_2, \ldots, \theta_N)$.

The hidden process $v = \{v_1, \ldots, v_\ell\}$, which is the source of all events in $x_1, x_2, \ldots x_N$ (besides the background events; cf. Step 4), is modeled as follows. The number ℓ of points in v is geometrically distributed with parameter $\lambda \operatorname{vol}(S)$:

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

where $\operatorname{vol}(S)$ is the multi-dimensional volume of set S. (We motivate this choice of prior in Part I (Dauwels *et al.*, 2009a,b).) In the particular case of bump models in the time-frequency domain, the space S is defined as:

$$S = \{(t, f) : t \in [t_{\min}, t_{\max}] \text{ and } f \in [f_{\min}, f_{\max}]\}, \tag{2}$$

and therefore

$$\operatorname{vol}(S) = (t_{\max} - t_{\min})(f_{\max} - f_{\min}). \tag{3}$$

Each point v_k for $k = 1, \ldots, \ell$ is uniformly distributed in S:

$$p(\tilde{t}, \tilde{f}|\ell) = \operatorname{vol}(S)^{-\ell}, \tag{4}$$

where $\tilde{t}$ and $\tilde{f}$ are the positions of the hidden events $(v_k)_{k=1,\ldots,\ell}$ in time and frequency respectively. The amplitudes, widths and heights of the bumps v_k are independently and identically distributed according to priors p_w, $p_{\Delta t}$ and $p_{\Delta f}$ respectively. In the following, we will discard those priors since they are irrelevant. With those choices, the prior of the hidden process v equals:

$$p(v, \ell) = p(\ell)p(v|\ell) \propto (1 - \lambda \operatorname{vol}(S))\lambda^{\ell}, \tag{5}$$

where the priors for the amplitudes, widths and heights of the bumps v_k have been discarded for convenience.

From the hidden process v, the point processes $(x_i)_{i=1,\ldots,N}$ are generated as follows (see Fig. 4). We first generate N identical copies of v. Next the amplitudes, widths and heights of the bumps are replaced by random independent draws from priors p_w, $p_{\Delta t}$ and $p_{\Delta f}$ respectively. Again, those priors are irrelevant for what follows, and we will omit them. Next, each event is removed with probability p_d ("deletion"), independently of the other events. The probability mass associated with the (remaining) n_k copies of $(v_k)_{k=1,\ldots,\ell}$ is given by:

$$p(n_k) = p_d^{N-n_k}(1 - p_d)^{n_k}. \tag{6}$$

We will need the product of $p(n_k)$ for $k = 1, \ldots, \ell$:

$$\prod_{k=1}^{\ell} p(n_k) = p_d^{N\ell - L^{\text{tot}}}(1 - p_d)^{L^{\text{tot}}}, \tag{7}$$

where L^{tot} is the total number of events in the N point processes $(x_i)_{i=1,\ldots,N}$:

$$L^{\mathrm{tot}} = \sum_{i=1}^{N} L_i. \tag{8}$$

At last, the resulting N sequences are shifted over $(\delta_i)_{i=1,\ldots,N} = (\delta_{ti}, \delta_{fi})_{i=1,\ldots,N}$, and the occurrence times and frequencies are slightly perturbed, resulting in the sequences $(x_i)_{i=1,\ldots,N}$. As we pointed out earlier, there might be a non-trivial timing and frequency offset between the bump models (see Fig. 3 and Fig. 4). The parameters $(\delta_{ti}, \delta_{fi})$ are introduced in the model to account for such offsets. The offsets between v and $(x_i)_{i=1,\ldots,N}$ may be modeled as bivariate Gaussian random variables with mean vectors $(\delta_{ti}, \delta_{fi})$ and diagonal non-isotropic covariance matrices $V_i = \mathrm{diag}(s_{ti}, s_{fi})$. It is reasonable to assume that the offsets in time are independent from the offsets in frequency, and vice versa. Therefore, we use diagonal matrices V_i. Statistical dependencies between the perturbations in time and frequency may be modeled by non-diagonal covariance matrices V_i. Such extensions are straightforward, and we do not consider them here.

We adopt the improper priors $p(\delta_{ti}) = 1 = p(\delta_{fi})$ for $(\delta_{ti})_{i=1,\ldots,N}$ and $(\delta_{fi})_{i=1,\ldots,N}$ respectively, and conjugate priors for s_{ti} and s_{fi}, i.e. scaled inverse chi-square distributions:

$$p(s_{ti}) = \frac{(s_{t0}\nu_t/2)^{\nu_t/2}}{\Gamma(\nu_t/2)} \frac{e^{-\nu_t s_{t0}/2s_{ti}}}{s_{ti}^{1+\nu_t/2}} \tag{9}$$

$$p(s_{fi}) = \frac{(s_{f0}\nu_f/2)^{\nu_f/2}}{\Gamma(\nu_f/2)} \frac{e^{-\nu_f s_{f0}/2s_{fi}}}{s_{fi}^{1+\nu_f/2}}, \tag{10}$$

where ν_t and ν_f are the degrees of freedom, and s_{t0} and s_{f0} are the width of the scaled inverse chi-square distributions, and $\Gamma(x)$ is the Gamma function.

In (Dauwels *et al.*, 2009b) we normalized the parameters (δ_t, s_t) and (δ_f, s_f) by the width and height of the bumps respectively, in order to take the size of the bumps into account. For simplicity, we will discard such normalization factors in the following. They can easily be incorporated in the statistical model, and we will briefly address this issue in Section 5.

It is also noteworthy that in (Dauwels *et al.*, 2009b) the variance of the time and frequency perturbations in the generative process is defined as $s_t/2$ and $s_f/2$ respectively (instead of s_t and s_f), so that the variance between the two observed sequences x_1 and x_2 is given by s_t and s_f. Therefore, when comparing results from bivariate and N-variate SES, a factor of two needs to be taken into account.

For later convenience, we will introduce some more notation. We denote by $v_{c_{ij}}$ the hidden event that generated x_{ij} (the jth event in point process x_i).

The function c is hence a clustering function, that groups the events x_{ij} into different clusters. Since there is at most one event from point process i in cluster k, the clustering function c fulfills the constraints:

$$\sum_{j=1}^{L_i} \delta\Big[c_{ij} - k\Big] \leq 1, \forall i, k. \tag{11}$$

Note that certain hidden events $(v_k)_{k=1,\ldots,\ell}$ may not have any copies, since all N copies may have been deleted. Therefore, the function c does not necessarily take ℓ different values. Without loss of generality, we will assume that c takes values in $\{1, 2, \ldots, L\}$, where L is the number of clusters and $0 \leq L \leq \ell$. Note that the number L of clusters is at most L^{tot}, i.e., the total number of events; this maximum number occurs when each event is a cluster, and hence all clusters are of size 1. With this definition of L, the number of copies n_1, $\ldots$, n_L are non-zero, whereas $n_{L+1} = n_{L+2} = \cdots = n_\ell = 0$. We introduce the index set $\mathcal{K}$ of clusters with $n_k > 1$:

$$\mathcal{K} = \{k \in \{1, 2, \ldots, L\} : n_k > 1\}. \tag{12}$$

We denote by $\mathcal{C}_k$ the set of n_k copies of v_k, and denote its index set by $\mathcal{I}_k$:

$$\mathcal{C}_k = \{x_{ij} : c_{ij} = k\} \text{ and } \mathcal{I}_k = \{(i, j) : c_{ij} = k\}. \tag{13}$$

The fraction ρ of missing events in the clusters can be computed as:

$$\rho = 1 - \frac{\sum_{k=1}^{L} n_k}{LN} = 1 - \frac{\bar{n}}{N}, \tag{14}$$

where $\bar{n}$ is the average number of events per cluster. Another important statistics is the distribution $(p_j)_{j=1}^{N}$ of the number of events per cluster:

$$p_j = \frac{\sum_{k=1}^{L} \delta[n_k - j]}{L}, \quad j = 1, 2, \ldots, N \tag{15}$$

In this notation, the overall probabilistic model may be written as:

$$\begin{aligned} p(x, c, v, \theta, \ell) \propto\, & p(s_t)p(s_f)(1 - \lambda \,\mathrm{vol}(S))(\lambda \, p_d^N)^\ell p_d^{-L^{\text{tot}}}(1 - p_d)^{L^{\text{tot}}} \\ & \cdot \prod_{i=1}^{N} \prod_{j=1}^{L_i} \mathcal{N}\Big(t_{ij} - \tilde{t}_{c_{ij}}; \delta_{ti}, s_{ti}\Big) \mathcal{N}\Big(f_{ij} - \tilde{f}_{c_{ij}}; \delta_{fi}, s_{fi}\Big). \end{aligned} \tag{16}$$

Note that the parameters N, $(L_i)_{i=1,\ldots,N}$ and L^{tot} are fixed for given point processes. Likewise, for given clustering c, the parameters L and $(n_k)_{k=1,\ldots,L}$ are fixed. The total number of deletions is given by $L^{\text{del,tot}} = N\ell - L^{\text{tot}}$. The number of hidden events v_k without copies is given by $L^{\text{del}} = \ell - L$.

As in Part I and II, we can marginalize the statistical model $p(x, c, v, \theta, \ell)$ *analytically* w.r.t. v and ℓ (Dauwels *et al.*, 2009a,b), resulting in $p(x, c, \theta)$ (cf.

Appendix A):

$$\begin{aligned} p(x, c, \theta) \propto\, & \gamma\, \beta^{NL} p(s_t) p(s_f) \\ & \cdot \prod_{k \in \mathcal{K}} \prod_{(i,j) \in \mathcal{I}_k} \mathcal{N}\Big(t_{ij} - \bar{t}_k; \delta_{ti}, s_{ti}\Big)\, \mathcal{N}\Big(f_{ij} - \bar{f}_k; \delta_{fi}, s_{fi}\Big), \end{aligned} \tag{17}$$

where

$$\bar{t}_k = \frac{\sum\limits_{(i,j) \in \mathcal{I}_k} w_{ti}(t_{ij} - \delta_{ti})}{\sum\limits_{(i,j) \in \mathcal{I}_k} w_{ti}} \tag{18}$$

$$\bar{f}_k = \frac{\sum\limits_{(i,j) \in \mathcal{I}_k} w_{fi}(f_{ij} - \delta_{fi})}{\sum\limits_{(i,j) \in \mathcal{I}_k} w_{fi}}, \tag{19}$$

with $w_{ti} = s_{ti}^{-1}$ and $w_{fi} = s_{fi}^{-1}$. Interestingly, the parameters $(\bar{t}_k, \bar{f}_k)$ may be interpreted as the coordinates of the center of the n_k copies associated with v_k (see Fig. 5(b)); in other words, those n_k copies may be viewed as a cluster of events, whose center is located at $(\bar{t}_k, \bar{f}_k)$. As easily can be shown, the latter parameters are also the maximum likelihood (ML) estimates of the time and frequency of the hidden event v_k, when the copies of v_k and the parameters (δ_i, s_i) are given $(i = 1, 2 \ldots, N)$. In practice, the latter are usually not given, and therefore, the coordinates $\bar{t}_k$ and $\bar{f}_k$ depend on how the events x_{ij} are assigned to hidden events v_k; in other words, those parameters depend on the clustering c.

The parameters β and γ in (17) are defined as:

$$\beta = p_d\, \sqrt[N]{\lambda}, \tag{20}$$

and

$$\gamma = \Big(p_d^{-1}\,(1 - p_d)\Big)^{L^{\mathrm{tot}}} (1 - \lambda \mathrm{vol}(S))\, \frac{1}{1 - \lambda \mathrm{vol}(S) p_d^N}. \tag{21}$$

Note that we defined similar parameters γ and β in Part I and II (Dauwels *et al.*, 2009a,b); the N-variate statistical model (17) is a natural extension of the pairwise statistical models of (Dauwels *et al.*, 2009a,b). The constant γ does not depend on c or the SES parameters δ_i and s_i (with $i = 1, \ldots, N$), and therefore, it is irrelevant for estimating the latter parameters and the clusters; we will discard γ in the following.

So far, we have not yet taken background events into account. They can be modeled as follows. Besides the hidden process v, we generate the background events as a point process $\tilde{v}$ of length $\tilde{\ell}$. We define the prior $p(\tilde{v}, \tilde{\ell})$ similarly as

$p(v, \ell)$ (5):

$$p(\tilde{v}, \tilde{\ell}) = p(\tilde{\ell}) p(\tilde{v} | \tilde{\ell}) \propto (1 - \tilde{\lambda} \, \mathrm{vol}(S)) \tilde{\lambda}^{\tilde{\ell}}. \tag{22}$$

As a result, some of the events x are generated from v (according to Step 1 to 3), and the other events (background events) are the process $\tilde{v}$. We will denote by χ the fraction of background events in x. To account for the background events, we will now assume that c takes values in $\{0, 1, 2, \ldots, L\}$, where $c_{ij} = 0$ iff x_{ij} is a background event. From a given clustering function c, we can easily infer the background events (and hence also $\tilde{\ell}$). An important statistics is the fraction χ of background events, which can also easily be computed form c.

We can include background events in statistical model (17) by multiplying it with the prior $p(\tilde{v}, \tilde{\ell})$ (22), resulting in:

$$\begin{aligned} p(x, c, \theta) \propto\, & \beta^{NL} \tilde{\beta}^{\tilde{\ell}} p(s_t) p(s_f) \\ & \cdot \prod_{k \in \mathcal{K}} \prod_{(i,j) \in \mathcal{I}_k} \mathcal{N}\Big(t_{ij} - \bar{t}_k; \delta_{ti}, s_{ti}\Big) \mathcal{N}\Big(f_{ij} - \bar{f}_k; \delta_{fi}, s_{fi}\Big), \end{aligned} \tag{23}$$

where the parameter $\tilde{\beta} = \tilde{\lambda}$ (cf. (20)).

The exponent of β and $\tilde{\beta}$ in (23) does clearly depend on c, and as a result, the parameters β and $\tilde{\beta}$ affect the inference of c and the SES parameters. As in Part I and II, we will interpret those parameters in terms of cost functions; the expressions $\log \beta$ and $\log \tilde{\beta}$ are part of the cost associated to each cluster and background event respectively. In all our experiments, we found that the setting $\tilde{\beta} = 10^{-20}$ yields satisfactory results.

3.3 Interpretation in Terms of Cost Functions

We can gain additional insight by considering the logarithm of statistical model (23):

$$\begin{aligned} -\log p(x, c, \theta) = & -\log p(s_t) - \log p(s_f) - LN \log \beta - \tilde{\ell} \log \tilde{\beta} \\ & + \sum_{k \in \mathcal{K}} \sum_{(i,j) \in \mathcal{I}_k} \Big(\frac{1}{2} \log 2\pi s_{ti} + \frac{1}{2 s_{ti}} (t_{ij} - \bar{t}_k - \delta_{ti})^2 \\ & + \frac{1}{2} \log 2\pi s_{fi} + \frac{1}{2 s_{fi}} (f_{ij} - \bar{f}_k - \delta_{fi})^2 \Big) + \zeta, \end{aligned} \tag{24}$$

where ζ is an irrelevant constant. The expression (24) may be considered as a cost function that associates certain costs with each event and cluster; we provided a similar viewpoint in Part I and II (Dauwels *et al.*, 2009a,b). The unit cost d_0 associated with each of the L clusters is given by:

$$d_0 = -N \log \beta. \tag{25}$$

Likewise, the unit cost $\tilde{d}_0$ associated with each of the background events is given by:

$$\tilde{d}_0 = -\log \tilde{\beta}. \tag{26}$$

The unit cost of each event x_{ij} associated with a cluster k of size $n_k > 1$ equals:

$$\begin{aligned} d(t_{ij}, f_{ij}; c, \theta) =& \frac{1}{2} \log 2\pi s_{ti} + \frac{1}{2s_{ti}} (t_{ij} - \bar{t}_k - \delta_{ti})^2 \\ &+ \frac{1}{2} \log 2\pi s_{fi} + \frac{1}{2s_{fi}} (f_{ij} - \bar{f}_k - \delta_{fi})^2. \end{aligned} \tag{27}$$

This cost depends on the choice c of clusters and on the parameters θ. Indeed, the parameters $\bar{t}_k$ and $\bar{f}_k$ are dependent on c and θ as follows:

$$\bar{t}_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{L_i} \delta[c_{ij} - k] \, w_{ti} (t_{ij} - \delta_{ti})}{\sum_{i=1}^{N} \sum_{j=1}^{L_i} \delta[c_{ij} - k] \, w_{ti}} \tag{28}$$

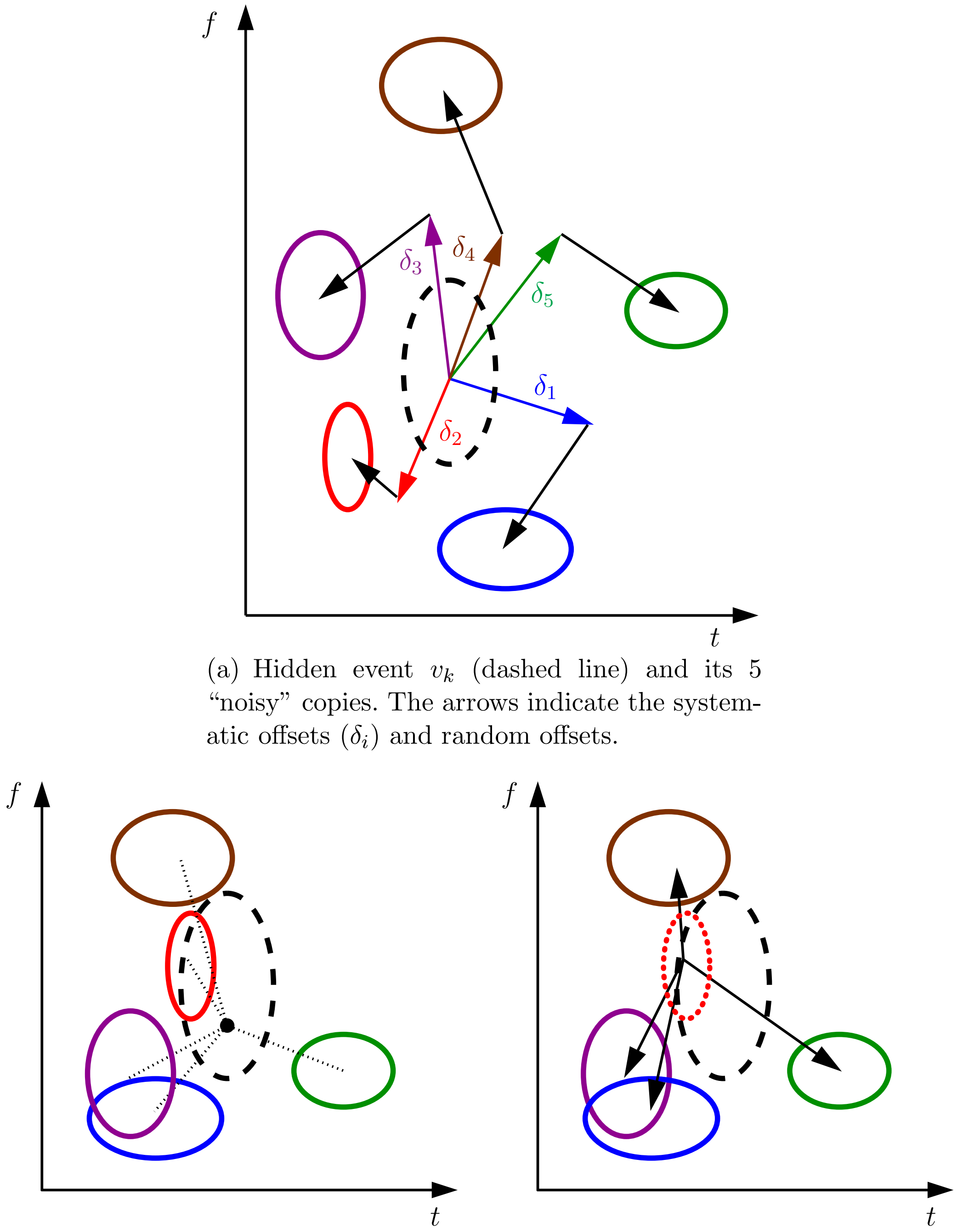
$$\bar{f}_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{L_i} \delta[c_{ij} - k] \, w_{fi} (f_{ij} - \delta_{fi})}{\sum_{i=1}^{N} \sum_{j=1}^{L_i} \delta[c_{ij} - k] \, w_{fi}}, \tag{29}$$

with $w_{ti} = s_{ti}^{-1}$ and $w_{fi} = s_{fi}^{-1}$. The distance $d(t_{ij}, f_{ij})$ (27) is illustrated in Fig. 5(b).

Note that the second and fourth term in the RHS of (27) are normalized Euclidian distances. Since the point processes $(x_i)_{1,\ldots,N}$ are defined on the time-frequency plane (see Fig. 4), the (normalized) Euclidean distance is indeed a natural metric. In some applications, the point process may be defined on more general spaces, in particular, curved spaces; in such situations, one may adopt non-Euclidean distance measures. We refer to (Dauwels *et al.*, 2008) for an example. To simplify the notation, we will define the unit cost of each event x_{ij} associated with a cluster k of size $n_k = 1$ as $d(t_{ij}, f_{ij}; c, \theta) = 0$.

We also define costs $d(s_t) = -\log p(s_t)$, $d(s_f) = -\log p(s_f)$, and $d(\theta) = d(s_t) + d(s_f)$. With those definitions of unit costs, we can rewrite (24) as:

$$-\log p(x, c, \theta) = d(\theta) + L d_0 + \tilde{\ell} \tilde{d}_0 + \sum_{k=1}^{L} \sum_{(i,j) \in \mathcal{I}_k} d(t_{ij}, f_{ij}; c, \theta) + \zeta. \tag{30}$$



(a) Hidden event v_k (dashed line) and its 5 “noisy” copies. The arrows indicate the systematic offsets (δ_i) and random offsets.

(b) For the sake of clarity, the offsets δ_i (cf. (a)) have been eliminated. The dot corresponds to the center of the 5 events (cf. (28) (29)), and the dotted lines indicate the distances $d(t_{ij}, f_{ij})$ (27).

(c) After eliminating the offsets δ_i, event 2 (dotted) lies the closest to the hidden event (dashed), and it serves as exemplar. The arrows indicate the distances $d(t_{ij}, f_{ij}, t_{i'j'}, f_{i'j'})$ (48).

Fig. 5. Exemplar representation. A cluster with 5 events, generated from a hidden event (dashed lines). For the sake of clarity, we have eliminated the offsets δ_i in (b) and (c). The exemplar is marked in dotted lines in (c).

This expression can be written as a function of c as follows:

$$-\log p(x,c,\theta) = d(\theta) + d_0 \max_{ij} c_{ij} + \tilde{d}_0 \sum_{ij} \delta[c_{ij}] \\ + \sum_{k=1}^{\max_{ij} c_{ij}} \sum_{i=1}^{N} \sum_{j=1}^{L_i} \delta[c_{ij} - k]\, d(t_{ij}, f_{ij}; c, \theta) + \tilde{\zeta}, \tag{31}$$

where $d(t_{ij}, f_{ij}; c, \theta)$ is given by (27). Clearly, the RHS of (31) depends on c in a non-linear fashion (cf. (27), (28) and (29)).

4 Statistical Inference

A reasonable approach to infer (c,θ) is maximum a posteriori (MAP) estimation:

$$(\hat{c}, \hat{\theta}) = \underset{(c,\theta)}{\operatorname{argmax}} \log p(x,c,\theta), \tag{32}$$

subject to (11). There is no closed form expression for (32), therefore, we need to resort to numerical methods. A simple technique to try to find (32) is cyclic maximization: We first choose initial values $\hat{\theta}^{(0)}$, and then perform the following updates for $\kappa \geq 1$ until convergence:

$$\hat{c}^{(\kappa)} = \underset{c}{\operatorname{argmax}} \log p(x, c, \hat{\theta}^{(\kappa-1)}) \tag{33}$$

$$\hat{\theta}^{(\kappa)} = \underset{\theta}{\operatorname{argmax}} \log p(x, \hat{c}^{(\kappa)}, \theta), \tag{34}$$

where (33) is determined subject to (11). The update (34) of the parameters θ is straightforward, and may be carried out by cyclic maximization; we refer to Appendix B for more details. The update (33) is far less straightforward; it involves an intractable optimization problem. We circumvent this issue by solving a related tractable optimization problem. In the following, we describe that problem.

The update (33) may be expanded as:

$$\hat{c}^{(\kappa)} = \underset{c}{\operatorname{argmin}} \Big(d_0 \max_{ij} c_{ij} + \tilde{d}_0 \sum_{ij} \delta[c_{ij}] \\ + \sum_{k=1}^{\max_{ij} c_{ij}} \sum_{i=1}^{N} \sum_{j=1}^{L_i} \delta[c_{ij} - k]\, d(t_{ij}, f_{ij}; c, \hat{\theta}^{(\kappa-1)}) \Big), \tag{35}$$

which is also determined subject to (11).

The optimization problem (35) is hard to solve directly, and therefore, we will solve a related tractable optimization problem instead. In order to formulate the latter, we introduce the following binary variables:

- b_k is equal to one iff cluster k is non-empty ($n_k > 0$ and $k = 1, 2, \ldots, L^{\text{tot}}$),
- b_{ijk} is equal to one iff the x_{ij} belongs to cluster k, i.e., $c_{ij} = k$,
- e_{ij} is equal to one iff x_{ij} is a background event.

We will now rewrite (35) using that notation, which will allow us to simplify the combinatorial problem in Section 4.2.

The binary variables b are related through the constraints:

$$b_k = \min\left(1, \sum_{i=1}^{N}\sum_{j=1}^{L_i} b_{ijk}\right), \quad \forall k. \tag{36}$$

These non-linear inequality constraints are equivalent to the linear constraints:

$$b_{ijk} \leq b_k, \quad \forall i, j, k, \quad \text{ and } \sum_{i=1}^{N}\sum_{j=1}^{L_i} b_{ijk} \geq b_k, \quad \forall k. \tag{37}$$

The constraints (11) correspond to:

$$\sum_{j=1}^{L_i} b_{ijk} \leq 1, \quad \forall i, k. \tag{38}$$

Moreover, an event is either a background event, or belongs to a cluster, which can be encoded by the constraints:

$$\sum_{k}^{L} b_{ijk} + e_{ij} = 1, \quad \forall i, j, \tag{39}$$

In this representation, we can rewrite (35) as:

$$\begin{aligned}\left(\hat{b}^{(\kappa)}, \hat{e}^{(\kappa)}\right) = \underset{b,e}{\operatorname{argmin}} \Bigg(& d_0 \sum_{k=1}^{L^{\text{tot}}} b_k + \tilde{d}_0 \sum_{i=1}^{N}\sum_{j=1}^{L_i} e_{ij} \\ & + \sum_{k=1}^{L^{\text{tot}}}\sum_{i=1}^{N}\sum_{j=1}^{L_i} b_{ijk}\, d(t_{ij}, f_{ij}; b, \hat{\theta}^{(\kappa-1)}) \Bigg),\end{aligned} \tag{40}$$

subject to (37)–(39). As we pointed out earlier, the number of clusters is at most L^{tot}, i.e., the total number of events.

If x_{ij} is associated with a cluster k of size $n_k > 1$, and hence $\sum_{i=1}^N \sum_{j=1}^{L_i} b_{ijk} > 1$, the expression $d(t_{ij}, f_{ij}; b, \hat{\theta}^{(\kappa-1)})$ in (40) is given by (27) (see Fig. 5(b)), with $\theta = \hat{\theta}^{(\kappa-1)}$ and:

$$\bar{t}_k = \frac{\sum_{i=1}^N \sum_{j=1}^{L_i} b_{ijk}\, \hat{w}_{ti}^{(\kappa-1)}(t_{ij} - \hat{\delta}_{ti}^{(\kappa-1)})}{\sum_{i=1}^N \sum_{j=1}^{L_i} b_{ijk}\, \hat{w}_{ti}^{(\kappa-1)}} \tag{41}$$

$$\bar{f}_k = \frac{\sum_{i=1}^N \sum_{j=1}^{L_i} b_{ijk}\, \hat{w}_{fi}^{(\kappa-1)}(f_{ij} - \hat{\delta}_{fi}^{(\kappa-1)})}{\sum_{i=1}^N \sum_{j=1}^{L_i} b_{ijk}\, \hat{w}_{fi}^{(\kappa-1)}}, \tag{42}$$

where $\hat{w}_{ti}^{(\kappa-1)} = \left(\hat{s}_{ti}^{(\kappa-1)}\right)^{-1}$ and $\hat{w}_{fi}^{(\kappa-1)} = \left(\hat{s}_{fi}^{(\kappa-1)}\right)^{-1}$; otherwise $d(t_{ij}, f_{ij}; b, \hat{\theta}^{(\kappa-1)}) = 0$.

By exponentiating the objective function in (40), and adding the priors in θ, we obtain the statistical model:

$$p(x, b, e, \hat{\theta}^{(\kappa-1)}) \propto p(\hat{s}_t^{(\kappa-1)})p(\hat{s}_f^{(\kappa-1)}) \prod_{k=1}^{L^{\text{tot}}} \left(\beta^N\right)^{b_k} \prod_{i=1}^{N} \prod_{j=1}^{L_i} \tilde{\beta}^{e_{ij}}$$
$$\cdot \prod_{k=1}^{L^{\text{tot}}} \prod_{i=1}^{N} \prod_{j=1}^{L_i} \left(\mathcal{N}\!\left(t_{ij} - \bar{t}_k; \delta_{ti}, s_{ti}\right) \mathcal{N}\!\left(f_{ij} - \bar{f}_k; \delta_{fi}, s_{fi}\right)\right)^{b_{ijk}}, \tag{43}$$

which is equivalent to (23). Likewise, the constrained combinatorial optimization problem (37)–(40) is equivalent to (35); it is a non-linear combinatorial optimization problem, since the objective function (RHS of (40)) is non-linear in b. Since the problem is *intractable*, we will simplify it: we linearize the objective function by using an exemplar representation. The resulting linear combinatorial optimization problem (a.k.a. integer linear program or ILP) can then be solved exactly by integer linear programming. In other words, we will approximate the original non-linear and intractable integer program (37)–(40) (which is hard to solve directly) into a linear and tractable integer program (which is much easier to solve).

4.2 Related (Tractable) Optimization Problem

Ideally, we wish to find the cluster centers $(\bar{t}_k, \bar{f}_k)$ (41) (42) that minimize the combined total cost $d(t_{ij}, f_{ij})$ (27) of all events in the cluster. That is an intractable problem, and we simplify it as follows. Rather than searching through all possible cluster centers, we only consider events x_{ij} as potential cluster centers. In other words, we consider a restricted subset of potential centers. More specifically, we approximate the cluster center $(\bar{t}_k, \bar{f}_k)$ (41) (42) by the event x_{ij} of the same cluster that lies the closest to the center, after eliminating the offset $(\delta_{ti}, \delta_{fi})$ (see Fig. 5(c)). The events x_{ij} that serve as cluster centers are referred to as "exemplars". This approach is inspired by other exemplar-based clustering algorithms, including affinity propagation (Frey *et al.*, 2007) and recent extensions (Lashkari *et al.*, 2008; Givoni *et al.*, 2009).

As a result, the non-linear cost $d(t_{ij}, f_{ij})$ (27) is approximated by a cost that is independent of b; the non-linear objective function (RHS of (40)) becomes linear in b, and hence the non-linear combinatorial optimization problem (40) is approximated by an integer linear program. We now derive this integer linear program. Similarly to the variables b_k and b_{ijk}, we introduce the following binary variables:

- b_{ij} is equal to one iff x_{ij} is an exemplar,
- $b_{iji'j'}$ is equal to one iff x_{ij} is associated with exemplar $x_{i'j'}$.

In this formulation, we approximate the intractable optimization problem (37)–(40) by the following integer linear program in b:

$$\begin{aligned}\left(\hat{b}^{(\kappa)}, \hat{e}^{(\kappa)}\right) = \operatorname*{argmin}_{b,e} \Big(d_0 \sum_{i=1}^{N} \sum_{j=1}^{L_i} b_{ij} + \tilde{d}_0 \sum_{i=1}^{N} \sum_{j=1}^{L_i} e_{ij} \\ + \sum_{i,i'=1}^{N} \sum_{j=1}^{L_i} \sum_{j'=1}^{L_{i'}} b_{iji'j'}\, d(t_{ij}, f_{ij}, t_{i'j'}, f_{i'j'}; \hat{\theta}^{(\kappa-1)}) \Big), \end{aligned} \tag{44}$$

subject to

$$\sum_{i'j'} b_{iji'j'} + b_{ij} + e_{ij} = 1, \quad \forall i, j, \tag{45}$$

$$\sum_{j=1}^{L_i} b_{iji'j'} \leq b_{i'j'}, \quad \forall i, i' \neq i, j', \tag{46}$$

$$b_{ijij'} = 0, \quad \forall i, j, j', \tag{47}$$

where

$$\begin{aligned} d(t_{ij}, f_{ij}, t_{i'j'}, f_{i'j'}; \hat{\theta}^{(\kappa-1)}) = \frac{1}{2} \log 2\pi \hat{s}_{ti}^{(\kappa-1)} + \frac{1}{2\hat{s}_{ti}^{(\kappa-1)}} (t_{ij} - t_{i'j'} - \hat{\delta}_{ti}^{(\kappa-1)})^2 \\ + \frac{1}{2} \log 2\pi \hat{s}_{fi}^{(\kappa-1)} + \frac{1}{2\hat{s}_{fi}^{(\kappa-1)}} (f_{ij} - f_{i'j'} - \hat{\delta}_{fi}^{(\kappa-1)})^2. \end{aligned} \tag{48}$$

By exponentiating the objective function in (44), and adding the priors in θ, we obtain the statistical model:

$$p(x,b,e,\hat{\theta}^{(\kappa-1)}) \propto p(\hat{s}_t^{(\kappa-1)})p(\hat{s}_f^{(\kappa-1)}) \prod_{i=1}^{N}\prod_{j=1}^{L_i} \left(\beta^N\right)^{b_{ij}} \tilde{\beta}^{e_{ij}}$$
$$\cdot \prod_{i,i'=1}^{N}\prod_{j=1}^{L_i}\prod_{j'=1}^{L_{i'}} \Big(\mathcal{N}\big(t_{ij}-t_{i'j'};\hat{\delta}_{ti}^{(\kappa-1)},\hat{s}_{ti}^{(\kappa-1)}\big)\mathcal{N}\big(f_{ij}-f_{i'j'};\hat{\delta}_{fi}^{(\kappa-1)},\hat{s}_{fi}^{(\kappa-1)}\big)\Big)^{b_{iji'j'}}. \tag{49}$$

The objective function (44) is a linear approximation of the non-linear objective function (35), and similarly, the statistical model (49) is an approximation of $p(x,c,\hat{\theta}^{(\kappa-1)})$ (23). The resulting linear integer problem (44)–(47) is much easier to solve than the non-linear integer problem (37)–(40). Note that both problems lead to similar results, since exemplars are often close to the cluster center (cf. Fig. 5(c)).

The sum $\sum_{ij} b_{ij}$ in (44) is equal to the number of exemplars; therefore, the first term in (44) assigns a cost d_0 to each exemplar. Likewise, the second term in (44) assigns a cost $\tilde{d}_0$ to each background event. The third term in (44) associates the cost (48) to each event x_{ij}, based on its associated exemplar $x_{i'j'}$. This cost is independent of b, and consequently, the objective function (44) is linear in b.

The constraints (45) ensure that each event is either an exemplar, or is associated with one exemplar (and not more than *one* exemplar), or is a background event (insertion). The constraints (46) encode the fact that an event x_{ij} can only be associated to an exemplar $x_{i'j'}$ ($b_{iji'j'} = 1$) iff the latter is indeed an exemplar ($b_{i'j'} = 1$); they also ensure that at most one event x_{ij} from x_i can be associated with an exemplar $x_{i'j'}$. Finally, the constraints (47) ensure that an event x_{ij} cannot be associated with an exemplar from the same point process x_i; without those constraints, multiple events from the same point process x_i may belong to the same cluster, which is not allowed.

The combinatorial optimization problem (44)–(47) is an integer linear program in b and e, since the objective function (44) and constraints (45)–(47) are linear in the variables b and e. More specifically, it is a *binary* linear program, since all variables are binary. Instead of solving the intractable problem (35) or equivalently (37)–(40), we solve the tractable problem (44)–(47); in particular, we solve it by means of off-the-shelf integer programming software.

As an alternative, we have also implemented the max-product algorithm (and various refinements) to solve (44)–(47), similarly as for bivariate SES (Dauwels *et al.*, 2009a,b). Unfortunately, that approach leads to poor results for N-variate SES. In particular, it does not always convergence, and sometimes it yields solutions that violate the constraints (45)–(47); those issues might be due to the fact that N-wise alignment is significantly more complex than pairwise alignment.

Note that it is straightforward to determine estimates $\hat{\rho}$, $(\hat{p}_j)_{j=1}^N$, and $\hat{\chi}$ from $\hat{b}$ and $\hat{e}$. Also, an estimate $\hat{c}$ can easily be computed from $\hat{b}$ and $\hat{e}$, as follows. We number all exemplars from 1 to L, in arbitrary order. We then set $\hat{c}_{ij} = k$ if x_{ij} is the kth exemplar or if it is associated with the kth exemplar. Likewise we set $\hat{c}_{ij} = 0$ if x_{ij} is a background event ($\hat{e}_{ij} = 1$). The resulting estimate $\hat{c}$ is an approximation of (35). With this estimate of $\hat{c}$, we can eventually refine the estimate θ, following rule (34).

The resulting SES inference algorithm is summarized Table 2.

5 Extensions

So far, we have developed N-variate SES for the particular example of bump models in time-frequency domain. The statistical model (23), and equivalently the cost function (31), may easily be generalized, and it may be applied to different kinds of point processes. One simply needs to define the cost functions d in (31) in a suitable manner. We will briefly outline several potential extensions and alternative applications.

- As we pointed out earlier, in (Dauwels *et al.*, 2009b) we normalized the parameters (δ_t, s_t) and (δ_f, s_f) by the width and height of the bumps respectively, in order to take the size of the bumps into account. Such normalization factors can easily be incorporated in cost function (31); the unit cost $d(t_{ij}, f_{ij}; c, \theta)$ of an event x_{ij} is then defined as:

$$\begin{aligned} d(t_{ij}, f_{ij}; c, \theta) = & \frac{1}{2}\log 2\pi \bar{s}_{ti} + \frac{1}{2\bar{s}_{ti}}(t_{ij} - \bar{t}_k - \bar{\delta}_{ti})^2 \\ & + \frac{1}{2}\log 2\pi \bar{s}_{fi} + \frac{1}{2\bar{s}_{fi}}(f_{ij} - \bar{f}_k - \bar{\delta}_{fi})^2, \end{aligned} \tag{50}$$

  with $\bar{\delta}_{ti} = \delta_{ti}\,\Delta t_k$, $\bar{\delta}_{fi} = \delta_{fi}\,\Delta f_k$, $\bar{s}_{ti} = s_{ti}\,\Delta t_k^2$, and $\bar{s}_{fi} = s_{fi}\,\Delta f_k^2$, where Δt_k and Δf_k are the average width and height respectively of the bumps x_{ij} in cluster k.
- As outlined in (Dauwels *et al.*, 2009b, Section 6), one can easily incorporate differences in amplitude, width and height between the bumps of the different point processes. Moreover, the bumps may be oblique, i.e., they are not necessarily parallel to the time and frequency axes.
- Until now we have considered bump models in the time-frequency domain. However, the statistical model (23), and equivalently the cost function (31), is readily extendable to point processes in other Euclidean spaces, e.g., three-dimensional spatial point processes or one-dimensional point processes in

INPUT:
One-dimensional or multi-dimensional point processes $(x_i)_{i=1}^N$ and parameters β, $\tilde{\beta}$, ν_t, ν_f, s_{t0}, s_{f0}, $(\hat{\delta}_{ti}^{(0)})_{i=1}^N$, $(\hat{\delta}_{fi}^{(0)})_{i=1}^N$, $(\hat{s}_{ti}^{(0)})_{i=1}^N$, $(\hat{s}_{fi}^{(0)})_{i=1}^N$.

ALGORITHM:
Iterate the following two steps until convergence or the available time has elapsed:

(1) Update the clustering $(\hat{b}, \hat{e})$ (and equivalently $\hat{c}$) by ILP:

$$\begin{aligned}\big(\hat{b}^{(\kappa)}, \hat{e}^{(\kappa)}\big) = \operatorname*{argmin}_{b,e} \big(d_0 \sum_{i=1}^{N} \sum_{j=1}^{L_i} b_{ij} + \tilde{d}_0 \sum_{i=1}^{N} \sum_{j=1}^{L_i} e_{ij} \\ + \sum_{i,i'=1}^{N} \sum_{j=1}^{L_i} \sum_{j'=1}^{L_{i'}} b_{iji'j'}\, d(t_{ij}, f_{ij}, t_{i'j'}, f_{i'j'}; \hat{\theta}^{(\kappa-1)})\big),\end{aligned}$$

subject to (45)–(47).

(2) Update the SES parameters:

Solve the equations:

$$\begin{aligned}\hat{\delta}_{ti}^{(\kappa)} &= \frac{1}{L_i} \sum_{j=1}^{L_i} \big(t_{ij} - \bar{t}_{\hat{c}_{ij}^{(\kappa)}}\big) \\ \hat{\delta}_{fi}^{(\kappa)} &= \frac{1}{L_i} \sum_{j=1}^{L_i} \big(f_{ij} - \bar{f}_{\hat{c}_{ij}^{(\kappa)}}\big) \\ \hat{s}_{ti}^{(\kappa)} &= \frac{\nu_t s_{t0} + L_i\, \hat{s}_{ti,\text{sample}}^{(\kappa)}}{\nu_t + L_i + 2} \\ \hat{s}_{fi}^{(\kappa)} &= \frac{\nu_f s_{f0} + L_i\, \hat{s}_{fi,\text{sample}}^{(\kappa)}}{\nu_f + L_i + 2},\end{aligned}$$

with

$$\bar{t}_k = \frac{\sum_{(i,j)\in\hat{\mathcal{I}}_k^{(\kappa)}} \hat{w}_{ti}^{(\kappa)} (t_{ij} - \hat{\delta}_{ti}^{(\kappa)})}{\sum_{(i,j)\in\hat{\mathcal{I}}_k^{(\kappa)}} \hat{w}_{ti}^{(\kappa)}} \qquad \bar{f}_k = \frac{\sum_{(i,j)\in\hat{\mathcal{I}}_k^{(\kappa)}} \hat{w}_{fi}^{(\kappa)} (f_{ij} - \hat{\delta}_{fi}^{(\kappa)})}{\sum_{(i,j)\in\hat{\mathcal{I}}_k^{(\kappa)}} \hat{w}_{fi}^{(\kappa)}}.$$

OUTPUT:
Clustering $(\hat{b}, \hat{e})$ (and equivalently $\hat{c}$) and SES parameters $\hat{\rho}$, $\hat{\chi}$, $(\hat{p}_j)_{j=1}^N$, $(\hat{\delta}_{ti})_{i=1}^N$, $(\hat{\delta}_{fi})_{i=1}^N$, $(\hat{s}_{ti})_{i=1}^N$, $(\hat{s}_{fi})_{i=1}^N$.

Table 2
Inference algorithm for N-variate SES.

time domain. We will consider an example of the latter in Section 7; the unit cost $d(t_{ij}; c, \theta)$ of an event x_{ij} may then defined as:

$$d(t_{ij}; c, \theta) = \frac{1}{2} \log 2\pi s_{ti} + \frac{1}{2s_{ti}} (t_{ij} - \bar{t}_k - \delta_{ti})^2, \tag{51}$$

where we did not take the widths of the events into account. If the latter need to be taken into account, we may normalize the parameters (δ_{ti}, s_{ti}) as in (50).

- In some applications, the point processes may be defined on *curved* manifolds, and non-Euclidean distances are then more natural. For instance, the point processes may be defined on planar closed curves. We refer to (Dauwels *et al.*, 2008) for an example. The unit cost $d(t_{ij}; c, \theta)$ of an event x_{ij} may then be defined by:

$$d(t_{ij}; c, \theta) = \frac{1}{2} \log 2\pi s_{ti} + \frac{1}{2s_{ti}} \Big(g(t_{ij}, \bar{t}_k) - \delta_{ti}\Big)^2, \tag{52}$$

where $\bar{t}_k$ is the "center" of cluster k, defined as:

$$\bar{t}_k = \operatorname{argmin}_t \sum_{(i,j) \in \mathcal{I}_k} \frac{1}{s_{ti}} \Big(g(t_{ij}, t) - \delta_{ti}\Big)^2, \tag{53}$$

where g is an arbitrary function, potentially non-linear; for $g(x, y) = x - y$, we recover (51).

6 Analysis of Synthetic Data

We investigate the robustness and reliability of N-variate SES by means of synthetic data. We consider one-dimensional and two-dimensional point processes, as in Part I (Dauwels *et al.*, 2009a) and II (Dauwels *et al.*, 2009b) respectively. We discuss the results for one-dimensional and two-dimensional point processes in Section 6.1 and 6.2 respectively.

6.1 One-Dimensional Point Processes

We randomly generated 1'000 sets of $N = 5$ one-dimensional point processes according to the generative process outlined in Section 3.

We tested several values of the parameters p_d, δ_{ti}, and s_{ti} (σ_{ti}), for $i = 1, 2, \ldots,$ 5. In particular, we tested the values $p_d = 0, 0.1, \ldots, 0.4$, and $\sigma_{ti} =$ 10ms, 30ms, and 50ms (for $i = 1, 2, \ldots, 5$), $t_{\min} =$ 0ms, and $t_{\max} = \ell_0 \cdot$ 100ms. The length ℓ was chosen as $\ell = \ell_0/(1 - p_d)$, where we tested the values $\ell_0 = 40, 100$. With this choice, the expected length of the point processes is ℓ_0, independently of p_d. In one set of experiments, we set $\delta_{ti} = 0$, for $i = 1, 2, \ldots, 5$. In a second set, the offsets δ_{ti} are drawn uniformly within $[-50\text{ms}, 50\text{ms}]$. In each case, we did not insert events (cf. Step 4).

We used the initial values $\hat{\delta}_{ti}^{(0)} = 0$ms, and $\hat{s}_{ti}^{(0)} = (20\text{ms})^2$, $(30\text{ms})^2$, for $i = 1, 2, \ldots, 5$. The parameter β was identical for all parameter settings, i.e., $\beta = 0.04$; it was optimized to yield the best overall results. We used an uninformative prior for δ_{ti} and s_{ti},, i.e., $p(\delta_{ti}) = p(s_{ti}) = 1$, for $i = 1, 2, \ldots, 5$. One could test various initial values of $\hat{\delta}_{ti}$, for $i = 1, 2, \ldots, 5$. However, the number of initial conditions grows exponentially with N, and therefore, it is not really practical to test multiple values for each $\hat{\delta}_{ti}^{(0)}$. For example, if we test 3 values for each $\hat{\delta}_{ti}^{(0)}$, we need to test a total of $3^5 = 243$ initial values $\hat{\delta}_{ti}^{(0)}$. Alternatively, one may use a small random subset of initial values; for the sake of conciseness, we do not consider that approach here.

We set $\tilde{\beta} = 10^{-20}$, as in all our simulations in this paper; for the synthetic data, no background events were inferred, i.e., $\chi = 0$ for all parameter settings.

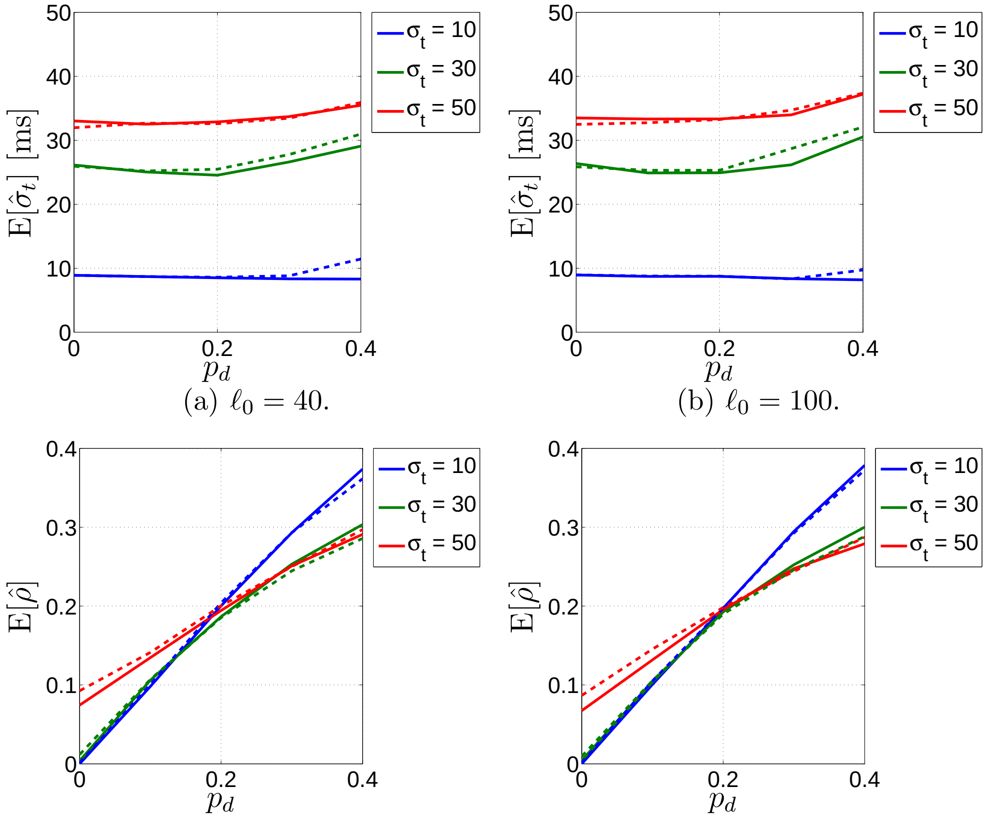
In order to assess the SES measures $S = s_t, \rho$, we compute for each above mentioned parameter setting the expectation $\mathrm{E}[S]$ and normalized standard deviation $\overline{\sigma}[S] = \sigma[S]/\mathrm{E}[S]$. Those statistics are computed by averaging over 1'000 sets of 5 point processes, randomly generated according to the generative process outlined in Section 3.

The results are summarized in Fig. 6. From this figure 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 $p_d > 0.2$. However, the bias is significantly smaller than for bivariate SES (see (Dauwels *et al.*, 2009a)).
- The estimates of s_t do only weakly depend on p_d, and vice versa.
- The estimates of s_t and p_d only weakly depend on δ_{ti} (curved vs. solid lines); they are robust to lags δ_t. Note that one could further reduce this dependency by testing various initial values $\hat{\delta}_{ti}^{(0)}$. However, the number of initial conditions grows exponentially with N, as we mentioned earlier; therefore, this approach is not really practical.
- The estimates of s_t and p_d are less biased for larger ℓ_0.

We have also observed from our experiments (not shown here):

- The estimates of δ_t are unbiased for all considered values of δ_t, s_t, and p_d.
- 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.



(a) $\ell_0 = 40$.

(b) $\ell_0 = 100$.

(c) $\ell_0 = 40$.

(d) $\ell_0 = 100$.

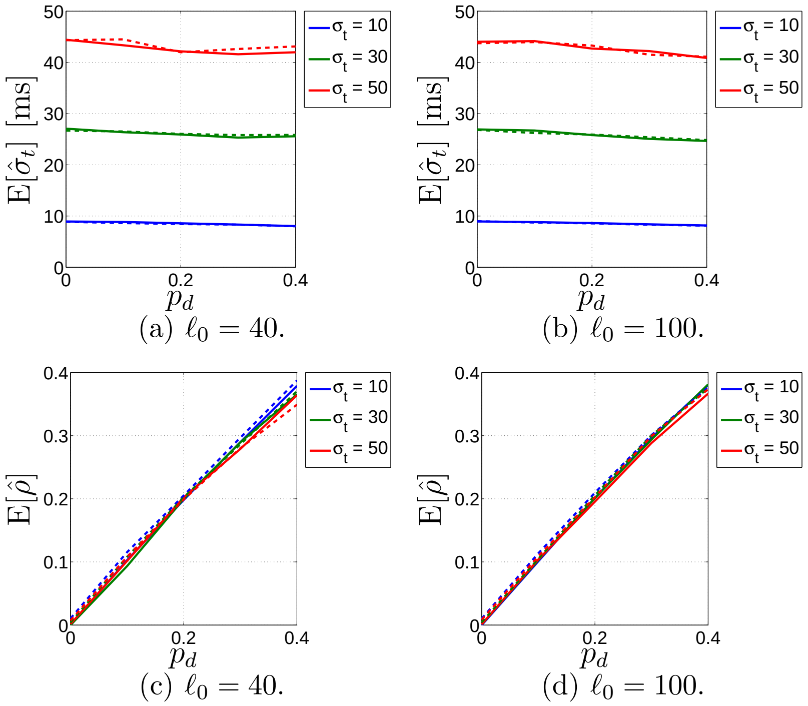
Fig. 6. Results for N-variate stochastic event synchrony for one-dimensional point processes: the figure shows the expected value E[$\hat{\sigma}_t$] and E[$\hat{\rho}$] for the parameter settings $\ell_0 = 40, 100$, $\sigma_t = 10, 30, 50$ms, and $p_d = 0, 0.1, \ldots, 0.4$. The solid lines are for zero delays δ_{ti}, whereas the dotted lines are for offsets δ_{ti} drawn uniformly within $[-50\text{ms}, 50\text{ms}]$.

- The normalized standard deviation of the SES parameters decreases as the length ℓ_0 increases, as expected.

6.2 Two-Dimensional Point Processes

Similarly as in the one-dimensional case, we randomly generated 1'000 sets of $N = 5$ two-dimensional point processes according to the generative process outlined in Section 3.

We considered several values of the parameters p_d, δ_{ti}, s_{ti} (σ_{ti}), δ_{fi}, and s_{fi} (σ_{fi}), for $i = 1, 2, \ldots, 5$. In particular, we tested the values $p_d = 0, 0.1, \ldots, 0.4$, and $\sigma_{ti} = 10$ms, 30ms, and 50ms, $\sigma_{fi} = 1$Hz, 2.5Hz, 5Hz (for $i = 1, 2, \ldots, 5$), $t_{\min} = 0$ms, and $t_{\max} = \ell_0 \cdot 100$ms, $f_{\min} = 0$Hz, and $f_{\max} = \ell_0 \cdot 1$Hz. The length ℓ was chosen as $\ell = \ell_0/(1 - p_d)$, where we tested the values $\ell_0 = 40, 100$. With this choice, the expected length of the point processes is ℓ_0, independently of



(a) $\ell_0 = 40$.

(b) $\ell_0 = 100$.

(c) $\ell_0 = 40$.

(d) $\ell_0 = 100$.

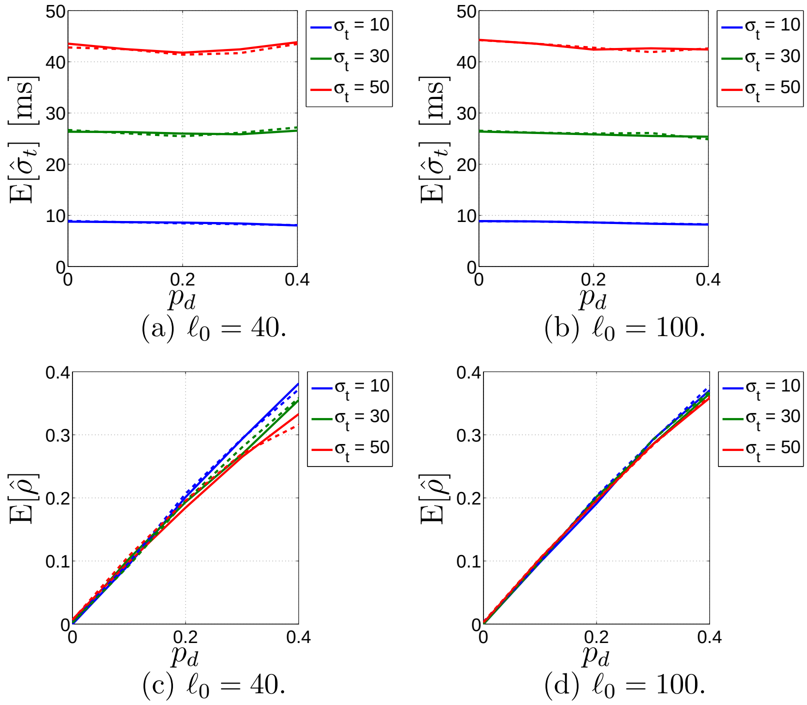
Fig. 7. Results for synthetic data: the figure shows the expected value $\mathrm{E}[\hat{\sigma}_t]$ and $\mathrm{E}[\hat{\rho}]$ and the normalized standard deviation $\bar{\sigma}[\hat{\sigma}_t]$ and $\bar{\sigma}[\hat{\rho}]$ for the parameter settings $\ell_0 = 40, 100$, $\sigma_t = 10, 30, 50$ms, $p_d = 0, 0.1, \ldots, 0.4$, and $\sigma_f = 1$Hz. The solid lines are for zero delays δ_{ti}, whereas the dotted lines are for offsets δ_{ti} and δ_{fi} drawn uniformly within $[-50\text{ms}, 50\text{ms}]$ and $[-5\text{Hz}, 5\text{Hz}]$ respectively. The curves for zero and random delays are practically coinciding.

p_d. In one set of experiments, we set $\delta_{ti} = 0 = \delta_{fi}$, for $i = 1, 2, \ldots, 5$. In a second set, the offsets δ_{ti} and δ_{fi} are drawn uniformly within $[-50\text{ms}, 50\text{ms}]$ and $[-5\text{Hz}, 5\text{Hz}]$ respectively. In each case, we did not insert events (cf. Step 4).

We used the initial values $\hat{\delta}_{ti}^{(0)} = 0$ms, $\hat{\delta}_{fi}^{(0)} = 0$Hz, $\hat{s}_{ti}^{(0)} = (20\text{ms})^2$, $(30\text{ms})^2$, and $\hat{s}_{fi}^{(0)} = (2\text{Hz})^2$, for $i = 1, 2, \ldots, 5$. The parameter β was identical for all parameter settings, i.e., $\beta = 0.01$; it was optimized to yield the best overall results. We used an uninformative prior for δ_{ti}, δ_{fi}, s_{ti}, and s_{fi}, i.e., $p(\delta_{ti}) = p(s_{ti}) = p(\delta_{fi}) = p(s_{fi}) = 1$.

The results are summarized in Fig. 7 to 9. Overall, they are quite similar to the ones for one-dimensional point processes (cf. Fig. 6). We observe the following:

- The estimates of s_t and p_d are slightly biased. However, the bias is significantly smaller than in the one-dimensional case, as for bivariate SES (cf. Fig.4 and Section 7 in (Dauwels *et al.*, 2009b)).
- 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



(a) $\ell_0 = 40$.

(b) $\ell_0 = 100$.

(c) $\ell_0 = 40$.

(d) $\ell_0 = 100$.

Fig. 8. Results for synthetic data: the figure shows the expected value $\mathrm{E}[\hat{\sigma}_t]$ and $\mathrm{E}[\hat{\rho}]$ and the normalized standard deviation $\bar{\sigma}[\hat{\sigma}_t]$ and $\bar{\sigma}[\hat{\rho}]$ for same the parameter settings as in Fig. 7, but now with $\sigma_f = 2.5$Hz. Again, the curves for zero and random delays are practically coinciding.

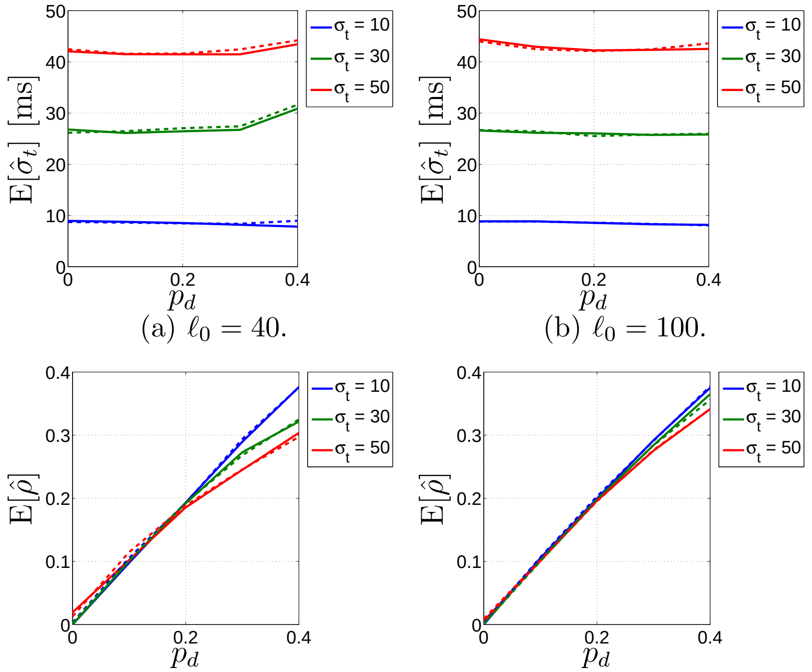
frequency, and hence are aligned incorrectly. The bias is about the same as for bivariate SES (see (Dauwels *et al.*, 2009b)).

- The estimates of s_t do only weakly depend on p_d, and vice versa.
- The estimates of s_t and p_d are robust to lags δ_t and frequency offsets δ_f, since the latter can be estimated reliably.
- The estimates of s_t and p_d are less biased for larger ℓ_0.

We have also observed from our experiments (not shown here):

- The estimates of δ_t and δ_f are unbiased for all considered values of δ_t, δ_f, s_t, s_f, and p_d.
- 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 normalized standard deviation of the SES parameters decreases as the length ℓ_0 increases, as expected.

In summary, by means of the N-variate SES inference method, one may reliably and robustly determine the timing dispersion s_t and event reliability ρ of a set of N (one-dimensional or multi-dimensional) point processes. As we also observed in (Dauwels *et al.*, 2009a,b) for bivariate SES, the timing dispersion and the number of event deletions are slightly underestimated due



(a) $\ell_0 = 40$. (b) $\ell_0 = 100$.

(c) $\ell_0 = 40$. (d) $\ell_0 = 100$.

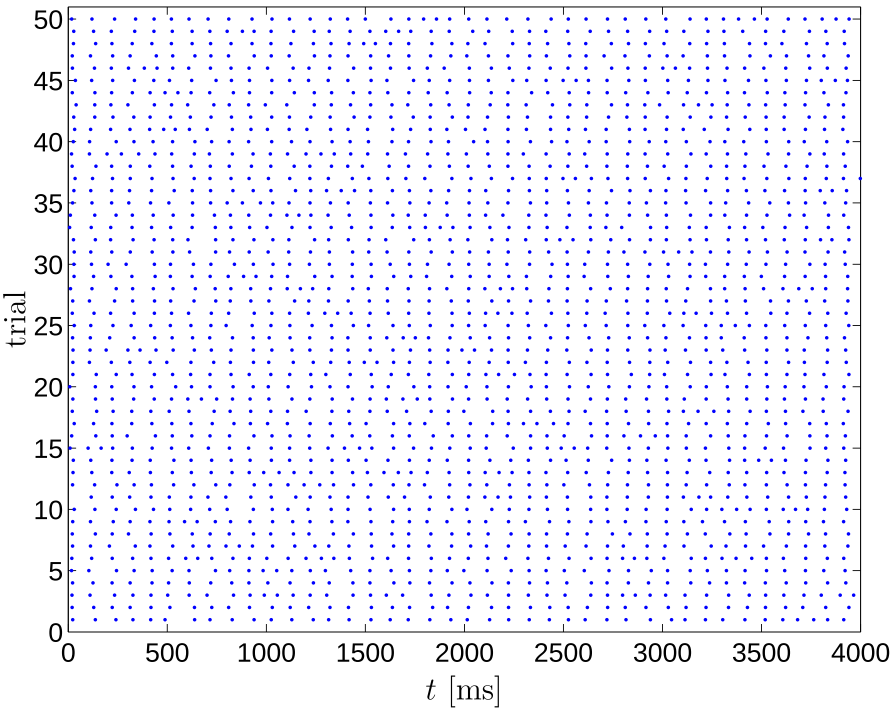
Fig. 9. Results for synthetic data: the figure shows the expected value $\mathrm{E}[\hat{\sigma}_t]$ and $\mathrm{E}[\hat{\rho}]$ and the normalized standard deviation $\bar{\sigma}[\hat{\sigma}_t]$ and $\bar{\sigma}[\hat{\rho}]$ for same the parameter settings as in Fig. 7, but now with $\sigma_f = 5$Hz. Again, the curves for zero and random delays are practically coinciding.

to the ambiguity inherent in event synchrony. However, this bias is smaller for N-variate SES than for bivariate SES, especially for one-dimensional point processes.

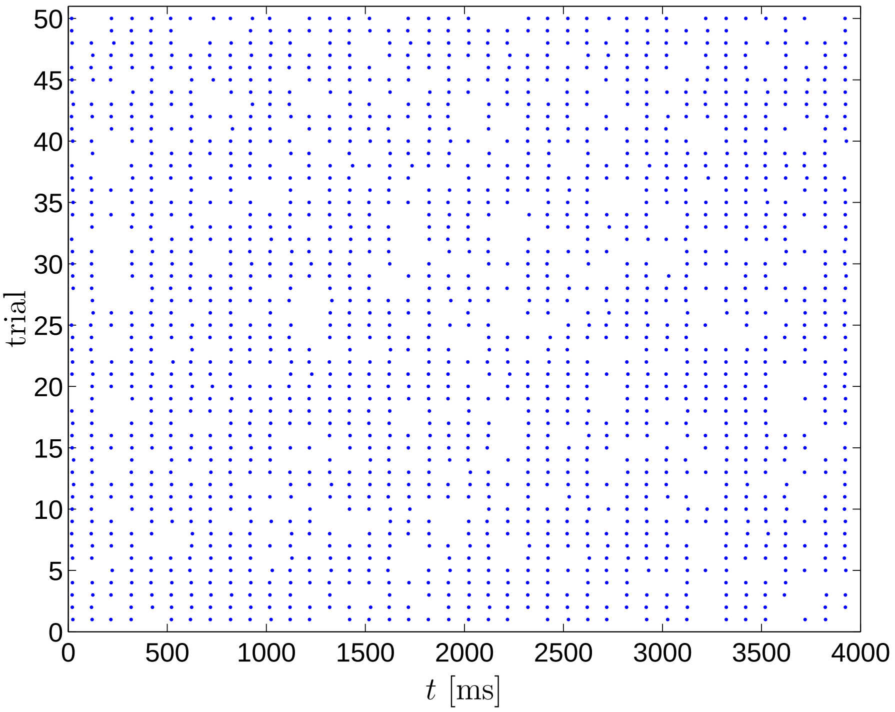
7 Application: Firing Reliablity of a Neuron

We consider here an application of SES that we also investigated in (Dauwels *et al.*, 2009a): We use SES to quantify the firing reliability of neurons. We again consider the Morris-Lecar neuron model (Morris *et al.*, 1981), which exhibits properties of type I and II neurons (Gutkin *et al.*, 1998; Tsumoto *et al.*, 2007; Tateno *et al.*, 2004). The spiking behavior differs in both neuron types, as illustrated in Fig. 10. In type II neurons, the timing jitter is small, but spikes tend to drop out. In type I neurons, on the other hand, fewer spikes drop out, but the dispersion of spike times is larger. In other words, type II neurons prefer to stay coherent or to be silent, on the other hand, type I neurons follow the middle course between those two extremes (Robinson, 2003).

In (Dauwels *et al.*, 2009a) we applied bivariate SES to the data of Fig. 10. Here we apply N-variate SES to the same data set. (We refer to (Dauwels *et al.*,

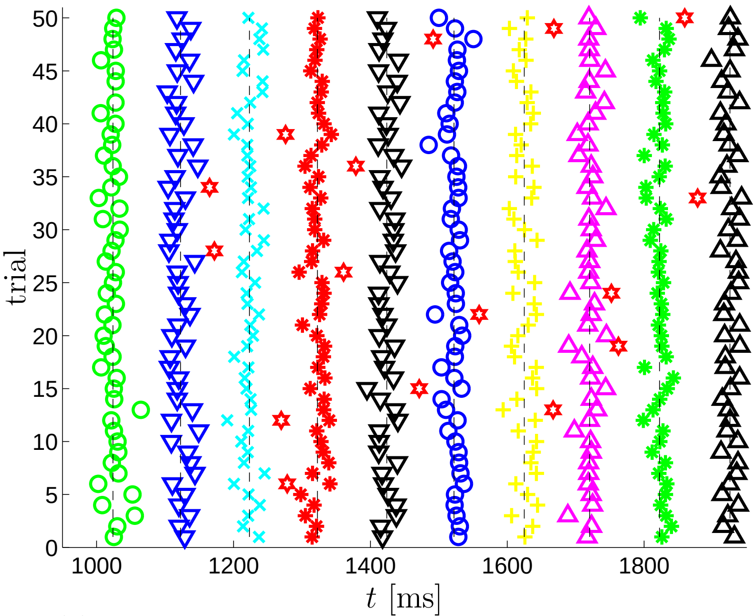


(a) Spike trains from type I neuron.

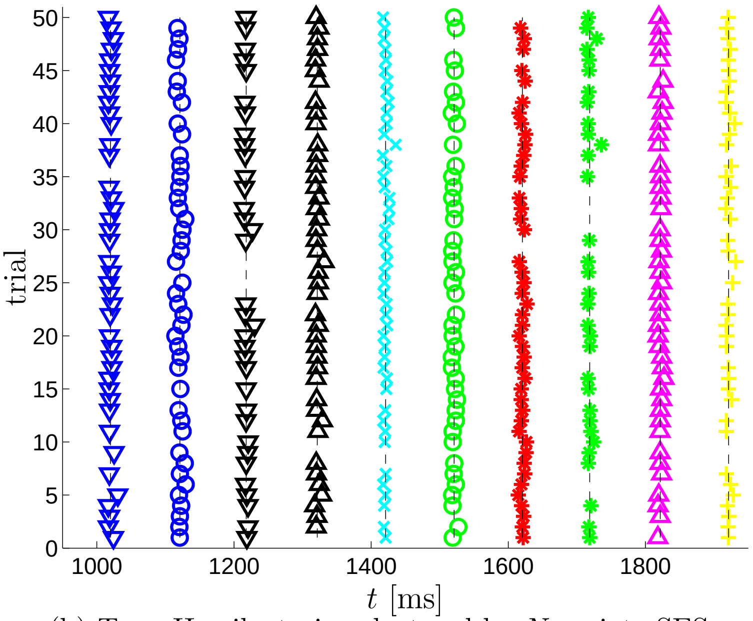


(b) Spike trains from type II neuron.

Fig. 10. Raster plots of spike trains from type I (top) and type II (bottom) neurons; in each case 50 spike trains are shown.



(a) Type I spike trains clustered by N-variate SES.



(b) Type II spike trains clustered by N-variate SES.

Fig. 11. Results of N-variate SES with $\beta = 10^{-3}$; for the sake of clarity, we show the range $t \in [950, 1950]$. Raster plots of spike trains from type I (top) and type II (bottom) neurons. Each cluster is indicated by a different combination of color and marker type (e.g., star, circle); background events are marked by red hexagons.

2009a) for more details on that data set.) In particular, we apply N-variate SES to the 50 trials simultaneously.

As an illustration, we show results of N-variate SES in Fig. 11. Each cluster is indicated by a different combination of color and marker type (e.g., star, circle); background events are marked by red hexagons.

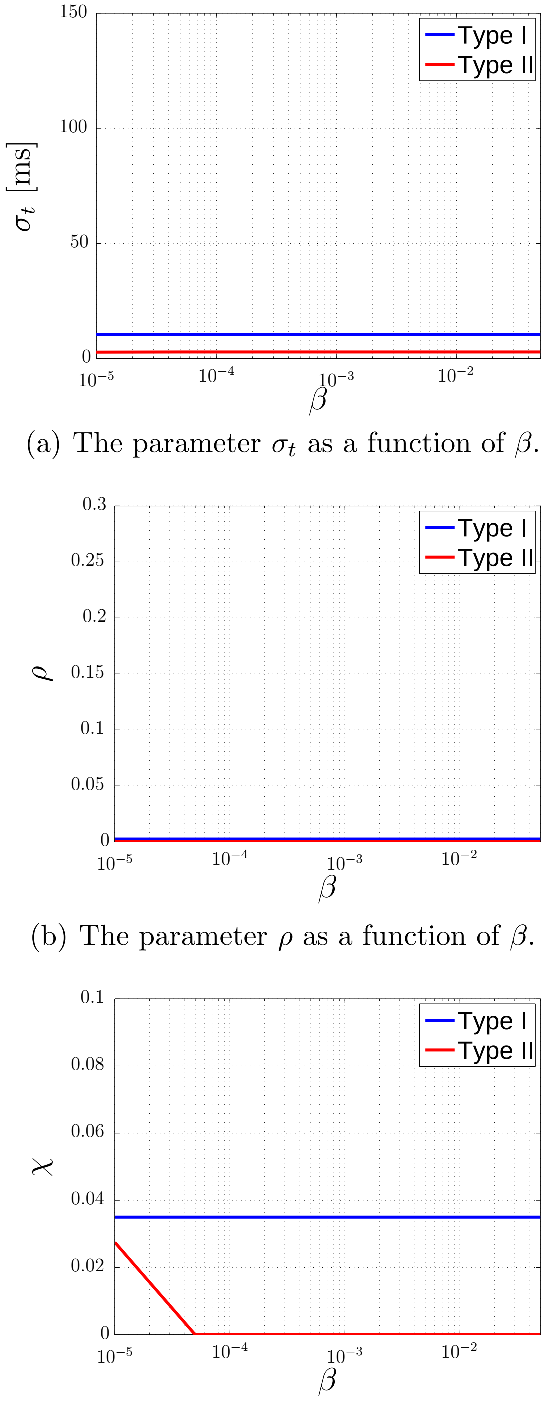
We choose the parameters in the N-variate SES algorithm as follows. We set $\hat{\delta}_t^{(0)} = 0$, and $\hat{s}_t^{(0)} = (3\text{ms})^2$, $(5\text{ms})^2$, $(7\text{ms})^2$ and $(9\text{ms})^2$. Each initialization of $(\hat{\delta}_t^{(0)}, \hat{s}_t^{(0)})$ may lead to a different solution $(\hat{c}, \hat{\delta}_t, \hat{s}_t)$; we choose the most probable solution, i.e., the one that has the largest value $p(x, \hat{c}, \hat{\delta}_t, \hat{s}_t)$ (23). We set $\tilde{\beta} = 10^{-10}$. Larger values of $\tilde{\beta}$ lead to a prohibitively large number of background events, whereas smaller values yield no background events at all.

We computed the SES parameters for different values of β. Fig. 12 shows how s_t (σ_t), ρ, and χ (fraction of background events) depend on β for both neuron types. From those figures it becomes immediately clear that the parameters s_t (σ_t) and ρ hardly depend on β. For values of $\beta < 10^{-4}$, some of the events from the Type II neuron are considered as background events, which is obviously incorrect (cf. Fig. 11(b)). Therefore, only values $\beta > 10^{-4}$ should be considered.

The parameter ρ is significantly smaller in type I than in type II neurons, in contrast, s_t is vastly larger. This agrees with our intuition: since in type II neurons spikes tend to drop out, ρ should be larger. On the other hand, since the timing dispersion of the spikes in type I is larger, we expect s_t to be larger in those neurons. We have made the same observations in (Dauwels *et al.*, 2009a).

Table 3 compares the numerical results for bivariate and N-variate SES. As we pointed out earlier, bivariate SES defines the variance of the perturbations in the generative process as $s_t/2$ (instead of s_t), so that the variance between the two *observed* sequences x_1 and x_2 is given by s_t. In Table 3 the standard deviations ($\sigma_t = \sqrt{s_t}$) in the *generative process* are reported, both for bivariate and N-variate SES; the values in (Dauwels *et al.*, 2009a) differ by a factor $\sqrt{2}$, since there the standard deviation between the observed sequences is reported.

In (Dauwels *et al.*, 2009a) we assessed the reliability of the bivariate SES estimates by means of bootstrapping (Efron *et al.*, 1993); we follow a similar procedure here for the N-variate SES estimates. In particular, for both types of neurons we generated 1,000 sets of 50 spike trains; we followed the generative process of Fig. 4, with the N-variate SES parameters of the actual spike trains, i.e., $(\rho, \sigma_t, \chi) = (10.6, 0.0025, 0.035)$ and $(\rho, \sigma_t, \chi) = (2.8, 0.18, 0.0)$ for type I and II neurons respectively. Next we applied N-variate SES to the resulting 1,000 sets of 50 spike trains. The expected value and normalized standard



(a) The parameter σ_t as a function of β.

(b) The parameter ρ as a function of β.

(c) The parameter χ as a function of β.

Fig. 12. The parameters σ_t, ρ, and χ estimated from spike trains of type I and type II Morris-Lecar neurons (cf. Fig. 10): the top, middle, and bottom figure show how σ_t, ρ, and χ respectively depend on β.

deviation $\bar{\sigma}$ of those estimates is reported in Table 3. We can observe that the expected value corresponds well with the actual value, and the normalized standard deviations are small. Therefore, the N-variate SES estimates can be considered reliable.

We first discuss the results for the type I neuron: the estimate of σ_t from bivariate and N-variate SES is almost identical; however, the estimate of ρ is much smaller for N-variate SES than for bivariate SES. Interestingly, N-variate SES inferred that about 3.5% of the events are background events, which accounts for the larger estimate $\rho = 0.029$ from the bivariate approach. In other words, type I neurons almost never fail to fire (firing reliability of 99.75%), however, additional spikes may occur (3.5% of the spikes). It is noteworthy that this insight was obtained by N-variate SES, and could not be revealed through bivariate SES; this example thus illustrates that N-variate SES not only yields more accurate estimates of the SES parameters (cf. Section 6), but also can lead to a more refined and detailed analysis. Interestingly, the normalized standard deviation $\bar{\sigma}[\rho]$ is much larger for N-variate SES than for bivariate SES, since ρ is much smaller for N-variate SES. However, the standard deviation $\sigma[\rho]$ is small and about the same for both models.

We now elaborate on the results for the type II neuron. The N-variate approach leads to larger and smaller estimates of σ_t and ρ respectively than the bivariate approach. None of the events are considered as background events ($\chi = 0$). We have manually counted the number of deletions in Fig. 11(b), and obtained $\rho = 0.184$. The N-variate estimate of ρ is exact and clearly the most reliable, whereas the bivariate approach overestimates the number of deletions. Since the N-variate approach considers all point processes simultaneously ($N = 50$), it infers the hidden process v more reliably than bivariate SES, and is able to associate events x more accurately with hidden events v_k.

We have observed that the N-variate SES algorithm converges after at most three iterations, both for type I and type II neurons. In each of those iterations, one updates the decision variables b, c, and e, and the SES parameters θ. Since we allowed a maximum number of 30 iterations, we can conclude that the algorithm always has converged in our experiments.

8 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 (Jeong, 2004; Dauwels *et al.*, 2010b) for a review). In this section, we apply SES to detect subtle perturbations in EEG synchrony of MCI patients. We considered

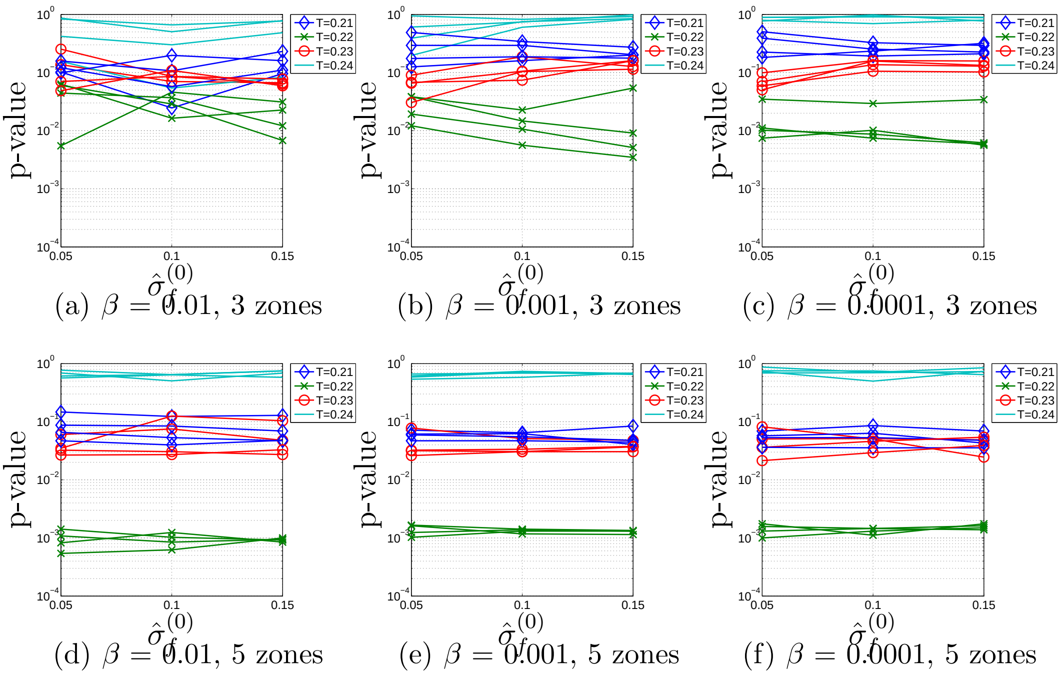
	Bivariate SES		N-variate SES	
Statistics	Type I	Type II	Type I	Type II
σ_t	10.7	1.91	10.6	2.81
$\mathrm{E}[\sigma_t]$	10.8	1.91	10.3	2.87
$\bar{\sigma}[\sigma_t]$	1.8%	1.8%	3.7%	3.0%
ρ	0.0290	0.270	0.0025	0.184
$\mathrm{E}[\rho]$	0.0283	0.273	0.0036	0.186
$\bar{\sigma}[\rho]$	12%	3.1%	90%	15%
χ	–	–	0.035	0.0
$\mathrm{E}[\chi]$	–	–	0.037	0.0
$\bar{\sigma}[\chi]$	–	–	11%	0.0%

Table 3
Estimates of bivariate SES estimates (ρ, σ_t) and N-variate SES parameters (ρ, σ_t, χ). Also shown are the results from the bootstrapping analysis of those estimates, in particular, the expected values and the normalized standard deviations $\bar{\sigma}$. The expected values practically coincide with the actual estimates and the normalized standard deviations are small; therefore, the estimates may be considered reliable.

this application also in (Dauwels *et al.*, 2009b), where we applied bivariate SES. We analyze here the same EEG data set, and use the same preprocessing and bump modeling procedures as in (Dauwels *et al.*, 2009b). The only difference is that we here apply N-variate SES instead of bivariate SES.

We first conducted a similar statistical analysis as in (Dauwels *et al.*, 2009b). The main results of that analysis are summarized in Fig. 13 and 14; they contain p-values obtained by the Mann-Whitney test for the parameters ρ and s_t respectively. This test indicates whether the parameters take different values for the two subject populations. More precisely, low p-values indicate large difference in the medians of the two populations. The p-values are shown for $\hat{\sigma}_t^{(0)} = \sqrt{s_{0,t}} = 0.1, 0.15, \ldots, 0.25$, $\hat{\sigma}_f^{(0)} = \sqrt{s_{0,f}} = 0.05, 0.1, \ldots, 0.15$, $\beta =$ 0.01, 0.001, 0.0001, $T =$ 0.21, 0.22, 0.23, 0.24, and the number of zones $N_R =$ 3 and 5.

The results for the parameters ρ and s_t are quite similar to the results obtained with bivariate SES (see (Dauwels *et al.*, 2009b)). The lowest p-values for ρ are



(a) $\beta = 0.01$, 3 zones (b) $\beta = 0.001$, 3 zones (c) $\beta = 0.0001$, 3 zones

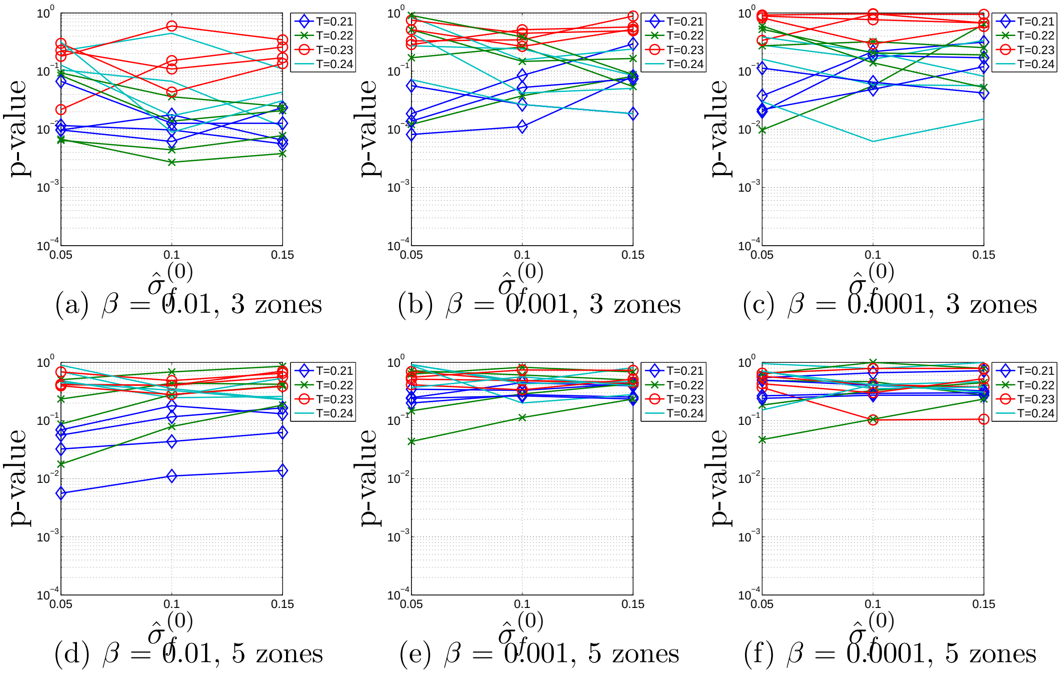
(d) $\beta = 0.01$, 5 zones (e) $\beta = 0.001$, 5 zones (f) $\beta = 0.0001$, 5 zones

Fig. 13. p-values obtained by the Mann-Whitney test for the parameter ρ for $\hat{\sigma}_t^{(0)} = \sqrt{s_{0,t}} = 0.1, 0.15, \ldots, 0.25$, $\hat{\sigma}_f^{(0)} = \sqrt{s_{0,f}} = 0.05, 0.01, 0.15$, $\beta = 0.01$, 0.001, 0.0001, $T = 0.21$, 0.22, 0.23, 0.24 and the number of zones $N_R = 3$ and 5. The p-values seem to vary little with $\sigma_t^{(0)}$, $\sigma_f^{(0)}$ and β, but are more dependent on T and the number of zones. The lowest p-values are obtained for $T = 0.22$ and $N_R = 5$ zones; the corresponding statistical differences are highly significant.

obtained for $T = 0.22$, $\beta = 0.01$, and $N_R = 5$ (see Fig. 13(e)), i.e., the smallest value is $\mathrm{p} = 5.4 \cdot 10^{-4}$; in bivariate SES, the smallest p-value ($\mathrm{p} = 2.1 \cdot 10^{-4}$) was obtained for for $T = 0.22$, $\beta = 0.001$, and $N_R = 5$ (see (Dauwels *et al.*, 2009b)). Similarly as in bivariate SES, the results depend strongly on T (cf. (Dauwels *et al.*, 2009b)); we provided an explanation for this dependency in (Dauwels *et al.*, 2009b). Interestingly, the results depend much less on $\hat{\sigma}_t^{(0)}$, $\hat{\sigma}_f^{(0)}$, and β than in bivariate SES.

From bivariate and N-variate SES analysis (see Fig. 13), we can conclude that the statistical differences in ρ are highly significant, especially for $T = 0.22$ and $N_R = 5$: There is a significantly higher degree of non-correlated activity in MCI patients, more specifically, a high number of non-coincident, non-synchronous oscillatory events. As in (Dauwels *et al.*, 2009b), we did not observe a strongly significant effect on the timing jitter s_t of the coincident events (see Fig. 14): very few p-values for s_t are smaller than 0.001, which suggests there are no strongly significant differences in s_t.

The N-variate SES model allows us to analyze the results for ρ in more detail; we have investigated how the statistics of bump clusters differ in controls subjects and MCI patients. More specifically, we considered the relative fre-



(a) $\beta = 0.01$, 3 zones

(b) $\beta = 0.001$, 3 zones

(c) $\beta = 0.0001$, 3 zones

(d) $\beta = 0.01$, 5 zones

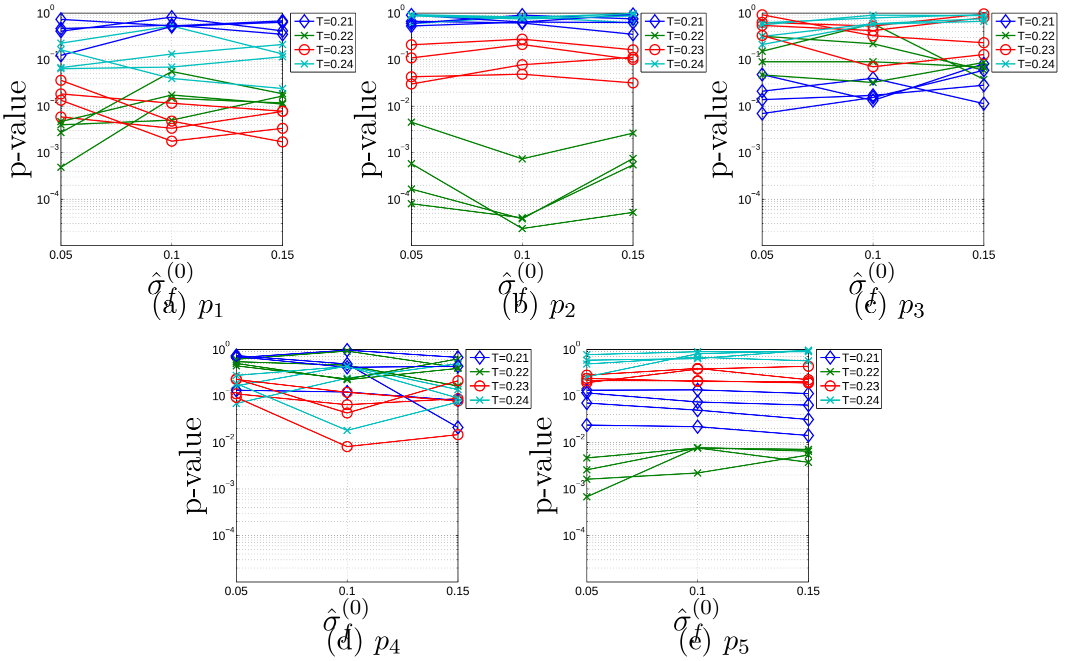
(e) $\beta = 0.001$, 5 zones

(f) $\beta = 0.0001$, 5 zones

Fig. 14. p-values obtained by the Mann-Whitney test for the parameter s_t for $\hat{\sigma}_t^{(0)} = \sqrt{s_{0,t}} = 0.1, 0.15, \ldots, 0.25$, $\hat{\sigma}_f^{(0)} = \sqrt{s_{0,f}} = 0.05, 0.01, 0.15$, $\beta = 0.01$, 0.001, 0.0001, $T = 0.21$, 0.22, 0.23, 0.24 and the number of zones $N_R = 3$ and 5. Very few p-values are smaller than 0.001, which suggests there are no strongly significant differences in s_t.

quency $p_j = p(n_k = j)$ of bump clusters of size $n_k = j$, for $j = 1, 2, \ldots, N_R$. The results are summarized in Fig. 15, for the parameter settings that yielded the smallest p-values for ρ ($N_R = 5$ and $\beta = 0.01$). From those figures, we can observe strongly significant differences in clusters of size 1, 2 and 5 for $T = 0.22$; specifically, in MCI patients there are fewer clusters of size 5 and more clusters of size 1 and 2. As a result, the fraction of missing events ρ is larger in MCI patients, as we mentioned earlier. The smallest p-values for the parameter p_2 is $\text{p} = 2 \cdot 10^{-5}$, which is substantially smaller than the smallest p-value for ρ ($\text{p} = 2.1 \cdot 10^{-4}$ for bivariate SES and $\text{p} = 5.4 \cdot 10^{-4}$ for N-variate SES).

In (Dauwels *et al.*, 2010a) we have applied a variety of classical synchrony measures to the same EEG data set. The main results of that analysis are summarized in Table 4. The most significant results were obtained with the the full-frequency direct transfer function (ffDTF), which is a Granger measure (Pereda *et al.*, 2005), resulting in a p-value of about 10^{-3} (Mann-Whitney test). In (Dauwels *et al.*, 2009b) we have combined ρ with ffDTF as features to distinguish MCI from control subjects (see Fig. 16). About 84% of the subjects are correctly classified. Here we combine ffDTF with the parameter p_2, computed by N-variate SES; the classification rate slightly improves to about 87%. This result is encouraging, however, it is too weak to allow us to



(a) p_1 (b) p_2 (c) p_3

(d) p_4 (e) p_5

Fig. 15. p-values obtained by the Mann-Whitney test for the parameters $p_i = p(n_k = i)$ (with $i = 1, 2, \ldots, 5$) for $\hat{\sigma}_t^{(0)} = \sqrt{s_{0,t}} = 0.05, 0.1, 0.15$, $\hat{\sigma}_f^{(0)} = \sqrt{s_{0,f}} = 0.01, 0.15, \ldots, 0.25$, $\beta = 0.01$, $T = 0.22$, and the number of zones $N_R = 5$.

Measure	Correlation	Coherence	Phase Coherence	Corr-entropy	Wave-entropy	
p-value	**0.025***	**0.029***	**0.041***	**0.032***	0.096	
Measure	MVAR coherence	Partial Coherence	PDC	DTF	ffDTF	dDTF
p-value	0.15	0.16	0.60	0.29	**0.0012****†	**0.029***
Measure	Kullback-Leibler	Rényi	Jensen-Shannon	Jensen-Rényi	I_W	I
p-value	0.065	0.067	0.069	0.074	0.052	0.060
Measure	N^k	S^k	H^k	S-estimator	Omega complexity	
p-value	**0.029***	**0.045***	0.052	**0.042***	0.079	
Measure	Hilbert Phase	Wavelet Phase	Evolution Map	Instantaneous Period	GFS	
p-value	0.96	0.082	0.64	0.73	**0.031***	
Measure	s_t (bivariate)	ρ (bivariate)	s_t (multivariate)	ρ (multivariate)	p_2 (multivariate)	
p-value	0.19	**0.00021****	**0.018***	**0.00054****	$\mathbf{2 \cdot 10^{-5}}^{**}$	

Table 4
Sensitivity for prediction of MCI: uncorrected p-values for Mann-Whitney test; * and ** indicate $p < 0.05$ and $p < 0.005$ respectively; † indicates p-values that remain significant after post-correction. The results for the standard measures and bivariate SES (s_t and ρ) are from (Dauwels *et al.*, 2010a) and (Dauwels *et al.*, 2009b) respectively, and we refer to those references for more details.

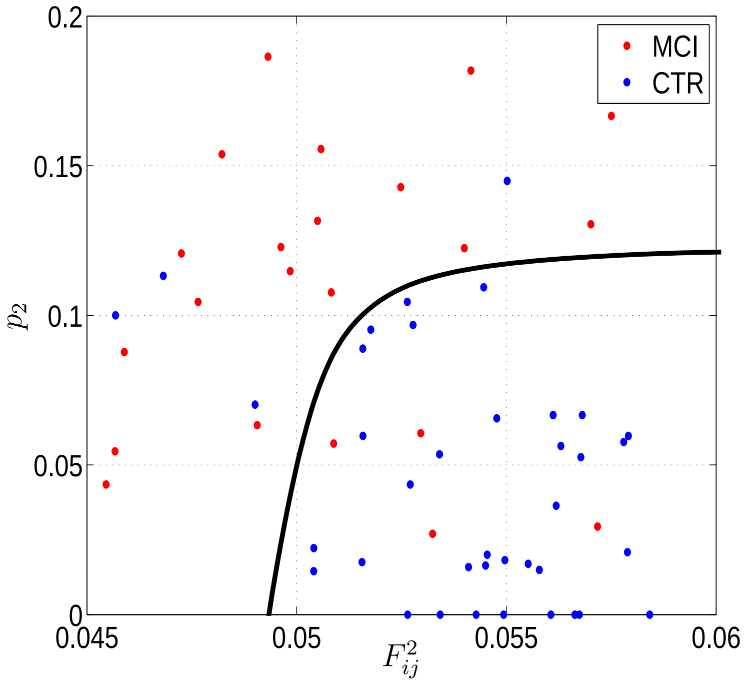


Fig. 16. Combining p_2 with ffDTF as features to distinguish MCI from age-matched control subjects. Note that ffDTF is a similarity measure whereas p_2 is a dissimilarity measure. The (ffDTF, p_2) pairs of the MCI and control subjects tend towards the left top corner and bottom right corner respectively. The smooth curve (solid) yields a classification rate of 87%.

predict AD reliably. We would need to combine synchrony measures with complementary statistics, for example, spectral features. We refer to (Dauwels *et al.*, 2010b) for more information on potential extensions. Moreover, the results would of course need to be verified on more datasets.

In summary, N-variate SES helped us to better understand the results from the bivariate SES analysis (Dauwels *et al.*, 2009b): the fraction of missing events ρ is larger in MCI patients, since in those patients there are fewer bump clusters of size 5 and more clusters of size 1 and 2. Moreover, N-variate SES allowed us to further improve the classification of MCI patients vs. control subjects from 84% to 87%.

9 Conclusion

We proposed an approach to determine the similarity of $N > 2$ (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 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.

In bivariate SES, we apply the max-product algorithm for aligning pairs of sequences. As we have observed, the max-product algorithm performs poorly for aligning $N > 2$ sequences; we have also experimented with various refinements of the max-product algorithm, and none of them yielded satisfactory results. An interesting topic for future research is to develop extensions of the max-product algorithm that lead to optimal or close-to-optimal N-wise alignments. Such message passing algorithms are often simpler and faster than integer linear programming techniques.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful feedback that has substantially improved the paper.

A Appendix: Derivation of the SES Model

In this appendix, we derive the N-variate SES model (17).

We first marginalize $p(x, c, v, \theta, \ell)$ (16) over v; it is noteworthy that only the Gaussian terms $\mathcal{N}(\cdot)$ in (16) depend on v. In the following, we will focus on those terms. For a given hidden event v_k, three cases are possible:

- *All* copies of v_k were deleted ($n_k = 0$). There are no Gaussian terms associated with v_k in (16), therefore, the expression (16) may be considered as constant w.r.t. v_k. Integrating (16) over v_k then leads to a term $\mathrm{vol}(S)$. There are L^{del} such terms, since the number of hidden events v_k without copies is given by L^{del}.
- There is *one* copy of v_k ($n_k = 1$), and therefore, there is only *one* Gaussian term in (16) that corresponds to v_k. Integrating that term over v_k results in the (trivial) term 1.
- There are *more than one* copies of v_k ($n_k > 1$), and as a consequence, there are several Gaussian terms in (16) that correspond to v_k. As easily can be shown, integrating the n_k Gaussian terms over v_k yields the terms:

$$\prod_{(i,j)\in\mathcal{I}_k} \mathcal{N}\Big(t_{ij} - \bar{t}_k; \delta_{ti}, s_{ti}\Big)\, \mathcal{N}\Big(f_{ij} - \bar{f}_k; \delta_{fi}, s_{fi}\Big), \tag{A.1}$$

where

$$\bar{t}_k = \frac{\sum\limits_{(i,j)\in\mathcal{I}_k} w_{ti}(t_{ij} - \delta_{ti})}{\sum\limits_{(i,j)\in\mathcal{I}_k} w_{ti}} \tag{A.2}$$

$$\bar{f}_k = \frac{\sum\limits_{(i,j)\in\mathcal{I}_k} w_{fi}(f_{ij} - \delta_{fi})}{\sum\limits_{(i,j)\in\mathcal{I}_k} w_{fi}}, \tag{A.3}$$

with $w_{ti} = s_{ti}^{-1}$ and $w_{fi} = s_{fi}^{-1}$.

In summary, marginalizing over v results in L^{del} terms $\text{vol}(S)$, and also in terms of the form (A.1), where there is one such term for each cluster of size $n_k > 1$.

The result of marginalizing over v may then be written as:

$$\begin{aligned} p(x, c, \theta, \ell) \propto\, & p(s_t)p(s_f)(1 - \lambda\,\text{vol}(S))(\lambda\text{vol}(S)p_d^N)^{L^{\text{del}}} \\ & \cdot (\lambda p_d^N)^L p_d^{-L^{\text{tot}}}(1 - p_d)^{L^{\text{tot}}} \\ & \cdot \prod_{k\in\mathcal{K}} \prod_{(i,j)\in\mathcal{I}_k} \mathcal{N}\Big(t_{ij} - \bar{t}_k; \delta_{ti}, s_{ti}\Big)\,\mathcal{N}\Big(f_{ij} - \bar{f}_k; \delta_{fi}, s_{fi}\Big), \end{aligned} \tag{A.4}$$

where we used the decomposition:

$$\ell = L + L^{\text{del}}. \tag{A.5}$$

Now we marginalize over the length ℓ. The first term in the decomposition (A.5) is fixed for given clustering c. Therefore, marginalizing $p(x, c, \theta, \ell)$ (A.4) over ℓ is equivalent to marginalizing over L^{del}:

$$p(x, c, \theta) = \sum_{\ell=0}^{\infty} p(x, c, \theta, \ell) = \sum_{L^{\text{del}}=0}^{\infty} p(x, c, \theta, \ell) \tag{A.6}$$

$$\begin{aligned} \propto\, & p(s_t)p(s_f)(1 - \lambda\,\text{vol}(S)) \sum_{L^{\text{del}}=0}^{\infty} (\lambda\text{vol}(S)p_d^N)^{L^{\text{del}}} \\ & \cdot (\lambda p_d^N)^L p_d^{-L^{\text{tot}}}(1 - p_d)^{L^{\text{tot}}} \\ & \cdot \prod_{k\in\mathcal{K}} \prod_{(i,j)\in\mathcal{I}_k} \mathcal{N}\Big(t_{ij} - \bar{t}_k; \delta_{ti}, s_{ti}\Big)\,\mathcal{N}\Big(f_{ij} - \bar{f}_k; \delta_{fi}, s_{fi}\Big) \end{aligned} \tag{A.7}$$

$$\begin{aligned} \propto\, & p(s_t)p(s_f)(1 - \lambda\,\text{vol}(S))\frac{1}{1 - \lambda\text{vol}(S)p_d^N} \\ & \cdot (\lambda p_d^N)^L p_d^{-L^{\text{tot}}}(1 - p_d)^{L^{\text{tot}}} \\ & \cdot \prod_{k\in\mathcal{K}} \prod_{(i,j)\in\mathcal{I}_k} \mathcal{N}\Big(t_{ij} - \bar{t}_k; \delta_{ti}, s_{ti}\Big)\,\mathcal{N}\Big(f_{ij} - \bar{f}_k; \delta_{fi}, s_{fi}\Big). \end{aligned} \tag{A.8}$$

In (A.7) occurs a sum of a geometric series; since $|\lambda\text{vol}(S)p_d^N| = \lambda\text{vol}(S)p_d^N < 1$, we can apply the well-known formula for the sum of a geometric series, resulting in (A.8). By defining β and γ as in (20) and (21) respectively, we obtain statistical model (17).

B Appendix: Derivation of the SES Inference Algorithm

In this appendix, we derive the inference update (34) for N-variate SES.

The point estimates $\hat{\delta}_{ti}^{(\kappa)}$ and $\hat{\delta}_{fi}^{(\kappa)}$ are the (sample) mean of the timing and frequency offset respectively, computed between all events in x_{ij} and their associated cluster centers:

$$\hat{\delta}_{ti}^{(\kappa)} = \frac{1}{L_i}\sum_{j=1}^{L_i}\Big(t_{ij} - \bar{t}_{\hat{c}_{ij}^{(\kappa)}}\Big) \tag{B.1}$$

$$\hat{\delta}_{fi}^{(\kappa)} = \frac{1}{L_i}\sum_{j=1}^{L_i}\Big(f_{ij} - \bar{f}_{\hat{c}_{ij}^{(\kappa)}}\Big), \tag{B.2}$$

where L_i is the number of events in x_i, and $\bar{t}_{\hat{c}_{ij}^{(\kappa)}}$ and $\bar{f}_{\hat{c}_{ij}^{(\kappa)}}$ are the coordinates of the (inferred) hidden event $v_{\hat{c}_{ij}^{(\kappa)}}$ associated with x_{ij}. The latter hidden event is inferred as the center of the cluster associated with x_{ij}. The coordinates $\bar{t}_k$ and $\bar{f}_k$ are computed as:

$$\bar{t}_k = \frac{\sum_{(i,j)\in\hat{\mathcal{I}}_k^{(\kappa)}} \hat{w}_{ti}^{(\kappa)}(t_{ij} - \hat{\delta}_{ti}^{(\kappa)})}{\sum_{(i,j)\in\hat{\mathcal{I}}_k^{(\kappa)}} \hat{w}_{ti}^{(\kappa)}} \tag{B.3}$$

$$\bar{f}_k = \frac{\sum_{(i,j)\in\hat{\mathcal{I}}_k^{(\kappa)}} \hat{w}_{fi}^{(\kappa)}(f_{ij} - \hat{\delta}_{fi}^{(\kappa)})}{\sum_{(i,j)\in\hat{\mathcal{I}}_k^{(\kappa)}} \hat{w}_{fi}^{(\kappa)}}, \tag{B.4}$$

with $\hat{w}_{ti}^{(\kappa)} = \Big(\hat{s}_{ti}^{(\kappa)}\Big)^{-1}$, $\hat{w}_{fi}^{(\kappa)} = \Big(\hat{s}_{fi}^{(\kappa)}\Big)^{-1}$, and $\hat{\mathcal{I}}_k^{(\kappa)}$ is the index set of the set $\hat{\mathcal{C}}_k^{(\kappa)}$ of events in cluster k (as specified by $\hat{c}^{(\kappa)}$):

$$\mathcal{C}_k^{(\kappa)} = \Big\{x_{ij} : \hat{c}_{ij}^{(\kappa)} = k\Big\} \text{ and } \mathcal{I}_k^{(\kappa)} = \Big\{(i,j) : \hat{c}_{ij}^{(\kappa)} = k\Big\}. \tag{B.5}$$

The estimates $\hat{s}_{ti}^{(\kappa)}$ and $\hat{s}_{fi}^{(\kappa)}$ are obtained as:

$$\hat{s}_{ti}^{(\kappa)} = \frac{\nu_t s_{t0} + L_i \, \hat{s}_{ti,\text{sample}}^{(\kappa)}}{\nu_t + L_i + 2} \tag{B.6}$$

$$\hat{s}_{fi}^{(\kappa)} = \frac{\nu_f s_{f0} + L_i \, \hat{s}_{fi,\text{sample}}^{(\kappa)}}{\nu_f + L_i + 2}, \tag{B.7}$$

where $s_{ti,\text{sample}}^{(\kappa)}$ and $s_{fi,\text{sample}}^{(\kappa)}$ are computed as:

$$\hat{s}_{ti,\text{sample}}^{(\kappa)} = \frac{1}{L_i} \sum_{j=1}^{L_i} \left(t_{ij} - \bar{t}_{\hat{c}_{ij}^{(\kappa)}} - \hat{\delta}_{ti}^{(\kappa)} \right)^2 \tag{B.8}$$

$$\hat{s}_{fi,\text{sample}}^{(\kappa)} = \frac{1}{L_i} \sum_{j=1}^{L_i} \left(f_{ij} - \bar{f}_{\hat{c}_{ij}^{(\kappa)}} - \hat{\delta}_{fi}^{(\kappa)} \right)^2, \tag{B.9}$$

and where $\bar{t}_k$ and $\bar{f}_k$ are given by (B.3) and (B.4).

Interestingly, the RHS of (B.1) and (B.2) depends on $\hat{\delta}_{ti}^{(\kappa)}$ and $\hat{\delta}_{fi}^{(\kappa)}$ respectively, through (B.3) and (B.4). Therefore the equalities (B.1) and (B.2) need to be solved numerically; the same holds for (B.6) and (B.7). Those expressions may be evaluated numerically by alternating the updates (B.3) and (B.4) with (B.1) (B.2) (B.6) and (B.7), with as initial estimates $\hat{\delta}_{ti}^{(\kappa-1)}$, $\hat{\delta}_{fi}^{(\kappa-1)}$, $\hat{s}_{ti}^{(\kappa-1)}$, and $\hat{s}_{fi}^{(\kappa-1)}$. It can easily be shown that this procedure is guaranteed to converge to a local extremum; indeed, it is equivalent to cyclic maximization, where one conditional maximization is in θ (resulting in (B.1) (B.2) (B.6) and (B.7)) and the other is in the parameters $\{(\bar{t}_k, \bar{f}_k)\}_{k=1,\ldots,L}$ (resulting in (B.3) and (B.4)).

References

Abeles M., Bergman H., Margalit E., and Vaadia E., 1993. Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. J. Neurophysiol 70(4), 1629–1638.

Amari S., Nakahara H., Wu S., and Sakai Y., 2003. Synchronous firing and higher-order interactions in neuron pool. Neural Computation 15, 127–142.

Buzsáki G., 2006. Rhythms of the brain. Oxford University Press.

Chapman R., Nowlis G., McCrary J., Chapman J., Sandoval T., Guillily M., Gardner M., Reilly L., 2007. Brain event-related potentials: diagnosing early-stage Alzheimer's disease. Neurobiol. Aging 28, 194–201.

Cichocki A., Shishkin S., Musha T., Leonowicz Z., Asada T., and Kurachi T., 2005. EEG filtering based on blind source separation (BSS) for early diagnosis of Alzheimer's disease. Clin. Neurophys 116, 729–37.

Dauwels J., Vialatte F., Rutkowski T., and Cichocki A., 2008. Measuring Neural Synchrony by Message Passing. In *Advances in Neural Information Processing Systems* 20, Edited by Platt J., Koller D., Singer Y., Roweis S., Cambridge, MA: MIT Press 361–368.

Dauwels J., Tsukada Y., Sakumura Y., Ishii S., Aoki K., Nakamura T., Matsuda M., Vialatte F., and Cichocki A., 2008. On the synchrony of morphological and molecular signaling events in cell migration, International Conference on Neural Information Processing (ICONIP), Lecture Notes in Computer Science (LNCS) 5506: 469–477.

Dauwels J., Vialatte F., Weber T., and Cichocki A., 2009a. Quantifying Statistical Interdependence by Message Passing on Graphs: I. One-Dimensional Point Processes, Neural Computation 21(8): 2152–2202.

Dauwels J., Vialatte F., Weber T., and Cichocki A., 2009b. Quantifying Statistical Interdependence by Message Passing on Graphs: II. Multi-Dimensional Point Processes, Neural Computation 21(8): 2203–2268.

Dauwels J., Vialatte F., and Cichocki A., 2010a. A comparative study of synchrony measures for the early diagnosis of Alzheimer's disease based on EEG, Neuroimage 49:668–693.

Dauwels J., Vialatte F., and Cichocki A., 2010b. Diagnosis of Alzheimer's Disease from EEG Signals: Where Are We Standing?, Curr Alzheimer Research, Epub May 11 2010.

Delprat N., Escudié B., Guillemain P., Kronland-Martinet R., Tchamitchian P., and Torrésani B., 1992. Asymptotic wavelet and Gabor analysis: extraction of instantaneous frequencies. IEEE Trans. Information Theory 38, 644–664.

Gutkin B. S. and Ermentrout G. B., 1998. Dynamics of membrane excitability determine interspike interval variability: a link between spike generation mechanisms and cortical spike train statistics. Neural Computation 10, 1047–1065.

Efron, B. and Tibshirani, R., 1993. *An Introduction to the Bootstrap,* Chapman & Hall/CRC.

Frey B. and Dueck D., 2007. Clustering by passing messages between data points. Science 315:5814, 972–976.

Givoni I. and Frey B. J., 2009. A binary variable model for affinity propagation. Neural Computation 21:6, 1589–1600.

Goupillaud P., Grossman A., and Morlet J., 1984. Cycle-octave and related transforms in seismic signal analysis. Geoexploration 23, 85–102.

Harris F.J., 2004. *Multirate signal processing for communication systems.* Upper Saddle River, NJ: Prentice Hall PTR.

Herrmann C. S., Grigutsch M., and Busch N. A., 2005. EEG oscillations and wavelet analysis. In: Handy, T. (ed.), Event-Related Potentials: a Methods Handbook, Cambridge, MIT Press, 229–259.

Hogan M. J., Swanwick G. R. J., Kaiser J., Rowan M., and Lawlor B., 2003. Memory-related EEG power and coherence reductions in mild Alzheimer's disease. Int. J. Psychophysiol. 49.

Jeong J., 2004. EEG dynamics in patients with Alzheimer's disease. Clinical Neurophysiology 115, 1490–1505.

Lashkari D. and Golland P., 2008. Convex clustering with exemplar-based models. In *Advances in Neural Information Processing Systems* 20, Edited by Platt J., Koller D., Singer Y., Roweis S., Cambridge, MA: MIT Press 361–368.

Matsuda H., 2001. Cerebral blood flow and metabolic abnormalities in Alzheimer's disease. Ann. Nucl. Med. 15, 85–92.

Morris C. and Lecar H., 1981. Voltage oscillations in the barnacle giant muscle fiber. Biophys. J. 35, 193–213.

Musha T., Asada T., Yamashita F., Kinoshita T., Chen Z., Matsuda H., Uno M., Shankle W.R., 2002. A new EEG method for estimating cortical neuronal impairment that is sensitive to early stage Alzheimers disease. Clin. Neurophysiol 113, 1052–8.

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

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

Quiroga R. Q., Kraskov A., Kreuz T., and Grassberger P., 2002. Performance of different synchronization measures in real data: a case study on EEG signals. Physical Review E 65.

Robinson H. P. C., 2003. The biophysical basis of firing variability in cortical neurons, Chapter 6 in *Computational Neuroscience: A Comprehensive Approach,* Mathematical Biology & Medicine Series, Edited By Jianfeng Feng, Chapman & Hall/CRC.

Singer W., 2001. Consciousness and the binding problem, 2001. Annals of the New York Academy of Sciences 929, 123–146.

Stam C. J., 2005. Nonlinear dynamical analysis of EEG and MEG: review of an emerging field. Clinical Neurophysiology 116, 2266–2301.

Tallon-Baudry C., Bertrand O., Delpuech C., and Pernier J., 1996. Stimulus specificity of phase-locked and non-phase-locked 40Hz visual responses in human. Journal of Neuroscience 16, 4240–4249.

Tateno T. and Pakdaman K., 2004. Random dynamics of the Morris-Lecar neural model. Chaos 14(3).

Toups J., Fellous J.-M., Thomas P., Sejnowski T., and Tiesinga P., 2011. Finding the event structure of neuronal spike trains, Neural Computation 23(9), 2169–2208.

Tsumoto K., Kitajima H., Yoshinaga T., Aihara K., and Kawakami H., 2006. Bifurcations in Morris-Lecar neuron model. Neurocomputing 69, 293–316.

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

Varela F., Lachaux J. P., Rodriguez E., and Martinerie J., 2001. The brainweb: phase synchronization and large-scale integration. Nature Reviews Neuroscience 2(4), 229–39.

Vialatte F., Cichocki A., Dreyfus G., Musha T., Rutkowski T. M., and Gervais R., 2005. Blind source separation and sparse bump modelling of time-frequency representation of EEG signals: new tools for early detection of Alzheimer's disease. Proc. IEEE Workshop on Machine Learning for Signal Processing, 27–32.

Vialatte F., 2005. *Modélisation en bosses pour l'analyse des motifs oscillatoires reproductibles dans l'activité de populations neuronales : applications à l'apprentissage olfactif chez l'animal et à la détection précoce de la maladie d'Alzheimer,* PhD Thesis, Paris VI University, Paris.

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

Victor J. D. and Purpura K. P., 1997. Metric-space analysis of spike trains: theory, algorithms, and application. Network: Comput. Neural Systems 8(17), 127–164.

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