



Methods for alignment of multi-class signal sets

Patrik Wahlberg*, Göran Salomonsson

Signal Processing Group, Department of Electrosence, Lund University, Box 118, S-221 00 Lund, Sweden

Received 19 July 1999; received in revised form 2 December 2002

Abstract

The paper treats jitter estimation for alignment of a set of signals which contains several unknown classes of waveforms. The motivating application is epileptic EEG spikes, where alignment prior to clustering and averaging is desired. The assumption that the signal waveforms are unknown precludes the use of classical techniques, notably matched filtering. Instead we treat two other classes of methods. In the first class the jitter of each signal is estimated with aid of the whole data set, using the Rayleigh quotient of the sample correlation matrix. The main idea of the paper is the suggestion of two such methods, consisting respectively of mean value computation and maximization of the Rayleigh quotient as a function of translation of a given signal. In the second class of methods each signal is processed individually, and one such method is estimation of the jitter of a signal by its centre of gravity. By means of deduction, simulations and evaluation on real life epileptic EEG signals, we show that the first class of methods is preferable to the second. Simulations also show that the method of maximization of the Rayleigh quotient seems to be a generally good method, which gives small estimation error and is applicable in a wide range of circumstances. For seven investigated sets of real life EEG data, the maximization algorithm turned out to give the best results, and improved alignment in the majority of signal clusters.

© 2003 Elsevier Science B.V. All rights reserved.

Keywords: Time jitter; Multi-class signal sets; Clustering; Alignment; Rayleigh quotient; Epileptic EEG spikes

1. Introduction

There is in medical signal processing a problem of clustering a given set of multichannel EEG transients from epileptics, so-called *spikes* [24,9]. The spikes are important for diagnostics since they often originate from the *epileptic foci*. The result of the clustering can be used primarily to see whether there are distinct groups of spikes and to study their properties, and secondarily to compute SNR-improving

averages within clusters prior to source localization, e.g. by dipole analysis [20]. The selection of spikes to be clustered can be made automatically [2], or manually by a neurophysiologist. In either case, the signals of each cluster are not perfectly aligned in the time dimension. Such variability in synchronization is called *time jitter*, and its presence makes the clustering problem more difficult and leads to less distinct or unseparable clusters, since separation between clusters is sometimes subtle also in the absence of jitter. Beside the epileptic spike clustering application, the problem exists also for evoked potentials [26], where data sets sometimes consist of several classes.

Problems connected to time jitter and delay estimation have been treated in communication and detection

* Corresponding author. Tel.: +46-46-2229017;
fax: +46-46-2224718.

E-mail addresses: patrik.wahlberg@es.lth.se (P. Wahlberg),
salomonsson.g@telia.com (G. Salomonsson).

theory [3,6]. Also in the biomedical signal processing literature, the problem of time jitter has been treated in many respects. In the field of evoked potentials, there is an interest in synchronizing, i.e. compensation for jitter, before averaging [25,14,5,16], and the jitter values can also be of interest as medical information in itself, and their behaviour as a function of time [12]. In ECG signal processing, researchers are interested in alignment [1,8,11,10,13], and signal models are used which incorporate jitter in a wider class of signal transformations [22]. A problem related to alignment is characterization of the jitter in the sense of estimating its statistical parameters [15].

Most reported techniques for jitter compensation use the assumption that the signals to be aligned belong to one unknown class, or in telecommunication problems to one of several classes whose waveforms are known. In these situations matched filtering techniques can be used, with the ensemble average as template signal in the one-class case. In the clustering application however, the assumptions are weaker which makes the problem more difficult. The technique of matched filtering of each signal with the sample average as template [25] may lead to poor estimates in the case of several signal classes (the average may be zero). Once a good clustering exists, the one-class techniques can be employed to each class, but the actual determination of the classes, i.e. the clustering, is made more difficult if the signals are not aligned. Clearly then, there is a need for methods for jitter compensation prior to clustering multi-class data.

Another way of attacking the problem is to do *pairwise* alignment of all signal pairs. This gives a representation of all metrics between signals and calls for graph-theoretical clustering methods [7]. These have augmented computational complexity compared to k -means clustering algorithms [4] which work on linearly represented data. We assume in this paper the use of linearly represented data, i.e. we do not treat pairwise alignment and graph-theoretical clustering.

In this paper we work according to the principle of alignment prior to clustering. A method employing the opposite order of these operations, i.e. clustering prior to alignment within each cluster, has been reported [11]. This method uses a metric which is invariant to translations, and a modification of the ordinary k -means clustering algorithm. We show in this paper that this translation-invariant metric

requires a high SNR to work satisfactorily for clustering, i.e. to represent cluster differences sufficiently well to enable correct clustering. We are aware of no other translation-invariant metric which lacks this flaw. For small SNR, which generally is the case of the epileptic EEG spikes, the principle of alignment prior to clustering is therefore motivated.

The essential content of the present paper is the proposal of two methods for jitter estimation. We suggest using the Rayleigh quotient of the data correlation matrix in order to obtain a function which can be used to construct estimators, whereby the desired property of incorporation of the data sample statistics is solved. The function consists of evaluation of the Rayleigh quotient as a function of translation of a given signal. The Rayleigh quotient of the correlation matrix can be expressed in terms of the eigenvectors, and alignment computation using the first eigenvector has been treated in [1] for ECG processing. However, this method is not adapted to the case of several signal classes.

We give a review of classical estimation theory in this context, where we focus on the minimum mean square error estimator (MMSE), and the maximization of the a posteriori probability estimator (MAP), which both can be regarded as Bayes estimators [19,23], and also the maximum likelihood estimator (ML). However, the probability density functions required for computation of these estimators are inherently not available in our problem formulation, even though there is a connection between ML and one of our suggested estimators under certain circumstances. We use MMSE for comparison with realistic estimators in the simulations.

A straightforward method for estimation of jitter is to compute the centre of gravity in the time dimension of each signal. A bias is introduced if the non-jittered signal has centre of gravity which not equals zero, but this bias is of small concern since it is a constant which do not affect the average within clusters. What is worse is that the method also gives bias that varies as a function of the jitter, which deteriorates averages. Moreover, it does not use the statistics of the given data set, instead all signals are treated individually.

We will start by treating the case of scalar-valued functions (i.e. vectors when the time axis is discrete) and later we will generalize to vector-valued functions (matrices), which is necessary for the epileptic

spike application, where data is measured at several channels.

The paper is organized as follows. Section 2 treats problem formulation and notations, Section 3 presents two new methods which are based on the Rayleigh quotient of the sample correlation matrix, and Section 4 describes the centre of gravity method. In Section 5 the theory and notations are extended to vector-valued functions, numerical experiments are reported in Section 6, and Section 7 gives results for real life EEG data.

2. Problem formulation

We use t to denote continuous time, n to denote discrete time, and by $x(t)$ we denote a signal x at time t . In the clustering problem a set $\{x_k(t)\}_{k=1}^K$ of K signals $x_k(t) \in L_2(\mathbf{R})$ is given. We assume that each signal is a sum of a deterministic, unknown *pattern signal* $m_j(t) \in L_2$ from one of L clusters (or classes) denoted $\omega_1, \dots, \omega_L$, and Gaussian noise $e_k(t) \in L_2$. The patterns m_j are assumed to have derivatives of all orders of quickly decreasing L_2 norm, and the norms of the pattern functions are roughly equal $\|m_j\| \approx \|m_i\|$, and well separated. Using $\|m_j\| \approx \|m_i\|$, the separation between patterns can be formulated $|\langle m_i, m_j \rangle| \ll \|m_i\|^2$, $i \neq j$, where $\langle \cdot, \cdot \rangle$ denotes the L_2 Hilbert space inner product $\langle m_i, m_j \rangle = \int m_i(t)m_j(t) dt$. Further we assume that the pattern signal of x_k is delayed τ_k time units, where τ_k is a sample of a Gaussian stochastic variable of zero mean and variance σ_τ^2 . Statistical independence between the noise e and τ is assumed. The model is hence

$$x_k(t) = m_j(t - \tau_k) + e_k(t), \quad k = 1, \dots, K, \quad (1)$$

where for each k , the a priori class probabilities p_j are assumed equal $p_j = 1/L$, $1 \leq j \leq L$. Sometimes we drop the index k of $x_k(t)$ for convenience. The delayed signal $m(t - \tau)$ is also denoted $m_\tau(t)$. It can be shown that the effect of the time jitter is $E\{m_{j,\tau}(t)\} = m_j * f_\tau(t)$ [18], i.e. the mean value is the convolution between m_j and the jitter probability density f_τ , whereby it is clear that the jitter distorts the means of the pattern signals.

The problem we want to solve is to find an estimator $\hat{\tau}(x)$ which gives small variance

$$E\{(\hat{\tau}(x) - E\hat{\tau}(x))^2\}. \quad (2)$$

The estimates are intended to be used for alignment. For this purpose each raw signal is inversely translated an amount which equals its estimated jitter, i.e. the aligned signal is defined by

$$\tilde{x}(t) \stackrel{\text{def}}{=} x(t + \hat{\tau}(x)). \quad (3)$$

For implementation of the estimators we discretize the time axis to N (=odd) points symmetrically around the origin, which transforms the function $x(t)$ into an $(N \times 1)$ -vector denoted \mathbf{x} and defined by

$$\mathbf{x} = \left[x\left(-\frac{N-1}{2}T_s\right) \cdots x(0) \cdots x\left(\frac{N-1}{2}T_s\right) \right]^T. \quad (4)$$

Here T_s denotes the sampling time interval, which always is understood but from here on omitted in the notation, i.e. we denote $\mathbf{x} = [x(-(N-1)/2) \cdots x(0) \cdots x(N-1)/2]^T$. Likewise \mathbf{m}_j denotes the discretized $m_j(t)$. We assume the support of $m_{j,\tau}$ for possible jitter values τ is contained in the interval $[-(N-1)/2, (N-1)/2]$. The norm $\|\cdot\|$ sometimes denotes the L_2 norm, and sometimes the Euclidean norm in \mathbf{R}^N , depending on context. For sums over n denoted $\sum_n(\cdot)$, the limits $\sum_{n=-(N-1)/2}^{(N-1)/2}(\cdot)$ are understood.

There exists a way of attacking the multi-class jitter problem the opposite way of our method. Instead of alignment before clustering, clustering can be performed before alignment (which is then computed within each cluster) [11]. The metric used in the clustering method (e.g. the k -means method [4]) is then modified from the ordinary Euclidean metric to a translation-invariant metric, and cluster representatives are not obtained by averaging within clusters but by choosing the cluster vector that minimizes the sum of metrics to all other cluster vectors. The translation-invariant metric uses the normalized integral method [8,11,10,13], where the normalized integral of signal $x_k(t)$ can be defined as

$$X_k(t) \stackrel{\text{def}}{=} \frac{\int_{-\infty}^t x_k^2(\tau) d\tau}{\int_{-\infty}^{+\infty} x_k^2(\tau) d\tau}. \quad (5)$$

The metric between signals $x_k(t)$ and $x_l(t)$ is computed using the function $t' = \phi(t)$, defined by

$$X_k(t) = X_l(t'). \quad (6)$$

If $x_\ell(t)$ is a translated version of $x_k(t)$ the function ϕ is an affine function $\phi(t) = at + b$, whereby it is clear that the metric Δ defined by

$$\Delta^2(x_k, x_\ell) \stackrel{\text{def}}{=} \inf_{a,b} \int \|\phi(t) - at - b\|^2 dt, \quad (7)$$

which computes the deviation from affinity of the function ϕ , is invariant to translations (in practice a finite sum replaces the integral). In order to get symmetry, the modification $\Delta^*(x_k, x_\ell) = \frac{1}{2}(\Delta(x_k, x_\ell) + \Delta(x_\ell, x_k))$ can be used [11].

In order to investigate the metric Δ^* we performed a simulation study. We generated $K = 300$ signals from $L = 3$ classes, where each pattern was chosen as a discretized Hermite function [24] $\mathbf{m}_j = H_{j-1}$, i.e. $\mathbf{m}_1 = H_0 = a$ Gaussian, $\mathbf{m}_2 = H_1 =$ the first derivative of a Gaussian, and $\mathbf{m}_3 = H_2 =$ the second derivative of a Gaussian. The signal vectors had length $N = 49$ samples and the patterns \mathbf{m}_j were centred within the interval. The jitter density function f_τ was chosen Gaussian, with three variances $\sigma_\tau^2 = 1$, $\sigma_\tau^2 = 5$ and $\sigma_\tau^2 = 15$ samples² respectively, and the noise \mathbf{e} was chosen zero mean, white and Gaussian, giving a range of SNR values using the definition

$$\text{SNR} \stackrel{\text{def}}{=} 10 \log_{10} \left(\frac{1}{K} \sum_{k=1}^K \frac{\|\mathbf{x}_k - \mathbf{e}_k\|^2}{\|\mathbf{e}_k\|^2} \right). \quad (8)$$

We compared the within-cluster variation and the separation between cluster representatives, measured by (i) the Euclidean metric $\|x_k - x_\ell\|$ and (ii) the metric $\Delta^*(x_k, x_\ell)$, as a function of SNR. Fig. 1(a) shows the within-cluster variation, the metric Δ^* gives smaller values than the Euclidean metric which is expected since Δ^* is invariant to jitter. However, as seen in Fig. 1(b) also the separation between cluster representatives is smaller for Δ^* , which can lead to problems for a clustering algorithm. For each of the two metrics classification error rates were obtained using classification of each signal by minimization over metrics to cluster representatives, obtained using the true clustering (we avoided using a clustering algorithm since error then also depends on the clustering algorithm). The error rates are shown in Fig. 1(c), and it can be seen that the metric Δ^* requires quite a high SNR, in the order of 20–25 dB, to work satisfactorily, i.e. to reach error rates smaller than 0.1. For such large SNR values and $\sigma_\tau^2 \geq 5$, the metric Δ^* improves upon the error rates of the Euclidean metric,

as expected. The Euclidean metric is essentially independent of the SNR apart from negative values where the error rates increase. We conclude that the metric Δ^* works well only in the case of high SNR. Otherwise it mixes clusters such that the error rate upon nearest-representative classification increases. The epileptic EEG signals have generally SNR smaller than 20 dB, so for this application the metric Δ^* is not suitable and the principle of alignment prior to clustering is motivated.

3. Methods incorporating the sample statistics

In this section we will relate the jitter estimation problem to general estimation theory, and suggest two estimators which are based on the Rayleigh quotient of the correlation function of the given data set. The vectors \mathbf{x} (ignoring the k index) are regarded as samples of a stochastic variable X , and the posterior density function for τ conditioned on X is defined by

$$f_{\tau|X=\mathbf{x}}(t) \stackrel{\text{def}}{=} \frac{f_{\tau,X}(t, \mathbf{x})}{f_X(\mathbf{x})} = \frac{f_{X|\tau=t}(\mathbf{x}) \cdot f_\tau(t)}{f_X(\mathbf{x})}. \quad (9)$$

Here $f_{\tau,X}(t, \mathbf{x})$ denotes the joint density function for τ and X , $f_X(\mathbf{x})$ the marginal density for X , and $f_{X|\tau=t}(\mathbf{x})$ is the density function for X conditioned on $\tau = t$. The estimator which minimizes the mean square error (MSE), denoted $\hat{\tau}_{\text{mmse}}$, belongs to the class of Bayes estimators, and it can be shown to be equal to the expectation of the posterior density function [19,23, p. 54],

$$\hat{\tau}_{\text{mmse}}(\mathbf{x}) \stackrel{\text{def}}{=} E\{\tau|X = \mathbf{x}\} = \int t \cdot f_{\tau|X=\mathbf{x}}(t) dt. \quad (10)$$

In order to use this formula with aid of (9), it is clear that only the nominator $f_{X|\tau=t}(\mathbf{x}) \cdot f_\tau(t)$ of (9) is of interest since the denominator $f_X(\mathbf{x})$ is a multiplicative constant, which is independent of t and thus the integration. The constant $f_X(\mathbf{x})$ just makes the integral of the posterior density $f_{\tau|X=\mathbf{x}}(t)$ equal to one. If the function $f_{\tau|X=\mathbf{x}}(t)$ has a unique maximum around which it is symmetric, the estimator $\hat{\tau}_{\text{mmse}}$ reduces to the maximum a posteriori (MAP) estimator [19,23, p. 54], defined by

$$\hat{\tau}_{\text{map}}(\mathbf{x}) \stackrel{\text{def}}{=} \arg \max_t \{f_{\tau|X=\mathbf{x}}(t)\}. \quad (11)$$

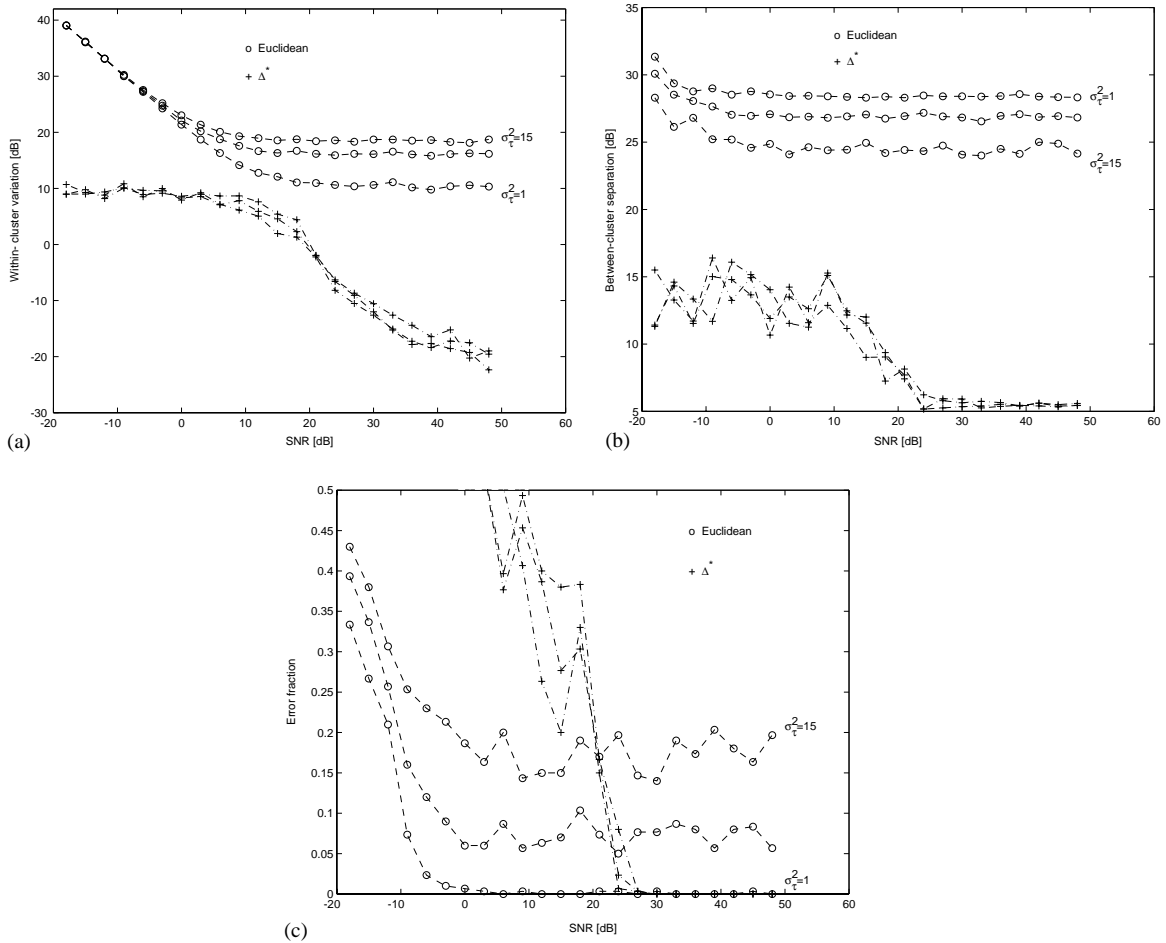


Fig. 1. Within-cluster variation (a), cluster separation (b) and error rates (c) for the two metrics $\|\cdot\|$ and Δ^* .

Hence, if the posterior density $f_{\tau|X=\mathbf{x}}(t)$ is Gaussian the estimates $\hat{\tau}_{\text{mmse}}$ and $\hat{\tau}_{\text{map}}$ coincide. The property that the posterior density is Gaussian is also a necessary condition for an estimator which obtains the Cramér–Rao lower bound of variance [23, p. 72]. For the jitter estimation problem, the posterior density $f_{\tau|X=\mathbf{x}}(t)$ is generally *not* Gaussian, even if τ and e are Gaussian, so $\hat{\tau}_{\text{mmse}}(\mathbf{x}) \neq \hat{\tau}_{\text{map}}(\mathbf{x})$ generally. The reason is that the delayed pattern signal $\mathbf{m}_{j,\tau}$ is a non-linear function of τ .

We consider the jitter sample τ as a stochastic variable, but instead it can be regarded as a fixed but unknown *parameter* of the signal \mathbf{x} . A natural estimator is in this case the maximum likelihood (ML)

estimator [23, p. 63], defined by

$$\hat{\tau}_{\text{ml}}(\mathbf{x}) \stackrel{\text{def}}{=} \underset{t}{\text{argmax}} \{f_{X|\tau=t}(\mathbf{x})\}. \quad (12)$$

It can be considered to be the limiting case of the MAP estimator when the density f_τ approaches uniform. If $f_{X|\tau=t}(\mathbf{x})$ is a sum of L white Gaussian densities and the patterns \mathbf{m}_j are sufficiently separated, the ML estimator is equivalent to the least square (LS) estimator [21]

$$\begin{aligned} \hat{\tau}_{\text{ls}}(\mathbf{x}) &\stackrel{\text{def}}{=} \underset{t}{\text{argmin}} \{\|\mathbf{x} - \mathbf{m}_t\|^2\} \\ &= \underset{t}{\text{argmin}} \{\|\mathbf{x}_{-t} - \mathbf{m}\|^2\}, \end{aligned} \quad (13)$$

where \mathbf{m} denotes the pattern from which \mathbf{x} originates, i.e. we use the model $\mathbf{x} = \mathbf{m}_{\tau_0} + \mathbf{e}$ with τ_0 the jitter value. The last equality is obtained from a change of variable. The requirement of sufficient separation between the patterns \mathbf{m}_j is necessary since otherwise $\min_{t,j} \|\mathbf{x} - \mathbf{m}_{j,t}\|^2$ may occur for $\mathbf{m}_j \neq \mathbf{m}$. The LS estimator can be constrained to be linear [6]. By evaluation of the square the estimator (13) can be implemented as the *matched filtering* estimator [6]

$$\hat{\tau}_{mf}(\mathbf{x}) = \arg \max_t \{\mathbf{x}_{-t}^T \mathbf{m}\}. \tag{14}$$

The estimators $\hat{\tau}_{ls}$ and $\hat{\tau}_{mf}$ are identical provided $\|\mathbf{x}_{-t}\|$ is constant as a function of t . This holds only approximately, for small t , using the assumption that the support of m_τ is contained in the interval $[-(N-1)/2, (N-1)/2]$. However, the approximation error can be assumed small since only small values of t are candidates for the maximization (14). Using the model $\mathbf{x} = \mathbf{m}_{\tau_0} + \mathbf{e}$, we have $\mathbf{x}_{-t}^T \mathbf{m} \approx \mathbf{m}_{\tau_0-t}^T \mathbf{m}$ if $\|\mathbf{e}\|$ is small compared to $\|\mathbf{m}\|$. Thus the function whose maximum defines the estimator (14) is approximately the autocorrelation of \mathbf{m} , with translated argument $(\tau_0 - \cdot)$. Using this approximation gives $\arg \max_t \{\mathbf{m}_{\tau_0-t}^T \mathbf{m}\} = \tau_0$, i.e. a correct estimate.

The exact implementation of the estimators (10) and (11) is unrealistic in the framework of our problem definition, since the factors of the posterior density $f_{\tau|X=\mathbf{x}}(t)$, i.e. $f_\tau(t)$ and $f_{X|\tau=t}(\mathbf{x})$ as a function of t for given \mathbf{x} , are unknown. They are not available unless one has knowledge of both the jitter density f_τ and the functions m_j , the determination of which belongs to the domain of the clustering problem. In fact, using the model (1) the factor $f_{X|\tau=t}(\mathbf{x})$ is

$$f_{X|\tau=t}(\mathbf{x}) = \sum_{j=1}^L p_j f_e(\mathbf{x} - \mathbf{m}_{j,t}), \tag{15}$$

where f_e is the density for the noise \mathbf{e} . It is seen that $f_{X|\tau=t}(\mathbf{x})$ depends on the unknown vectors $\{\mathbf{m}_j\}_{j=1}^L$. Considering the linear least square estimator [6] also leads to difficulties. Namely, the cross-correlation vector $E\{\tau X\}$ is then needed, and it can be shown to approximately equal $-\sigma_\tau^2(1/L) \sum_{j=1}^L \mathbf{m}'_j$ where \mathbf{m}'_j denotes the discretized derivative of \mathbf{m}_j . The cross-correlation vector can be estimated by $-\sigma_\tau^2(1/K) \sum_{k=1}^K \mathbf{x}'_k$, but this quantity may be zero, leading to an estimator identical to zero, and further σ_τ is not available.

Instead of the above estimators we shall formulate a criterion depending on the translation of a given signal, which is computed with aid of the sample correlation matrix of the data set. For the rest of this section we use the model

$$\mathbf{x} = \mathbf{m}_{\tau_0} + \mathbf{e}, \tag{16}$$

where $\mathbf{m} = \mathbf{m}_j$ for some j , i.e. \mathbf{x} belongs to class ω_j , is delayed τ_0 time units and corrupted by noise \mathbf{e} . Given the signal vector \mathbf{x} we wish to define a criterion of the matching of the translated signal $\mathbf{x}_t = [x(-(N-1)/2-t) \cdots x((N-1)/2-t)]^T$ to the signal space containing signals of the same pattern as \mathbf{x} and whose jitter values are statistically most significant. From such a criterion as a function of t , estimators of the actual jitter of \mathbf{x} can be constructed. A tentative criterion is the squared projection distance

$$h_1(\mathbf{x}, t, r) = \|\mathbf{x}_t - \mathbf{Q}_r \mathbf{Q}_r^T \mathbf{x}_t\|^2, \tag{17}$$

where $r \leq N$, $\mathbf{Q}_r = [\mathbf{q}_1 \cdots \mathbf{q}_r]$, and \mathbf{q}_i are eigenvectors of the sample correlation matrix

$$\mathbf{R} = \frac{1}{K} \sum_{k=1}^K \mathbf{x}_k \mathbf{x}_k^T, \tag{18}$$

corresponding to the eigenvalues λ_i . These are non-negative since \mathbf{R} is non-negative definite, and decreasingly ordered, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$. The criterion (17) measures the squared error of the projection of \mathbf{x}_t on the space spanned by the first r eigenvectors of \mathbf{R} , whereby small values of (17) will be reached when t is such that \mathbf{x}_t is close to the statistically most significant signal space. However, the criterion (17) requires a choice of r , which is non-trivial. In order to modify (17) for elimination of this decision we reformulate (17) according to

$$h_1(\mathbf{x}, t, r) = \|(I - \mathbf{Q}_r \mathbf{Q}_r^T) \mathbf{x}_t\|^2 = \|\mathbf{Q} \mathbf{D} \mathbf{Q}^T \mathbf{x}_t\|^2, \tag{19}$$

where $\mathbf{Q} = [\mathbf{q}_1 \cdots \mathbf{q}_N]$, and $\mathbf{D} = \text{diag}([0 \cdots 0 \ 1 \cdots 1])$ whose first r diagonal elements are zero. The essence of the criterion function h_1 lies in the monotonically increasing diagonal of \mathbf{D} , and it can as such be replaced by, using the notation $A = \text{diag}([\lambda_1 \cdots \lambda_N])$,

$$(\lambda_1 I - A)^{1/2} = \text{diag}([0 (\lambda_1 - \lambda_2)^{1/2} \cdots (\lambda_1 - \lambda_N)^{1/2}]), \tag{20}$$

which is a matrix that needs no decision of r , in contrast to \mathbf{D} . The corresponding criterion to be

minimized is

$$h_2(\mathbf{x}, t) = \|\mathbf{Q}(\lambda_1 I - \Lambda)^{1/2} \mathbf{Q}^T \mathbf{x}_t\|^2 = \lambda_1 \|\mathbf{x}_t\|^2 - \mathbf{x}_t^T \mathbf{R} \mathbf{x}_t. \quad (21)$$

The assumption of the support of m_{τ_0} being contained in the interval $[-(N-1)/2, (N-1)/2]$ gives that $\|\mathbf{x}_t\|$ is approximately constant as a function of t for values of t such that $|t + \tau_0| \leq |\tau_0|$. For other values of t this is not assured since $m_{t+\tau_0}$ may be translated out of $[-(N-1)/2, (N-1)/2]$. However, for the t values where h_2 is likely to have its minimum, i.e. around $t = -\tau_0$, $\|\mathbf{x}_t\|^2$ is approximately constant. Because of this, minimization of (21) can be replaced by maximization of

$$g(\mathbf{x}, t) \stackrel{\text{def}}{=} \mathbf{x}_t^T \mathbf{R} \mathbf{x}_t. \quad (22)$$

In words, the function (22) is computed by first translating \mathbf{x} to the right t time units and then evaluating the Rayleigh quotient defined by \mathbf{R} . More exactly, the Rayleigh quotient is defined by

$$\frac{\mathbf{x}_t^T \mathbf{R} \mathbf{x}_t}{\|\mathbf{x}_t\|^2}, \quad (23)$$

but since $\|\mathbf{x}_t\| \approx \|\mathbf{x}\|$ independently of t for $|t + \tau_0| \leq |\tau_0|$, the denominator is almost a multiplicative constant and does not influence the maximization.

The correlation matrix \mathbf{R} can be approximated with aid of a Taylor series expansion (see Appendix A) according to

$$\mathbf{R} \approx \mathbf{R}_e + \sum_{j=1}^L p_j (\mathbf{m}_j \mathbf{m}_j^T + \sigma_\tau^2 \mathbf{m}'_j \mathbf{m}'_j{}^T), \quad (24)$$

where \mathbf{R}_e is the correlation matrix of the noise vector \mathbf{e} , and \mathbf{m}'_j is the discretized derivative of m_j . Using (24), the function $g(\mathbf{x}, t)$ can be approximated by

$$g(\mathbf{x}, t) \approx \mathbf{x}_t^T \mathbf{R}_e \mathbf{x}_t + \sum_{j=1}^L p_j ((\mathbf{x}_t^T \mathbf{m}_j)^2 + \sigma_\tau^2 (\mathbf{x}_t^T \mathbf{m}'_j)^2). \quad (25)$$

For the next step we assume (i) that the noise is white and has variance $\sigma_e^2 \ll \|\mathbf{m}_j\|^2 \forall j$, and (ii) $\|\mathbf{m}'_j\| \ll \|\mathbf{m}_j\| \forall j$, i.e. the norms of the derivatives are a magnitude in order smaller than the signal norms. Hereby

(25), in turn, can be approximated by

$$g(\mathbf{x}, t) \approx \sigma_e^2 \|\mathbf{x}\|^2 + p(\mathbf{m}_{\tau_0+t}^T \mathbf{m})^2 + \sum_{m_j \neq m} p_j (\mathbf{m}_{\tau_0+t}^T \mathbf{m}_j)^2, \quad (26)$$

where p is the a priori probability of the cluster ω_j to which \mathbf{m} belong. The second term of (26), i.e. $g_1(t) = p(\mathbf{m}_{\tau_0+t}^T \mathbf{m})^2$ is p times the squared autocorrelation function of \mathbf{m} with translated argument $(\tau_0 + \cdot)$. This is a function which has a maximum at $t = -\tau_0$ and is symmetric around this point. The third term, $g_2(t) = \sum_{m_j \neq m} p_j (\mathbf{m}_{\tau_0+t}^T \mathbf{m}_j)^2$ is small compared to $g_1(t)$ when $|\tau_0 + t|$ is small. In fact, since in the orthogonal decomposition $\mathbf{m}_j = \alpha_j \mathbf{m} + \beta_j \tilde{\mathbf{m}}$ with $\tilde{\mathbf{m}} \perp \mathbf{m}$, $|\alpha_j| \ll 1$ must hold due to the assumed clear separation between \mathbf{m} and \mathbf{m}_j , we have $g_2(t) \approx \sum_{m_j \neq m} p_j (\alpha_j \mathbf{m}_{\tau_0+t}^T \mathbf{m})^2 \ll g_1(t)$ when $|\tau_0 + t|$ is small. From these considerations we conclude that $g(\mathbf{x}, t)$ is the sum of a constant and a term which is approximately proportional to the squared autocorrelation function $(\mathbf{m}_{\tau_0+t}^T \mathbf{m})^2$. It therefore has an approximately symmetric peak around $t = -\tau_0$. Since in the model (16) the jitter value is $\tau = \tau_0$, we have hereby heuristically motivated the estimator $\hat{\tau}_R(\mathbf{x}) = \arg \max_t \{g(\mathbf{x}, -t)\}$, where R denotes *Rayleigh*. Upon comparison between the function which this definition approximately maximizes, and the function which the matched filter (14) approximately maximizes, it is clear that they are close akin since the argument function of (14) is the square root of an approximation of $g(\mathbf{x}, t)$ minus a constant.

Using the analogy between the estimator $\hat{\tau}_R$ and the MAP estimator (11), we can also define an estimator corresponding to the MMSE estimator (10). This is the expectation of the normalized $g(\mathbf{x}, -t)$ regarded as a density function. Hence, after normalization of $g(\mathbf{x}, \cdot)$ according to

$$h(\mathbf{x}, n) \stackrel{\text{def}}{=} \frac{g(\mathbf{x}, n)}{\sum_k g(\mathbf{x}, k)}, \quad (27)$$

using discrete time, we have obtained two estimators, defined by

$$\hat{\tau}_{R1}(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{n=-(N-1)/2}^{(N-1)/2} n \cdot h(\mathbf{x}, -n) \quad (28)$$

and

$$\hat{\tau}_{R2}(\mathbf{x}) \stackrel{\text{def}}{=} \arg \max_n \{h(\mathbf{x}, -n)\}, \quad (29)$$

respectively. We will abbreviate these estimators R1 and R2.

4. A method which process each signal individually

A non-parametric method for jitter estimation is to compute the centre of gravity of the signal [17]. The centre of gravity jitter estimator is denoted $\hat{\tau}_{\text{cog}}$ and can be defined by

$$\hat{\tau}_{\text{cog}}(x) = \frac{\sum_n n|x(n)|}{\sum_n |x(n)|}. \quad (30)$$

A given signal $x(n)$ is compensated for jitter according to $\tilde{x}(n) = x(n + \hat{\tau}_{\text{cog}})$, which gives

$$\begin{aligned} \sum_n n|\tilde{x}(n)| &= \sum_n n|x(n + \hat{\tau}_{\text{cog}})| \\ &\approx \sum_n (n - \hat{\tau}_{\text{cog}})|x(n)| = 0. \end{aligned} \quad (31)$$

Using the assumption of the support of $m_{j,\tau}$ being contained in the interval $[-(N-1)/2, (N-1)/2]$, the approximation error in the second equality can be considered negligible (a small inexactitude is however present since $\hat{\tau}_{\text{cog}}$ is not an integer generally, i.e. the signal $x(t)$ is sampled at different time points before and after the second equality). Thus $\tilde{x}(n)$ has centre of gravity approximately in the origin. If a pattern signal m_j does not have centre of gravity in the origin, a bias is introduced between $\hat{\tau}_{\text{cog}}$ and the jitter τ which equals the centre of gravity of m_j . Since this bias is constant independently of τ , it does not disturb the possibility to identify clusters or averages within clusters, it only shifts all signals a constant value. However, since the summation interval $[-(N-1)/2, (N-1)/2]$ is finite and symmetric around the origin, a bias which varies with τ is introduced since the amount of noise before and after the signal $m_{j,\tau}$ are unequal after delay. Fig. 2 shows a pattern signal m_j and a delayed ($\tau=20$) noisy version x , with vertical marks for τ_{cog} , the centre of gravity of the noise-free delayed signal, and the estimate (30), respectively. The latter is biased towards the origin, and the bias increases with τ . Hereby optimal alignment is not obtained, which makes conditions for clustering worse and deteriorates the cluster

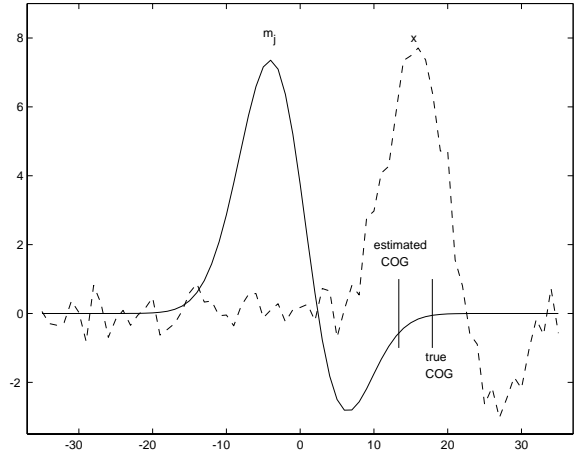


Fig. 2. Pattern signal m_j and a delayed ($\tau = 20$) noisy version x , with marks for its noise-free centre of gravity τ_{cog} and the estimate $\hat{\tau}_{\text{cog}}$ according to Eq. (30).

averages of aligned signals. The estimator (30) will be abbreviated COG henceforth.

Having defined the COG estimator, the question of its efficiency as compared to the estimators R1 and R2 can be posed. The following contention, which is derived in the appendix, claims the superiority of the estimator R1 as compared to a slightly modified version of COG, provided the SNR is high enough and a weak requirement on the pattern signal m is fulfilled.

Contention 1. *Provided (i) the pattern function m is symmetric or antisymmetric, (ii) the SNR is high enough, and (iii) the condition*

$$\sum_n \frac{(\mathbf{m}_{-n}^T \mathbf{m})^2}{\|\mathbf{m}\|^4} > L \quad (32)$$

is fulfilled, the estimator R1 has smaller variance than the estimator

$$\hat{\tau}_c = \frac{\sum_n n(x(n))^2}{\sum_n (x(n))^2}, \quad (33)$$

which is identical to COG after replacement of modulus with square.

5. Extension to vector-valued functions

The methods treated in Sections 3 and 4 can be extended to vector-valued functions, which is necessary

for the EEG application where data are measured at a multitude (often 20–30) channels. A simple extension is to align each channel individually. However, different jitter estimates for each channel are then obtained, and the signal delay pattern between the channels may be destroyed. This delay pattern may be of medical interest, so treating each channel individually is not a preferable method. Instead we will extend the methods to signal matrices where each channel (column) is translated equal values. By \mathbf{X} we denote an $(N \times C)$ —matrix whose columns consist of synchronously recorded signal vectors from C channels. The matrix \mathbf{X} is a sum of pattern matrix \mathbf{M} and noise matrix \mathbf{E} . The definition (22) can be rewritten

$$g(\mathbf{x}, t) = \mathbf{x}_t^T \mathbf{R} \mathbf{x}_t = \sum_{i=1}^N \lambda_i (\mathbf{q}_i^T \mathbf{x}_t)^2, \quad (34)$$

where the eigenvalue-eigenvector factorization $\mathbf{R} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T$ has been used. For matrix data an essentially identical formulation can be used,

$$g(\mathbf{X}, t) = \sum_{i=1}^{NC} \lambda_i \text{tr}^2(\mathbf{Q}_i^T \mathbf{X}_t), \quad (35)$$

using the matrix scalar product $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^T \mathbf{B})$. The norm corresponding to this scalar product is the Frobenius matrix norm, denoted $\| \cdot \|_F$. In (35) \mathbf{X}_t denotes a matrix the columns of which have been translated t time units. Here the matrices \mathbf{Q}_i , interpreted as $(NC \times 1)$ —vectors by piling the columns on top of each other according to $\mathbf{q}_i^{(v)} = [\mathbf{q}_{i,1}^T \ \cdots \ \mathbf{q}_{i,C}^T]^T$, (superscript (v) denotes *vector* and $\mathbf{q}_{i,j}$ denotes column j of \mathbf{Q}_i), are the eigenvectors of the correlation matrix of the data matrices interpreted as vectors, i.e. eigenvectors of

$$\mathbf{R}^{(v)} = \frac{1}{K} \sum_{k=1}^K \mathbf{x}_k^{(v)} \mathbf{x}_k^{(v)T}. \quad (36)$$

For the computation of the centre of gravity in the multi-channel case we use

$$\hat{t}_{\text{cog}} = \frac{\sum_{i=1}^C \sum_{n=1}^N n |x_{ni}|}{\sum_{i=1}^C \sum_{n=1}^N |x_{ni}|}, \quad (37)$$

where x_{ni} denotes the element at row n and column i of the matrix \mathbf{X} .

6. Numerical experiments

Comparison between the variances of COG, R1 and R2 for vector-valued functions:

We generated $K = 300$ matrices of $C = 15$ channels and $N = 29$ time samples from $L = 3$ classes. In order to obtain data that are realistic to the epileptic EEG spike application, the potential distribution over the channels was generated as potentials from a dipole inside a spherical conductor model of the head [20]. For each of the three classes the dipole parameters were fixed. The dipole location parameters were $r_1 = r_2 = r_3 = 0.5$, $\theta_1 = \theta_2 = \pi/3$, $\theta_3 = -\pi/3$, and $\phi_1 = \pi/4$, $\phi_2 = \phi_3 = -\pi/4$, for each class respectively, using a spherical coordinate system of radius r , azimuth θ and latitude ϕ . These dipole parameters correspond to well separated signal classes, with respect to potential distribution over the channels. For all classes the dipole moment was directed along the z axis and its amplitude was varying as a Gaussian function in the time interval $[-(N-1)/2, (N-1)/2]$.

The matrices were given Gaussian jitter of variance σ_τ^2 , whereupon coloured Gaussian noise of different variances was added, giving a range of SNR. The noise was chosen to be coloured in order to simulate realistic EEG noise which is coloured, and we generated it by filtering white noise through a Butterworth lowpass filter of degree ten and cut-off frequency $f = 0.125$. For matrix data, SNR is defined by

$$\text{SNR} \stackrel{\text{def}}{=} 10 \log_{10} \left(\frac{1}{K} \sum_{k=1}^K \frac{\|\mathbf{X}_k - \mathbf{E}_k\|_F^2}{\|\mathbf{E}_k\|_F^2} \right), \quad (38)$$

where $\| \cdot \|_F$ denotes the Frobenius matrix norm. Fig. 3 shows the results in terms of inverse class-averaged variance of the jitter estimation errors of methods COG, R1 and R2, the inverse class-averaged variance of the theoretically optimal Bayes MMSE estimator (10), and the inverse of σ_τ^2 . All quantities are displayed in dB. The variances of the jitter were chosen to be $\sigma_\tau^2 = 0.25$ in Fig. 3 (a), $\sigma_\tau^2 = 1$ in (b), $\sigma_\tau^2 = 5$ in (c), and $\sigma_\tau^2 = 15$ in (d).

All three methods gave small variability between classes. It can be seen that for several values of σ_τ^2

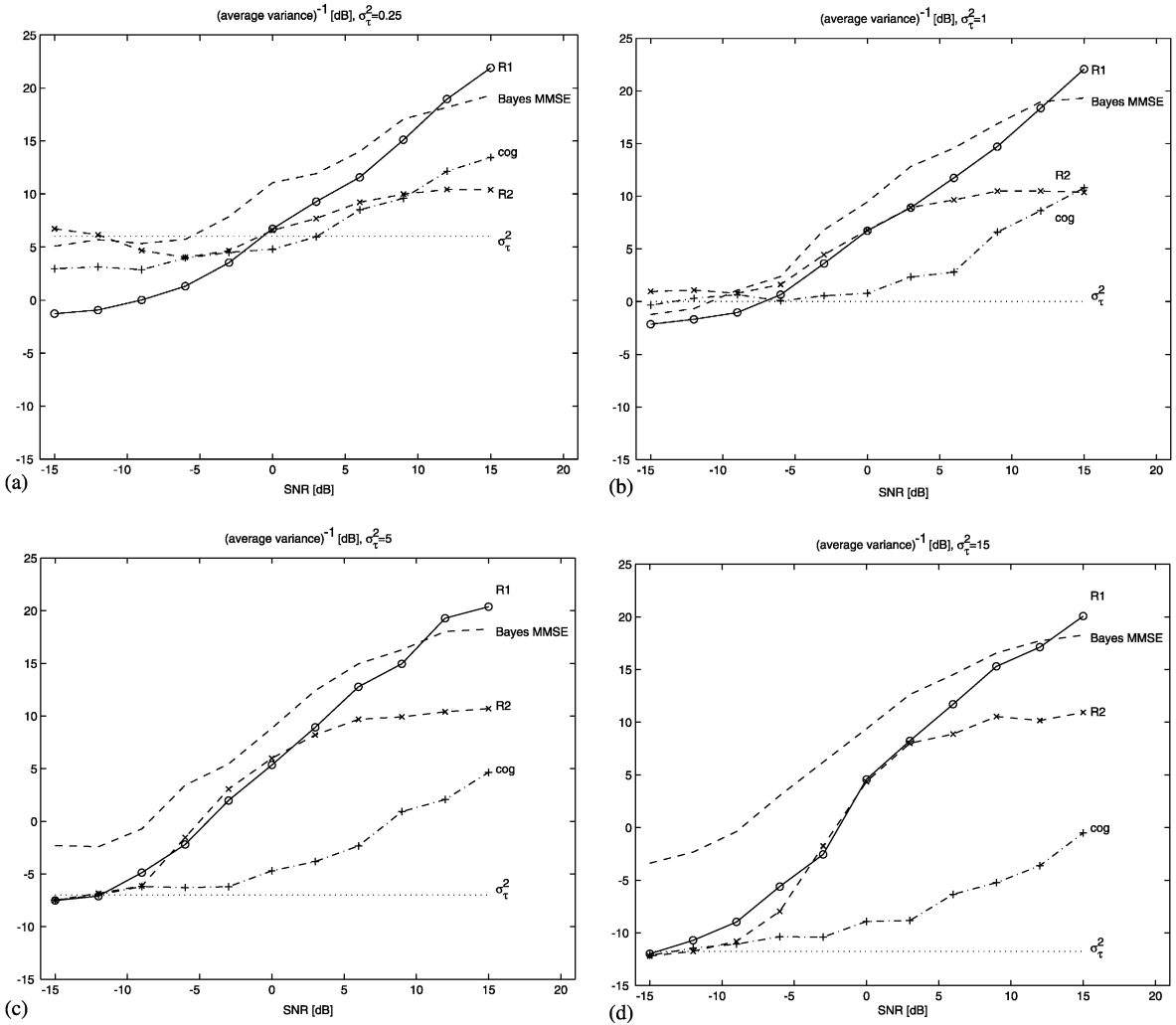


Fig. 3. Result of numerical experiment, variance results for $\sigma_\tau^2 = 0.25$ (a), $\sigma_\tau^2 = 1$ (b), $\sigma_\tau^2 = 5$ (c), and $\sigma_\tau^2 = 15$ (c). Key: COG (+), R1 (o), R2 (x), σ_τ^2 (dotted) and $\hat{\tau}_{\text{mmse}}$ (dashed).

and SNR the results are below the dotted line representing σ_τ^2 , which means that the method makes alignment *worse* than no processing. We observe that R1 seems to be the best method for large SNR and worse for small SNR, in particular for small jitter variances $\sigma_\tau^2 = 0.25$, $\sigma_\tau^2 = 1$, see Figs. 3(a) and (b). For the larger jitter variances $\sigma_\tau^2 = 5$, $\sigma_\tau^2 = 15$, R1 and R2 give about equal results, and both are better than COG. For the smaller jitter variances $\sigma_\tau^2 = 0.25$, $\sigma_\tau^2 = 1$ and small SNR, R1 is not a good method since it gives the worst deterioration of alignment for $\text{SNR} \leq 0$ dB, see

Fig. 3(a). R2 gives least deterioration of alignment for SNR values where deterioration of alignment occur. All estimators improve upon the Bayes MMSE estimator (10) for some values of SNR and σ_τ^2 . We believe this artifact is due to insufficient exactness in the computation of the Bayes MMSE error. Our conclusion from this experiment is that R2 seems to be the generally most useful method for vector-valued functions, since it never deteriorates alignment much, and for large σ_τ^2 it works not much worse than R1, which is the best method for large σ_τ^2 .

Table 1
Properties of real life data sets

Data set no.	F_s (Hz)	C	L	K	$\{K_j\}_{j=1}^L$
1	128	21	4	40	10, 10, 8, 12
2	128	21	3	29	8, 11, 10
3	128	21	3	50	7, 17, 4
4	128	26	3	50	15, 29, 6
5	250	26	2	50	31, 19
6	250	26	5	94	68, 14, 3, 6, 3
7	128	28	2	100	45, 55

7. Results on EEG data

We evaluated the methods COG, R1 and R2 on seven sets of clinically recorded epileptic EEG spikes. These data sets had been manually categorized by a neurophysiologist, and Table 1 gives properties of the data sets in terms of sampling frequency F_s , the number of channels C , the number of clusters L , the number of spikes K , and the number of spikes of each cluster K_j , $j = 1, \dots, L$. As can be seen the assumption of equal prior probability of clusters is not always fulfilled in practice, in particular data set no. 6 contains clusters of unequal size.

A signal time interval of 350 ms was cut out and used for COG estimation, and also for evaluation of the Rayleigh quotient with translations of each signal ± 200 ms. The signal matrices were (Frobenius-) normed, since clustering irrespective of signal amplitude was desired. We considered the clustering of the neurophysiologist as correct. We estimated that the SNR according to (38) for these data sets were approximately in the range $-6 \text{ dB} \leq \text{SNR} \leq 3 \text{ dB}$. In order to make comparisons between methods and between each method and raw data, we computed two criteria of clustering, μ_{ω_j} and s_{ω_j} , for each cluster ω_j and for each data set. The separation criterion μ_{ω_j} is defined as the squared Frobenius norm distance between the average of cluster ω_j and cluster ω_i , summed over all clusters ω_i , i.e.

$$\mu_{\omega_j} = \sum_{i \neq j} \|\bar{\mathbf{X}}_{\omega_j} - \bar{\mathbf{X}}_{\omega_i}\|_F^2, \quad (39)$$

where $\bar{\mathbf{X}}_{\omega_j}$ denotes the average of all \mathbf{X}_k such that $\mathbf{X}_k \in \omega_j$. This criterion is large if cluster ω_j is well

separated from the other clusters. The compactness criterion s_{ω_j} is defined to be the average squared Frobenius norm distance from the samples of cluster ω_j to their average,

$$s_{\omega_j} = \frac{1}{K_j} \sum_{\mathbf{X}_k \in \omega_j} \|\mathbf{X}_k - \bar{\mathbf{X}}_{\omega_j}\|_F^2. \quad (40)$$

The criterion s_{ω_j} is small if cluster ω_j is compact. It is clear that s_{ω_j} decreases and μ_{ω_j} increases or is left invariant if a set of signals is transformed from being poorly aligned to well aligned, since the within-cluster variability s_{ω_j} then becomes smaller while leaving average distances between clusters μ_{ω_j} approximately equal, or possibly augmenting them. Therefore a method is assessed from the sign and magnitude of the difference between criteria for processed data and raw data, i.e. Δs_{ω_j} and $\Delta \mu_{\omega_j}$ respectively. If Δs_{ω_j} is negative and $\Delta \mu_{\omega_j}$ positive or zero, improved alignment of cluster ω_j is indicated.

Table 2 shows the result of application of methods COG, R1 and R2. It can be seen that the methods COG and R1 does not give desirable results, in the sense that Δs_{ω_j} is positive for all clusters and data sets, i.e. clusters become less compact after alignment, and $\Delta \mu_{\omega_j}$ is negative for most clusters and data sets. On the other hand, method R2 gives improved alignment for data set nos. 3, 4 and 5, since Δs_{ω_j} is negative and $\Delta \mu_{\omega_j}$ is positive for all clusters of these data sets. For the other data sets R2 gives mixed results. For data set no. 1, Δs_{ω_j} is negative for two clusters and positive for the other two clusters, and $\Delta \mu_{\omega_j}$ is negative for all clusters. For data set no. 2, cluster ω_1 gets improved alignment (Δs_{ω_1} negative and $\Delta \mu_{\omega_1}$ positive), but for the other clusters no consistent improvement is obtained. For data set no. 6, consistent improvement

Table 2
Results for evaluation on real life data

Data set no.	Cluster	COG		R1		R2	
		Δs_{ω_j}	$\Delta \mu_{\omega_j}$	Δs_{ω_j}	$\Delta \mu_{\omega_j}$	Δs_{ω_j}	$\Delta \mu_{\omega_j}$
1	ω_1	0.16	-0.75	0.07	-0.78	-0.03	-0.29
	ω_2	0.10	-0.54	0.12	-0.82	-0.06	-0.21
	ω_3	0.19	-0.81	0.27	-1.38	0.10	-0.76
	ω_4	0.13	-0.91	0.26	-1.62	0.23	-0.97
2	ω_1	0.21	-0.50	0.31	-0.77	-0.02	0.27
	ω_2	0.10	-0.48	0.21	-0.80	0.32	-0.38
	ω_3	0.10	-0.48	0.23	-1.13	-0.03	-0.33
3	ω_1	0.10	-0.23	0.28	-0.58	-0.03	0.09
	ω_2	0.23	-0.29	0.42	-0.56	-0.01	0.03
	ω_3	0.08	-0.49	0.19	-1.17	-0.01	0.11
4	ω_1	0.03	-0.04	0.05	0.01	-0.04	0.05
	ω_2	0.08	0.04	0.12	-0.01	-0.03	0.14
	ω_3	0.03	0.06	0.02	0.09	-0.02	0.09
5	ω_1	0.15	-0.08	0.27	-0.12	-0.01	0.02
	ω_2	0.06	-0.08	0.07	-0.12	-0.03	0.02
6	ω_1	0.21	-1.00	0.35	-1.71	-0.01	-0.16
	ω_2	0.23	-0.93	0.14	-0.74	-0.06	0.05
	ω_3	0.34	-1.80	0.40	-2.18	0.54	-0.92
	ω_4	0.22	-1.34	0.37	-1.80	-0.024	-0.36
	ω_5	0.41	-1.38	0.51	-1.96	0.48	-0.99
7	ω_1	0.12	-0.13	0.23	-0.32	0.01	0.05
	ω_2	0.14	-0.13	0.19	-0.32	-0.01	0.05

is obtained for cluster ω_2 , otherwise mixed results. Clusters ω_3 , ω_4 and ω_5 are small (see Table 1) and their criteria values are of less interest than clusters ω_1 and ω_2 . For data set no. 7 cluster ω_2 is improved and the results of cluster ω_1 are mixed. In summary, 73% of clusters had improved compactness s_{ω_j} after alignment using method R2, and separations μ_{ω_j} were improved in 55% of the clusters. We conclude that R2 gave better results than R1 and COG, which in fact gave worse results than raw data. This result is consistent with the simulation result of Fig. 3(a) and (b), i.e. $\sigma_\tau^2 = 0.25$ and $\sigma_\tau^2 = 1$, for $\text{SNR} \leq -9$ dB. Method R2 improved alignment in the majority of clusters. However, alignment of several clusters were not consistently improved by any method.

In order to exemplify improvement of alignment visually we show in Figs. 4 and 5 two examples of signal clusters before and after alignment using method R2. Fig. 4 shows all waveforms of cluster ω_2 of data set no. 1 in the channel of largest energy before and after alignment. Fig. 5 shows all waveforms

of cluster ω_2 of data set no. 6 in the channel of largest energy before and after alignment. Clearly the alignment has improved after alignment using method R2 for these two clusters and channels.

The methods COG and R1 gave worse results than no processing for all seven data sets. In order to see whether the methods COG and R1 are applicable in the case of larger jitter values in data, we introduced *artificial* jitter of variance $\sigma_\tau^2 = 75$ (ms)² and applied each of the methods COG, R1 and R2. The results are given in Table 3, where it can be seen that COG and R1 give negative Δs_{ω_j} for many clusters and data sets, which means that these methods improve alignment if jitter variance is large enough. For COG, Δs_{ω_j} is negative in 68% of all clusters and $\Delta \mu_{\omega_j}$ is positive in 82% of all clusters. For R1, Δs_{ω_j} is negative in 36% of all clusters and $\Delta \mu_{\omega_j}$ is positive in 64% of all clusters. Nevertheless, the improvement in Δs_{ω_j} and $\Delta \mu_{\omega_j}$ is better for R2 than for COG and R1 for almost all clusters. We conclude that with large enough jitter, all three methods improve upon raw data. Method R2

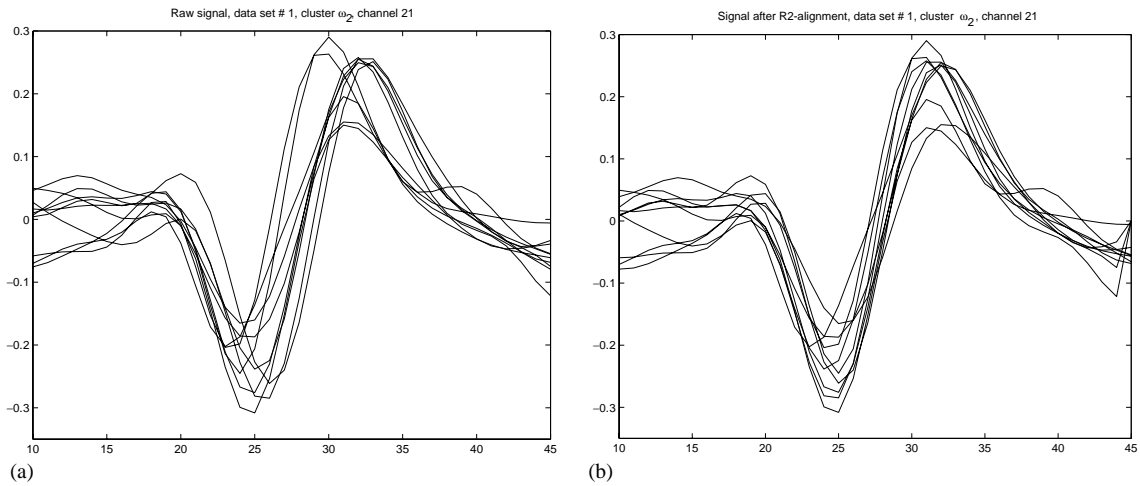


Fig. 4. Superposed signal waveforms of channel 21 of data set no. 1 and cluster ω_2 , before alignment (a), and after alignment with method R2 (b).

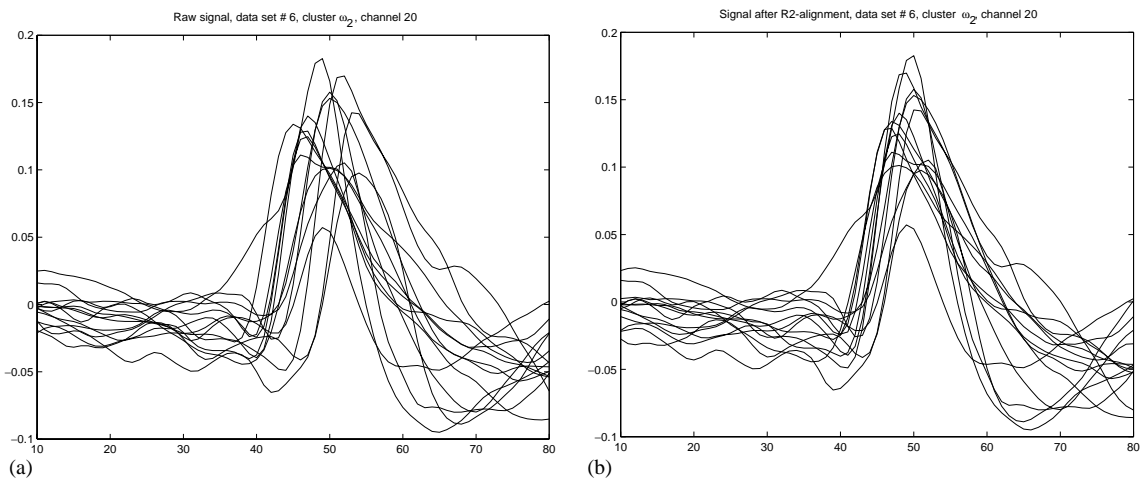


Fig. 5. Superposed signal waveforms of channel 20 of data set no. 6 and cluster ω_2 , before alignment (a), and after alignment using method R2 (b).

seems to be the method of choice also for large jitter variance.

Fig. 6 shows cluster ω_1 of data set no. 1 in the channel of largest energy for raw signal with artificial jitter, and COG, R1 and R2 aligned versions. It can be seen that all three methods improve alignment as compared to raw data, and also that R2 is better than R1, which in turn is better than COG.

8. Conclusions

We have treated methods for estimation of jitter in multi-class signal sets. Classical techniques like matched filtering are not applicable due to lack of knowledge of the signal waveforms, and approximate template signals can not be computed by averaging all signals. The center of gravity (COG) jitter estimate is a method which treats each signal individually. It

Table 3
Results for evaluation on real life data with artificial jitter

Data set no.	Cluster	COG		R1		R2	
		Δs_{ω_j}	$\Delta \mu_{\omega_j}$	Δs_{ω_j}	$\Delta \mu_{\omega_j}$	Δs_{ω_j}	$\Delta \mu_{\omega_j}$
1	ω_1	-0.29	1.26	-0.43	1.50	-0.50	2.81
	ω_2	-0.12	0.79	-0.15	0.79	-0.38	2.27
	ω_3	-0.09	1.39	-0.07	1.33	-0.48	2.99
	ω_4	0.02	0.87	0.12	0.56	0.11	1.43
2	ω_1	-0.05	0.50	0.02	0.33	-0.17	0.93
	ω_2	-0.01	0.17	0.05	-0.07	0.12	0.10
	ω_3	-0.00	0.50	0.13	0.02	0.00	0.47
3	ω_1	-0.02	0.11	0.10	-0.08	-0.17	0.26
	ω_2	0.02	0.11	0.12	0.03	-0.26	0.28
	ω_3	-0.04	0.21	0.05	-0.21	-0.04	0.63
4	ω_1	-0.15	0.21	-0.16	0.25	-0.27	0.31
	ω_2	-0.08	0.46	-0.07	0.43	-0.19	0.54
	ω_3	-0.05	0.28	-0.06	0.32	-0.09	0.31
5	ω_1	-0.05	0.12	0.03	0.11	-0.27	0.28
	ω_2	-0.13	0.12	-0.15	0.11	-0.25	0.28
6	ω_1	0.02	0.30	0.16	-0.44	-0.17	1.11
	ω_2	0.10	-0.18	0.04	-0.10	-0.14	0.17
	ω_3	0.24	-0.63	0.30	-1.03	0.15	-0.27
	ω_4	0.04	-0.18	0.16	-0.74	-0.18	0.11
	ω_5	0.08	-0.10	0.27	-0.97	0.35	-0.33
7	ω_1	-0.04	0.45	0.06	0.32	-0.11	0.68
	ω_2	-0.13	0.45	-0.12	0.32	-0.29	0.68

gives a bias which varies with the jitter value, leading to relatively large variance of the jitter estimation error.

The main features of this paper are the more elaborate methods that were developed using the statistics of the signal ensemble. After relating the problem to classical estimation paradigms and showing that these are not applicable under the present (weak) assumptions, we suggested two jitter estimators which use the Rayleigh quotient of the sample correlation matrix. For a given signal the Rayleigh quotient is evaluated while varying the translation, giving a function which has its largest values when the translated signal is close to signals which have the statistically most significant jitter values. By analogy with the estimators MMSE and MAP, we defined two estimators by mean value computation (R1) and maximization (R2) of the Rayleigh quotient function, respectively. Theoretically, we have found that R1 gives smaller variance

than COG (slightly modified) if the SNR is high enough.

The simulation results indicate that R2 is a generally well working method for vector-valued functions, which is of interest for application to multichannel EEG signals. For vector-valued functions R2 improves alignment in a range of SNR and jitter variance (σ_τ^2) values, and the improvement is larger than for COG. R1 gives in many cases even smaller estimation variance than R2, and thus improves alignment further, but method R1 has the flaw of giving large deterioration of alignment for small SNR and σ_τ^2 . Therefore R2 is the generally most reliable method.

When applying the methods to real life epileptic EEG signals, R2 gave the best results. It improved clustering in the majority of investigated clusters, albeit not all. The two other methods R1 and COG in fact deteriorated cluster compactness and separation as compared to no processing. A possible explanation for these phenomena may be that the SNR and the

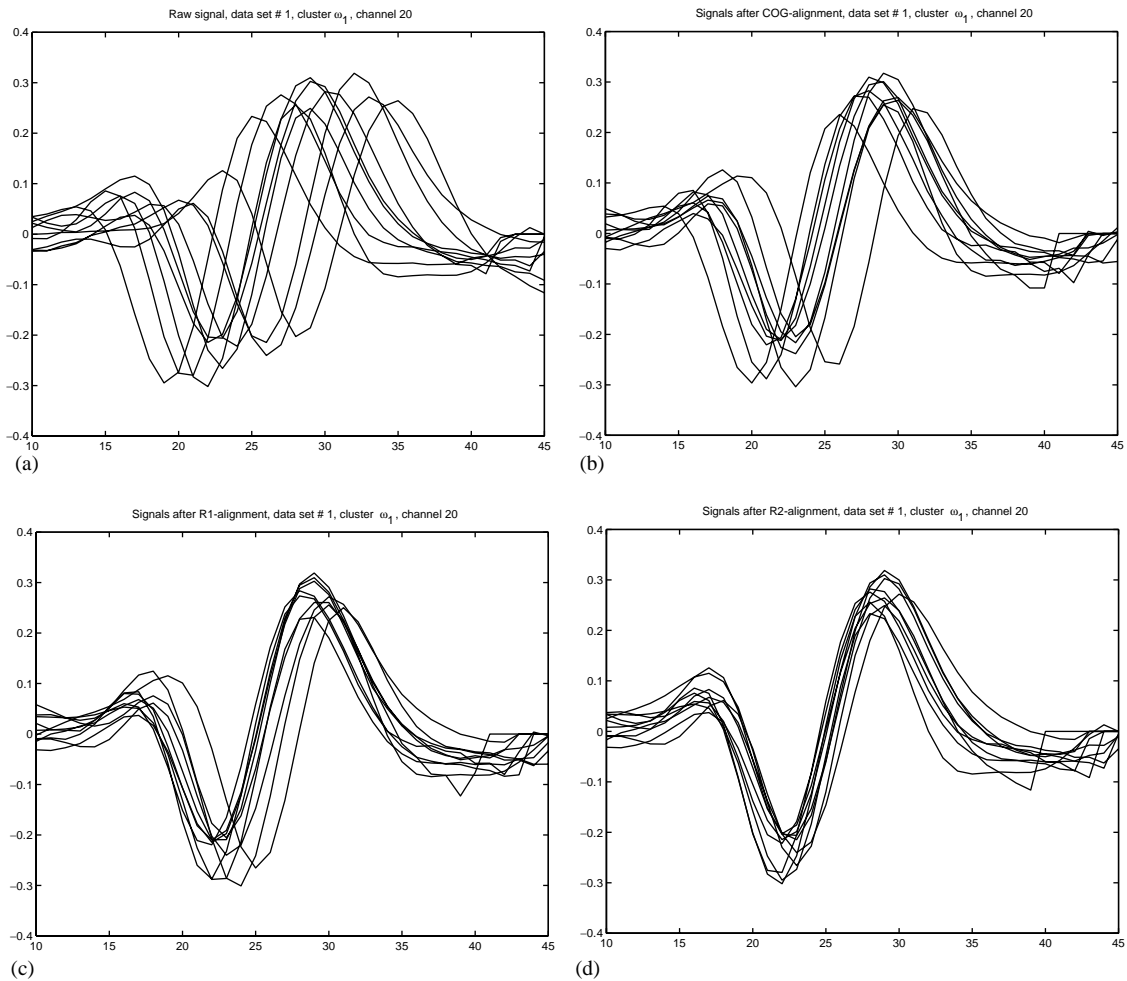


Fig. 6. Superposed signal waveforms of channel 20 of data set no. 1 and cluster ω_1 , before alignment (a), and after alignment using method COG (b), method R1 (c) and method R2 (d).

jitter variance σ_τ^2 are small for the real data sets, in which case the simulations predict that R1 and COG deteriorates alignment. For larger jitter variance than was actually present in the data, R1 and COG improves upon no processing, but apparently there is a risk in using them without prior knowledge of large jitter variance, since otherwise deterioration of clustering may be the result. The method R2 performed better than R1 and COG also in the case of large jitter variance, making it the preferable method for large as well as small jitter variance, as evaluated on these seven data sets.

Appendix A

A.1. Derivation of approximation (24)

The model (1) can be inserted into the definition of the correlation function for $x(t)$, giving

$$\begin{aligned} R(t_1, t_2) &\stackrel{\text{def}}{=} E\{x(t_1)x(t_2)\} \\ &= E\{(m_j(t_1 - \tau) + e(t_1)) \\ &\quad \times (m_j(t_2 - \tau) + e(t_2))\} \end{aligned}$$

$$= R_e(t_1, t_2) + \sum_{j=1}^L p_j E\{m_j(t_1 - \tau)m_j(t_2 - \tau)\}, \quad (\text{A.1})$$

where R_e denotes the correlation function of $e(t)$, and where we have used the assumptions of $e(t)$ having zero mean, and $e(t)$, the class ω_j and τ being independent. Further, assuming the existence of derivatives of $m_j(t)$, a first-order Taylor expansion gives $m_j(t - \tau) = m_j(t) - \tau m'_j(t) + o(\tau)$, which inserted in (A.1) gives

$$R(t_1, t_2) = R_e(t_1, t_2) + \sum_{j=1}^L p_j (m_j(t_1)m_j(t_2) + \sigma_\tau^2 m'_j(t_1)m'_j(t_2)) + \varepsilon, \quad (\text{A.2})$$

where the assumed property $E\{\tau\} = 0$ have been used, and ε is arbitrarily small if sufficient decrease of the norms of the higher derivatives $\|\frac{\partial^n}{\partial t^n} m_j\|$, $n = 2, 3, \dots$, is assumed. The sample correlation function $\hat{R}(t_1, t_2)$ is defined by

$$\hat{R}(t_1, t_2) = \frac{1}{K} \sum_{k=1}^K x_k(t_1)x_k(t_2), \quad (\text{A.3})$$

and it is an unbiased approximation of $R(t_1, t_2)$, the squared error of which decreases as a function of K due to the consistency of this estimator when the first four moments are assumed finite [21, p. 92]. If K is large enough the expected squared error is thus small. Discretizing time, (A.2) is approximated by the $(N \times N)$ -correlation matrix

$$\mathbf{R} \approx \mathbf{R}_e + \sum_{j=1}^L p_j (\mathbf{m}_j \mathbf{m}_j^T + \sigma_\tau^2 \mathbf{m}'_j \mathbf{m}'_j^T), \quad (\text{A.4})$$

where \mathbf{m}'_j denotes the discretized derivative of m_j .

A.2. Derivation of Contention 1

We use the signal model

$$x(n) = m(n - t) + e(n), \quad - (N - 1)/2 \leq n \leq (N - 1)/2, \quad (\text{A.5})$$

where $m = m_j$ for some j , e is zero mean Gaussian white noise of variance σ_e^2 , the jitter t has Gaussian density function $f_\tau(t)$, with variance σ_τ^2 . We assume m

is symmetric or antisymmetric around zero. According to the general assumptions of Section 2, the support of m_t is contained in the interval $[-(N - 1)/2, (N - 1)/2]$ for all jitter values t . The signal to noise ratio, denoted S , is defined by $S = \|\mathbf{m}\|^2 / (\sigma_e^2 N)$. In order to facilitate comparison with the estimator R1, we modify the estimator COG into

$$\hat{\tau}_c = \frac{\sum_n n(x(n))^2}{\sum_n (x(n))^2}, \quad (\text{A.6})$$

i.e. square is used instead of modulus. We abbreviate the estimator (A.6) COG₂. Its bias conditioned on $\tau = t$ is given by

$$\begin{aligned} E\{\hat{\tau}_c - \tau | \tau = t\} &= E\left\{ \frac{\sum_n (n - t)(x(n))^2}{\sum_n (x(n))^2} \right\} \\ &= \frac{\sum_n (n - t)E\{(x(n))^2\}}{\sum_n E\{(x(n))^2\}} + \mathcal{O}(S^{-1}) \\ &\approx \frac{\sum_n nm^2(n) - t\sigma_e^2 N}{\|\mathbf{m}\|^2 + \sigma_e^2 N} + \mathcal{O}(S^{-1}). \end{aligned} \quad (\text{A.7})$$

The second equality is motivated by

$$\begin{aligned} \sup_{\sigma_e \rightarrow 0} \frac{1}{\sigma_e} \left(E\left\{ \frac{\sum_n (n - t)(m(n - t) + e(n))^2}{\sum_n (m(n - t) + e(n))^2} \right\} - \frac{\sum_n (n - t)E\{(m(n - t) + e(n))^2\}}{\sum_n E\{(m(n - t) + e(n))^2\}} \right) \\ \approx \sup_{\sigma_e \rightarrow 0} \left(E\left\{ \frac{\sum_n (n - t)(2m(n - t)e(n) + e^2(n))}{\sigma_e \|\mathbf{m}\|^2 + \sum_n (2m(n - t)e(n) + e^2(n))} \right\} - \frac{1}{\sigma_e} \cdot \frac{(-t)N\sigma_e^2}{(\|\mathbf{m}\|^2 + N\sigma_e^2)} \right) < \infty, \end{aligned} \quad (\text{A.8})$$

using Schwartz' inequality and $\sum_n (n - t)m^2(n - t) \approx \sum_n nm^2(n) = 0$, which is due to the assumptions m_t being contained in the interval $[-(N - 1)/2, (N - 1)/2]$ and m (anti-)symmetric. This approximation is used also in the last equality of (A.7), whereby $E\{\hat{\tau}_c - \tau | \tau = t\} \approx -t \cdot 1 / (1 + S)$ if S is large. Using the

formula $E\{\hat{\tau}_c - \tau\} = \int f_\tau(t) \cdot E\{\hat{\tau}_c - \tau | \tau = t\} dt$ now gives $E\{\hat{\tau}_c - \tau\} \approx 0$, because f_τ is even, so COG₂ is approximately unbiased. A similar computation yields for the estimator R1

$$\begin{aligned} E\{\hat{\tau}_{R1} - \tau | \tau = t\} &= \frac{\sum_n (n-t) E\{g(\mathbf{x}, -n)\}}{\sum_n E\{g(\mathbf{x}, -n)\}} + \mathcal{O}(S^{-1}) \\ &= \frac{\sum_n (n-t) \mathbf{m}_{t-n}^T \mathbf{R} \mathbf{m}_{t-n} - t N \sigma_e^2 \text{tr} \mathbf{R}}{\sum_n \mathbf{m}_{t-n}^T \mathbf{R} \mathbf{m}_{t-n} + N \sigma_e^2 \text{tr} \mathbf{R}} + \mathcal{O}(S^{-1}). \end{aligned} \quad (\text{A.9})$$

For the first term of the right hand side nominator we have

$$\sum_n (n-t) \mathbf{m}_{t-n}^T \mathbf{R} \mathbf{m}_{t-n} \approx \sum_n n \mathbf{m}_{-n}^T \mathbf{R} \mathbf{m}_{-n} = 0, \quad (\text{A.10})$$

since $\mathbf{m}_{-n}^T \mathbf{R} \mathbf{m}_{-n}$ is an even function of n . Hereby

$$E\{\hat{\tau}_{R1} - \tau | \tau = t\} \approx -t \frac{1}{1 + S \frac{\sum_n \mathbf{m}_{t-n}^T \mathbf{R} \mathbf{m}_{t-n}}{\|\mathbf{m}\|^2 \text{tr} \mathbf{R}}}, \quad (\text{A.11})$$

and further

$$E\{\hat{\tau}_{R1} - \tau\} = \int f_\tau(t) E\{\hat{\tau}_{R1} - \tau | \tau = t\} dt \approx 0, \quad (\text{A.12})$$

since $\sum_n \mathbf{m}_{t-n}^T \mathbf{R} \mathbf{m}_{t-n}$ is an even function of t . Hence R1 is approximately unbiased. We now turn to the variance of the estimators. The approximation

$$E\{(\hat{\tau}_c - \tau)^2 | \tau = t\} \approx t^2 \frac{1}{(1+S)^2} \quad (\text{A.13})$$

holds with improved accuracy with increasing S , and yields

$$E\{(\hat{\tau}_c - \tau)^2\} \approx \frac{\sigma_\tau^2}{(1+S)^2} \quad (\text{A.14})$$

and, likewise,

$$E\{(\hat{\tau}_{R1} - \tau)^2\} \approx \frac{\sigma_\tau^2}{\left(1 + S \frac{\sum_n \mathbf{m}_{t-n}^T \mathbf{R} \mathbf{m}_{t-n}}{\|\mathbf{m}\|^2 \text{tr} \mathbf{R}}\right)^2}, \quad (\text{A.15})$$

where the approximation $\sum_n \mathbf{m}_{t-n}^T \mathbf{R} \mathbf{m}_{t-n} \approx \sum_n \mathbf{m}_{-n}^T \mathbf{R} \mathbf{m}_{-n}$ have been used. In order to compare

variances of COG₂ and R1, we focus on the quantity

$$\frac{\sum_n \mathbf{m}_{-n}^T \mathbf{R} \mathbf{m}_{-n}}{\|\mathbf{m}\|^2 \text{tr} \mathbf{R}}. \quad (\text{A.16})$$

With aid of the approximation

$$\mathbf{R} \approx \sigma_e^2 I + \frac{1}{L} \sum_{j=1}^L \mathbf{m}_j \mathbf{m}_j^T, \quad (\text{A.17})$$

obtained from (A.4) with further approximations, $\frac{1}{L} \sum_{j=1}^L \|\mathbf{m}_j\|^2 \approx \|\mathbf{m}\|^2$, $\mathbf{m}^T \mathbf{m}_j \approx 0$ when $\mathbf{m} \neq \mathbf{m}_j$, and $\sum_n \|\mathbf{m}_{-n}\|^2 = \alpha N \|\mathbf{m}\|^2$, where $\alpha < 1$ since with $|n|$ large enough, the support of \mathbf{m}_{-n} is no longer contained in the interval $[-(N-1)/2, (N-1)/2]$ ($\alpha \approx 1$ can be assumed), we obtain

$$\begin{aligned} \frac{\sum_n \mathbf{m}_{-n}^T \mathbf{R} \mathbf{m}_{-n}}{\|\mathbf{m}\|^2 \text{tr} \mathbf{R}} &= \frac{\sigma_e^2 \sum_n \|\mathbf{m}_{-n}\|^2 + \frac{1}{L} \sum_n (\mathbf{m}_{-n}^T \mathbf{m})^2}{\|\mathbf{m}\|^2 \sigma_e^2 N + \|\mathbf{m}\|^2 \frac{1}{L} \sum_{j=1}^L \|\mathbf{m}_j\|^2} \\ &= \alpha \frac{1 + S \frac{\sum_n (\mathbf{m}_{-n}^T \mathbf{m})^2}{\alpha L \|\mathbf{m}\|^4}}{1 + S \frac{1/L \sum_{j=1}^L \|\mathbf{m}_j\|^2}{\|\mathbf{m}\|^2}} \\ &\approx \alpha \frac{1 + S \frac{1}{\alpha L} \sum_n \frac{(\mathbf{m}_{-n}^T \mathbf{m})^2}{\|\mathbf{m}\|^4}}{1 + S}. \end{aligned} \quad (\text{A.18})$$

If we now assume $S \gg 1$, we obtain

$$\frac{\sum_n \mathbf{m}_{-n}^T \mathbf{R} \mathbf{m}_{-n}}{\|\mathbf{m}\|^2 \text{tr} \mathbf{R}} \approx \frac{1}{L} \sum_n \frac{(\mathbf{m}_{-n}^T \mathbf{m})^2}{\|\mathbf{m}\|^4} > 1, \quad (\text{A.19})$$

where the inequality holds provided

$$\sum_n \frac{(\mathbf{m}_{-n}^T \mathbf{m})^2}{\|\mathbf{m}\|^4} > L. \quad (\text{A.20})$$

The condition (A.20) is quite a weak assumption since the number of clusters L typically is small (≤ 10) and $(\mathbf{m}_{-n}^T \mathbf{m})^2 \approx \|\mathbf{m}\|^4$ when $|n|$ is small for a number of n which is larger than L . This is the case if the function $m(t)$ is varying slowly enough. Hereby (A.20) implies

$$\frac{\sum_n \mathbf{m}_{-n}^T \mathbf{R} \mathbf{m}_{-n}}{\|\mathbf{m}\|^2 \text{tr} \mathbf{R}} > 1, \quad (\text{A.21})$$

which inserted into (A.15), upon comparison with (A.14), gives

$$E\{(\hat{\tau}_{R1} - \tau)^2\} < E\{(\hat{\tau}_c - \tau)^2\}, \quad (\text{A.22})$$

i.e. the estimator R1 has smaller variance than COG₂ provided S is large enough. Hereby the contention has been derived. \square

References

- [1] R.A. Christensen, A.D. Hirschman, Automatic phase alignment for the Karhunen-Loève expansion, *IEEE Trans. Biomed. Eng.* 26 (2) (1979) 94–99.
- [2] A.A. Dingle, R.D. Jones, G.J. Carroll, W.R. Fright, A multistage system to detect epileptiform activity in the EEG, *IEEE Trans. Biomed. Eng.* 40 (12) (1993) 1260–1268.
- [3] L.E. Franks, *Signal Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
- [4] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, New York, 1990.
- [5] L. Gupta, D.L. Molfese, R. Tammana, P.G. Simos, Nonlinear alignment and averaging for estimating the evoked potential, *IEEE Trans. Biomed. Eng.* 43 (4) (1996) 348–356.
- [6] C.W. Helstrom, *Elements of Signal Detection and Estimation*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
- [7] A.K. Jain, R.C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [8] R. Jané, H. Rix, P. Caminal, P. Laguna, Alignment methods for averaging of high-resolution cardiac signals: a comparative study of performance, *IEEE Trans. Biomed. Eng.* 38 (6) (1991) 571–579.
- [9] B.H. Jansen, H.C. Huang, Automated morphological analysis by means of dynamic time-warping, *Electroencephalography Clin. Neurophysiol.* 60 (1985) 282–284.
- [10] S. Jesus, H. Rix, High resolution ECG analysis by an improved signal averaging method and comparison with beat-to-beat approach, *J. Biomed. Eng.* 10 (1988) 25–32.
- [11] S. Jesus, H. Rix, A. Varenne, Signal averaging using shape classification: application to high resolution ECG, in: *Proc. EUSIPCO -86*, The Hague, Netherlands, September 1986, pp. 1371–74.
- [12] X. Kong, N.V. Thakor, Adaptive estimation of latency changes in evoked potentials, *IEEE Trans. Biomed. Eng.* 43 (2) (1996) 189–197.
- [13] P. Laguna, H. Rix, P. Caminal, R. Jané, Performance Analysis of a Time Delay Estimate Between Two Noisy Transient Signals, in: *Proceedings of the 12th International Conference on IEEE-EMBS*, IEEE, New York, 1990, pp. 877–78.
- [14] C.D. McGillem, J.J. Aunon, C.A. Pomazala, Improved waveform estimation procedures for event-related potentials, *IEEE Trans. Biomed. Eng.* 32 (6) (1985) 371–379.
- [15] O. Meste, H. Rix, Jitter statistics estimation in alignment processes, *Signal Process.* 51 (1) (1996) 41–53.
- [16] J. Mochs, W. Kohler, T. Gasser, D.T. Pham, Novel approaches to the problem of latency jitter, *Psychophysiol.* 25 (2) (1988) 217–226.
- [17] A. Papoulis, *Signal Analysis*, McGraw-Hill, New York, 1977.
- [18] O. Rempelman, H.H. Ros, Coherent averaging technique: a tutorial review. Part 2: trigger jitter, overlapping responses and non-periodic stimulation, *J. Biomed. Eng.* 8 (1986) 30–35.
- [19] L.L. Scharf, *Statistical Signal Processing*, Addison-Wesley, Reading, MA, 1991.
- [20] M. Scherg, *Fundamentals of Dipole Source Potential Analysis*, in: F. Grandori, M. Hoke, G.L. Romani (Eds.), *Auditory Evoked Magnetic Fields*, Adv. Audiol. Basel, Karger, 1990, pp. 40–69.
- [21] H.W. Sorensen, *Parameter Estimation*, Marcel Dekker Inc, New York, 1980.
- [22] L. Sörnmo, Vectorcardiographic loop alignment and morphologic beat-to-beat variability, *IEEE Trans. Biomed. Eng.* 45 (12) (1998) 1401–1413.
- [23] H.L. van Trees, *Detection, Estimation, and Modulation Theory*, Vol. I, Wiley, New York, 1968.
- [24] P. Wahlberg, G. Salomonsson, Feature extraction and clustering of EEG epileptic spikes, *Comput. Biomed. Res.* 29 (October 1996) 382–394.
- [25] C.D. Woody, Characterization of an adaptive filter for the analyses of variable latency neuroelectric signals, *Med. Biol. Eng.* 5 (1967) 539–553.
- [26] G. Zouridakis, B.H. Jansen, N.N. Boutros, A fuzzy clustering approach to EP estimation, *IEEE Trans. Biomed. Eng.* 44 (8) (1997) 673–680.