Quantifying the Similarity of Multiple Multi-Dimensional Point Processes by Integer Programming with Application to Early Diagnosis of Alzheimer's Disease from EEG

Justin Dauwels and Theophane Weber
M.I.T.
Cambridge, MA 02139, USA
Email: {jdauwels, theo_weber}@mit.edu

François Vialatte and Andrzej Cichocki
RIKEN Brain Science Institute
Wako-shi, Saitama, 351-0106, Japan
Email: {fvialatte,cia}@brain.riken.jp

***Abstract*—A novel approach is proposed to quantify the similarity (or "synchrony") of multiple multi-dimensional point processes. It is based on a generative stochastic model that describes how two or more point processes are related to each other. Inference in that model is carried out by integer programming. As an application, the problem of diagnosing Alzheimer's disease (AD) from multi-channel EEG recordings is considered. The proposed method seems to be more sensitive to AD induced perturbations in EEG synchrony than classical similarity measures.**

I. Introduction

Suppose we wish to quantify the correlation among multiple signals. This question is for example important in the context of neuroscience. Indeed, it has frequently been reported that certain neurological diseases such as Alzheimer's disease (AD) causes brain signals from different brain regions (e.g., from left and right hemisphere) to be less correlated [1]. Therefore, if one is capable to reliably detect degradations in brain signal coherence, one might be able to diagnose such diseases. This is a challenging problem: one is mostly interested in detecting brain diseases as early as possible, and fluctuations in brain signal coherence are then usual very weak. In earlier work [2], we compared the EEG synchrony of early-stage AD patients to age-matched control patients using a variety of interdependence measures including the Pearson correlation coefficient (and several non-linear extensions based on the theory of kernels), phase synchrony indices, mutual information and related divergence measures (e.g., Kullback-Leibler divergence), Granger causality measures, and state-space based measures [3]; we found that only very few of them were able to capture the degradation of brain signal coherence in AD patients, most similarity measures amount to about the same numerical values for both groups of subjects. There might be two reasons for this: either the fluctuations are indeed very tiny, potentially hidden in measurement noise and artifacts, which makes it notoriously hard to detect them, or classical similarity measures are simply not suitable for the problem at hand. Following this last hypothesis, we developed in [4] a novel measure for similarity ("stochastic event synchrony" or SES in short) that captures alternative aspects of similarity. The underlying idea is very simple: first we extract from each signal a point process, i.e., a sequence of events, next we try to match events from one signal to events from other signals. The better the matching, the more similar the original signals are. This approach differs from the classical approaches mentioned earlier in one important point: classical measures are usually directly computed from the original signals (either in time or time-frequency domain), in contrast, we determine the similarity based on point processes extracted from those signals. Obviously, we thereby assume that those event sequences capture the most relevant characteristics of the signals. The idea behind SES is illustrated in Fig. 1: the top row shows the wavelet transform of EEG signals recorded by two different electrodes. One can clearly see several distinct regions of strong activity, and therefore, it is natural to approximate each wavelet transform by a sequence of (half-ellipsoid) basis functions ("bumps") [5], as depicted in the bottom row; each bump is described by five parameters: its center in time and frequency, in addition to its width, height, and amplitude. In order to quantify the similarity of the two EEG signals, we align the resulting "bump models" (see Fig. 2): bumps in one time-frequency map may not be present in the other map ("spurious" bumps); other bumps are present in both maps ("non-spurious bumps"), but appear at slightly different positions on the maps. The black lines in Fig. 2 connect the centers of non-spurious bumps, and hence, visualize the offset in position between pairs of non-spurious bumps. Stochastic event synchrony consists of five parameters that quantify the alignment of two bump models:

- ρ_{spur}: fraction of spurious bumps,
- δ_t and δ_f: average time and frequency offset respectively between non-spurious bumps,
- σ_t and σ_f: standard deviation of the time and frequency offset respectively between non-spurious bumps.

We align the two bump models and determine the above parameters by iterating the following two steps [4]:

1) For given estimates of δ_t, δ_f, σ_t, and σ_f, we align the two bump models (cf. Fig. 2 (bottom)) by iterative max-

product message passing on a graphical model [6]
2) Given this alignment, the SES parameters are updated by maximum a posteriori (MAP) estimation [4].

The parameters ρ_{spur} and σ_t quantify the synchrony between bump models (and hence, the original time-frequency maps); low ρ_{spur} and σ_t implies that the two time-frequency maps at hand are well synchronized.

So far, we have developed SES for *pairs* of signals (as in Fig. 1) [4]. In practice, however, one often needs to analyze many more signals at the same time. For example, EEG is usually recorded by an array of 21, 64 or 256 electrodes [7]. In principle, one may apply SES to each pair of signals, and average the SES parameters over all those pairs, resulting in a global measure for synchrony. However, multivariate SES allows us to investigate interactions between more than two signals; for example, it enables us to identify events that occur in all signals or in a subset of signals.

The extension of SES from pairs of signals to multiple signals is the subject of this paper. This non-trivial extension involves two issues: first of all, it is not directly clear how to extend Fig. 2 to multiple signals, or in other words, how can we extend the graphical model of bivariate SES to multivariate SES? Second, once we have designed such model, how can we perform statistical inference?

This paper is organized as follows. In the following section, we outline the statistical model underlying multivariate SES, and describe how to perform inference for that model in Section III. As an illustration, we use multivariate SES in Section IV to detect AD induced perturbations in EEG synchrony. At the end of the paper, we make some concluding remarks.

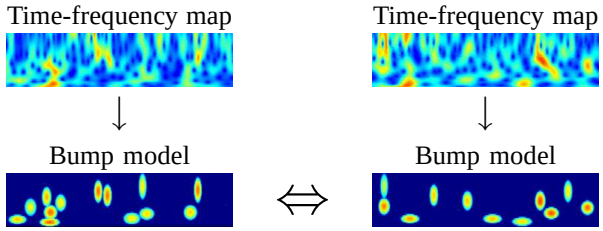
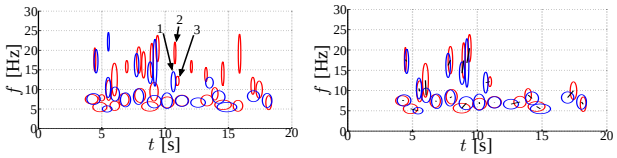


Fig. 1. Two-dimensional stochastic event synchrony.

II. Multivariate Model

We consider N signals $Y_1, \ldots, Y_N$ (for example, EEG signals) from which we extracted point processes $X_1, \ldots, X_N$ by some means, e.g., bump models (cf. Fig. 1). Each point process is a list of points ("events") in a given multi-dimensional set $S \subseteq \mathbb{R}^M$, i.e., $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}\}$ with $x_{i,k} \in S$ for $k = 1, \ldots, n_i$ and $i = 1 \ldots N$. Let us again consider the example of bump models (cf. Fig. 1, where



(a) Bump models of two EEG channels.

(b) Non-spurious bumps (ρ_{spur} = 27%); the black lines connect the centers of non-spurious bumps.

Fig. 2. Spurious and non-spurious activity.

$N = 2$). Intuitively speaking, N bump models X_i are well-synchronized if the bumps of each model appear in most other models, potentially with some small offset between them. In other words, if one overlays N well-synchronized bump models, the bumps naturally appear in clusters that contain precisely one bump from all (or almost all) bump models. This intuitive concept of similarity may readily be translated into a generative stochastic model. In that model, the N point processes X_i are treated as independent noisy observations of a hidden "mother" process $\tilde{X}$. An observation sequence X_i is obtained from $\tilde{X}$ by the following three-step procedure:

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

As a result, each X_i consists of a "noisy" copy of a non-empty subset of mother bumps. The point processes X_i may be considered well-synchronized if there only few deletions (cf. Step 2) and if the bumps of X_i are "close" to the corresponding mother bumps (cf. Step 3). One way to determine the synchrony of given point processes X_i is to first reconstruct the hidden mother process $\tilde{X}$, and to next determine the number of deletions and the average distance between the point processes X_i and the mother process $\tilde{X}$. Inferring the mother process is a high-dimensional estimation problem, the underlying stochastic model typically has a large number of local extrema. Therefore, we will use a slightly alternative procedure: we will assume that one bump in each cluster is an *identical* copy of a mother bump, the other bumps in that cluster are *noisy* copies of that mother bump. The identical copy, referred to as "exemplar", plays the role of "center" or "representative" of each cluster. We will assume, without loss of generality, that there is an exemplar for each mother bump. In addition, some point processes X_i may contain noisy copies of that mother bump, but this needs not always be the case, in other words, there might be clusters of size one, solely consisting of an exemplar. Note that under this assumption, the mother process $\tilde{X}$ is the list of all exemplars. The exemplar-based formulation amounts to the following inference problem: given the point processes X_i, we need to infer whether each bump is an exemplar or a noisy copy of some exemplar, with the constraint that each exemplar has at most one copy per point process X_i. Obviously, this inference problem also has potentially many locally optimal solutions, however, in contrast to the original (continuous) inference problem, we can find the global optimum by integer programming (see Section III).

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

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

$$p(M) = (1 - \lambda \operatorname{vol}(S))\big(\lambda \operatorname{vol}(S)\big)^M, \qquad (1)$$

where $\mathrm{vol}(S)$ is the multi-dimensional volume of set S.
- Each point $\tilde{x}_m$ for $m = 1, \ldots, M$ is uniformly distributed in S.

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

$$p(\tilde{X}) = (1 - \lambda \,\mathrm{vol}(S))\lambda^M. \tag{2}$$

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

- For each event $\tilde{x}_m$ in the mother process $\tilde{X}$, one of the point process X_i with $i \in \{1, \ldots, N\}$ is chosen at random, denoted by $X_{i(m)}$, and a copy of mother event $\tilde{x}_m$ is created in $X_{i(m)}$; this identical copy is referred to as "exemplar". For convenience, we will adopt a uniform prior $p\big(i(m) = i\big) = 1/N$ for $i = 1, \ldots, N$.
- Next, for each event $\tilde{x}_m$ in the mother process $\tilde{X}$ (with $m = 1, \ldots, M$), a "noisy" copy may be created in the point processes X_j with $j \neq i(m)$ (at most one copy per point process X_j). The number c_m of copies is modeled by a prior $p(c_m)$. In this paper, we choose as prior $p(c_m)$ a binomial distribution with $N-1$ trials and probability of success δ. Conditional on the number c_m of copies, the copies are attributed uniformly at random to other signals X_j, with the constraints of at most one copy per signal and $j \neq i(m)$; since there are $\binom{N-1}{c_m}$ possible attributions $\mathcal{A}_m \subseteq \{1, \ldots, i(m)-1, i(m)+1, \ldots, N\}$ with $|\mathcal{A}_m| = c_m$, the probability mass of an attribution $\mathcal{A}_m$ is $p(\mathcal{A}_m) = \binom{N-1}{c}^{-1}$. The process of generating a noisy copy x from a mother bump $\tilde{x}_m$ for a point process X_i is described by a conditional distribution $p(x|\tilde{x}; \theta_i)$, parameterized by some vector θ_i that may differ for each point process X_i. For the sake of simplicity, we will assume in this paper that $\theta_i = \theta$ for all i. The vector θ_i may be treated as a random vector with non-trivial (perhaps conjugate) prior $p(\theta_i)$. In the case of bump models, a simple mechanism to generate copies is to slightly shift the mother bump center while keeping the other mother bump parameters (width, height, and amplitude) fixed. The center offset may be modeled as a bivariate Gaussian random variable with non-zero mean vector $(\delta_{t,i}, \delta_{f,i})$ and diagonal non-isotropic covariance matrix $V_i = \mathrm{diag}(\sigma_{t,i}, \sigma_{f,i})$, and hence, $\theta_i = (\delta_{t,i}, \delta_{f,i}, \sigma_{t,i}, \sigma_{f,i})$.

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

$$\begin{aligned} p(\tilde{X}, X, \mathcal{I}, \theta) &= p(\theta)(1 - \lambda \,\mathrm{vol}(S))\lambda^M N^{-M} \\ &\cdot \prod_{m=1}^{M} \delta\big(x_{i(m),k(m)} - \tilde{x}_m\big) p(c_m) \binom{N-1}{c_m}^{-1} \\ &\cdot \prod_{(i,j) \in \mathcal{I}_m^{\text{copy}}} p(x_{i,j}|\tilde{x}_m; \theta). \end{aligned} \tag{3}$$

If the point processes $X = (X_1, \ldots, X_N)$ are well-synchronized, almost all processes X_i contain a copy of each mother bump $\tilde{x}_m$, and therefore, the sets $\mathcal{I}_m^{\text{copy}}$ are either of size $N-1$ or are slightly smaller. Moreover, in the case of bump models, the standard deviations $\sigma_{t,i}$ and $\sigma_{f,i}$ are then small. Therefore, given point processes $X = (X_1, \ldots, X_N)$, we wish to infer $\mathcal{I}$ and θ, since those variables contain information about similarity.

III. Inference as integer program

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

$$(\hat{\mathcal{I}}, \hat{\theta}) = \operatorname*{argmax}_{(\mathcal{I}, \theta)} \log p(\tilde{X}, X, \mathcal{I}, \theta). \tag{4}$$

There is no closed form expression for (4), therefore, we need to resort to numerical methods. A simple technique to try to find (4) is cyclic maximization: We first choose initial values $\hat{\theta}^{(0)}$, and then perform the following updates for $r \geq 1$ until convergence:

$$\hat{\mathcal{I}}^{(r)} = \operatorname*{argmax}_{\mathcal{I}} \log p(\tilde{X}, X, \mathcal{I}, \hat{\theta}^{(r-1)}) \tag{5}$$

$$\hat{\theta}^{(r)} = \operatorname*{argmax}_{\theta} \log p(\tilde{X}, X, \hat{\mathcal{I}}^{(r)}, \theta). \tag{6}$$

This procedure is guaranteed to converge as long as the conditional maximizations (5) and (6) have unique solutions, which is in practice most often the case.

The update (6) of the parameters θ is usually available in closed form, on the other hand, the update (5) is less trivial and requires some additional attention. We show that it can be written as an integer program, which may be solved by classical integer programming techniques or by max-product message passing on a sparse cyclic factor graph of $p(\tilde{X}, X, \mathcal{I}, \theta)$ (3). In order to formulate that integer program, we introduce the following variables:

- $b_{i,k}$ is a binary variable equal to one if and only if the k-th event of X_i is an exemplar.
- $b_{i,k,i',k'}$ is a binary variable equal to one if and only if the k-th event of X_i is a copy of the exemplar $x_{i',k'}$.
- $b_{i,i',k'}$ is a binary variable equal to one if and only if no event of X_i is a copy of exemplar $x_{i',k'}$.

Note that $b_{i,k,i,k'} = 0$ for all k and k' and $b_{i,i,k'} = 1$ for all i and k', since X_i must not contain a noisy copy of a mother event $\tilde{x}_m$ if it already contains the exemplar associated to $\tilde{x}_m$. With this parametrization, the conditional maximization (5) may be cast as the problem of inferring the above variables b:

$$\begin{aligned} \min_{b} \quad & \alpha \sum_{i, 1 \leq k \leq n_i} b_{i,k} + \beta \sum_{i, i' \neq i, 1 \leq k' \leq n_{i'}} b_{i,i',k'} \\ & + \sum_{i,i', 1 \leq k \leq n_i, 1 \leq k' \leq n_{i'}} -\log p(x_{i,k}|x_{i',k'}; \theta)\, b_{i,k,i',k'} \end{aligned} \tag{7}$$

subject to

$$\forall i,k, \qquad \sum_{i',k'} b_{i,k,i',k'} + b_{i,k} = 1 \tag{8}$$

$$\forall i, i' \neq i, k', \quad b_{i,i',k'} = b_{i',k'} - \sum_{1 \leq k \leq n_i} b_{i,k,i',k'}, \tag{9}$$

where

$$\alpha = \log \lambda + (N-1) \log \delta \text{ and } \beta = \log \left(\frac{1-\delta}{\delta} \right). \tag{10}$$

This optimization problem may be solved by max-product message passing on a sparse graph of $p(\tilde{X}, X, \mathcal{I}, \theta)$ (3), along the lines of the algorithm of [4]; we omit the details here due to space constraints. We implemented this approach for the problem described in Section IV, but observed that it does not perform well, the solutions it provide are suboptimal and not very useful. As an alternative, we solved the problem by classical integer programming techniques, using CPLEX (Ilog) [8]. Not only is this approach guaranteed to find the optimal solution, but rather unexpectedly, it also ran vastly faster (a factor of 10 to 100) than the max-product algorithm.

IV. Application to diagnosis of AD from EEG

We analyzed rest eyes-closed EEG data recorded from 21 sites on the scalp based on the 10–20 system [7]. The sampling frequency was 200 Hz, and the signals were band-pass filtered between 4 and 30Hz. The subjects comprises two study groups: the first consists of 22 patients diagnosed as suffering from mild cognitive impairment (MCI), who subsequently developed mild AD. The other group is a control set of 38 age-matched, healthy subjects who had no memory or other cognitive impairments. Pre-selection was conducted to ensure that the data were of a high quality, as determined by the presence of at least 20s of artifact free data. We aggregated the 21 bump models in five regions (frontal, temporal left and right, central, occipital) by means of the aggregation algorithm described in [5], resulting in a bump model for each of those five regions ($N = 5$).

As we pointed out earlier, in previous work [2], we have applied a large variety of *classical* synchrony measures (more than 30 in total, including various information-theoretic measures) to both data sets with the aim of detecting MCI induced perturbations in EEG synchrony; we observed that none of those classical measures except full frequency Directed Transfer Function (ffDTF) [3], which is a Granger causality measure, was able to detect significant loss of EEG synchrony in MCI patients. More precisely, all measures amount to Mann-Whitney p-values larger than 0.005 with the exception of ffDTF ($p = 0.0012$)—as a reminder, the smaller the p-value, the more the statistics of the quantity at hand differ in both groups. On the other hand, bivariate SES resulted in significant differences between both subject groups, in particular, there was a significant increase of ρ_{spur} ($p = 2 \cdot 10^{-4}$). This seems to indicate that there is an increase of unsynchronized spurious activity in AD patients. Interestingly and perhaps surprisingly, the timing jitter σ_t was in both subject groups about the same, in other words, the non-spurious activity is equally well synchronized in both populations.

The results for multivariate SES are summarized in Table I; we studied the following statistics:

- (Posterior) distribution $p(c_m = i|X) = p_i^c$ of the number of copies of each exemplar c_m, parameterized by $(p_0^c, p_1^c, \ldots, p_4^c)$,
- σ_t: standard deviation in time domain ("time jitter"),
- σ_f: standard deviation in frequency domain ("frequency jitter").

We also consider the linear combination h^c of all parameters p_i^c that optimally separates both subject groups. Interestingly, the latter statistic amounts to about the same p-value as the index ρ_{spur} of bivariate SES. The posterior $p(c_m|X)$ mostly differs in p_4^c: in MCI patients, the number of clusters of size five significantly decreases, which in turn causes an increase in ρ_{spur}. Combining h^c with ffDTF and σ_f allows to separate the two groups quite well (about 90% correctly classified), as shown in figures 3 and 4, far better than what can be achieved by means of classical similarity measures (about 75% correctly classified).

Stat.	p_0^c	p_1^c	p_2^c	p_3^c	p_4^c	h^c	σ_t	σ_f
p-value	0.07	0.08	0.23	0.03*	0.0013**	2.10^{-4}**	0.12	9.10^{-4}**

TABLE I
SENSITIVITY OF MULTIVARIATE SES FOR DIAGNOSING AD (P-VALUES FOR MANN-WHITNEY TEST; * AND ** INDICATE $p < 0.05$ AND $p < 0.005$ RESPECTIVELY)

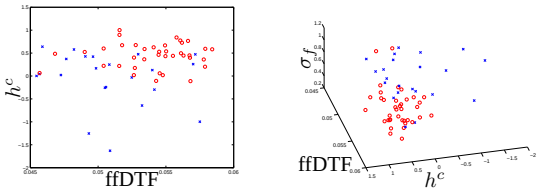


Fig. 3. Classification with SES,ffDTF (left) and SES, ffDTF and σ_F (right)

V. Conclusion

We proposed a novel approach to determine the synchrony of multiple multi-dimensional point processes; it is based on a graphical model that describes how the point processes are related through a common hidden "mother" process. The proposed technique may be used for various applications in neuroscience (e.g., in brain-computer interfaces), biomedical signal processing, and beyond.

References

[1] J. Jong, "EEG Dynamics in Patients with Alzheimer's Disease," *Clinical Neurophysiology*, 115:1490–1505 (2004).

[2] J. Dauwels, F. Vialatte, and A. Cichocki, "A Comparative Study of Synchrony Measures for the Early Detection of AD," *Proc. ICONIP 2007*, Kitakyushu, September 2007.

[3] E. Pereda, R. Q. Quiroga, and J. Bhattacharya, "Nonlinear Multivariate Analysis of Neurophsyiological Signals," *Progress in Neurobiology*, 77 (2005) 1–37.

[4] J. Dauwels, F. Vialatte, T. Rutkowski, and A. Cichocki, "Measuring Neural Synchrony by Message Passing," *Proc. NIPS 2007*, Vancouver, Canada, December 2007.

[5] F. Vialatte, C. Martin, R. Dubois, J. Haddad, B. Quenet, R. Gervais, and G. Dreyfus, "A Machine Learning Approach to the Analysis of Time-Frequency Maps, and Its Application to Neural Dynamics," *Neural Networks*, 2007, 20:194–209.

[6] H.-A. Loeliger, "An Introduction to Factor Graphs," *IEEE Signal Processing Magazine*, Jan. 2004, pp. 28–41.

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

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