Data Min Knowl Disc (2014) 28:1530–1553
DOI 10.1007/s10618-014-0371-0

Confidence bands for time series data

Jussi Korpela · Kai Puolamäki · Aristides Gionis

Received: 16 February 2014 / Accepted: 21 June 2014 / Published online: 27 July 2014

Abstract Simultaneous confidence intervals, or *confidence bands*, provide an intuitive description of the variability of a time series. Given a set of N time series of length M, we consider the problem of finding a confidence band that contains a $(1-\alpha)$-fraction of the observations. We construct such confidence bands by finding the set of $N-K$ time series whose envelope is minimized. We refer to this problem as the *minimum width envelope* problem. We show that the minimum width envelope problem is **NP**-hard, and we develop a greedy heuristic algorithm, which we compare to quantile- and distance-based confidence band methods. We also describe a method to find an effective confidence level α_{eff} and an effective number of observations to remove K_{eff}, such that the resulting confidence bands will keep the family-wise error rate below α. We evaluate our methods on synthetic and real datasets. We demonstrate that our method can be used to construct confidence bands with guaranteed family-wise error rate control, also when there is too little data for the quantile-based methods to work.

Keywords Simultaneous confidence interval · Confidence band · Time series · Multiplicity correction · Family-wise error rate

Responsible editors: Toon Calders, Floriana Esposito, Eyke Hüllermeier, Rosa Meo.

Electronic supplementary material The online version of this article (doi:10.1007/s10618-014-0371-0) contains supplementary material, which is available to authorized users.

J. Korpela (✉) · K. Puolamäki
Finnish Institute of Occupational Health, Topeliuksenkatu 41 a A, 00250 Helsinki, Finland
e-mail: jussi.korpela@ttl.fi

A. Gionis
Department of Information and Computer Science, Helsinki Institute for Information Technology (HIIT), Aalto University, PO Box 15400, 00076 Aalto, Helsinki, Finland

1 Introduction

Confidence intervals are typically used to describe a univariate distribution. However, the concept of confidence intervals can be extended and be used to describe also multivariate time series data. We focus on the traditional two dimensional value-versus-time representation of time series and speak of *confidence bands* that define a region of most probable observations. A time series is *extreme* if it at some point falls outside of the confidence band.

Confidence bands, such as the ones shown in Fig. 1, are useful for many reasons. They are easy to interpret both because of the direct analogy to univariate confidence intervals and because they are presented in the same format as the data. The visual representation is particularly effective as we can easily spot trends, patterns, and outliers when the data are presented as an image. In addition, automatic detection of outliers is readily implemented using simple thresholding.

However, as the length of the time series grows, false alarms become a problem. For example, consider a time series of length 10 where the time points are completely uncorrelated to each other. One naïve option is to form a 90 % confidence band, which consists of ten independently-computed scalar 90 % confidence intervals. The time series will lie completely within this band only $0.90^{10} = 35$ % of the time. In the remaining 65 % at least one of the time points lies outside the naïve confidence band. In a real world application these might trigger an alarm 65 % of the time. Hence, we should be able to control the number of false positives our method produces, i.e., to perform multiplicity correction. The standard way to perform the multiplicity correction is to define an error rate (e.g., the fraction of false positives f_p), a threshold (call it α), and devise a method that keeps the error rate at or below the given threshold. In the statistical literature there is a plethora of methods for multiplicity correction, but most of them are either specific to a certain statistical test or they consider only the p-values from a set of possibly correlated tests. Multiplicity correction is rarely applied to the construction of the confidence bands. In Sect. 6 we provide further discussion regarding the related work.

In this work we describe a method to compute multivariate confidence bands for time series data while controlling the *family-wise error rate* (FWER). The FWER is the probability of making one or more false discoveries. In particular, when saying "controlling the FWER at α," we mean that the probability of falsely marking a time series as extreme is at most α. The proposed methods are data-driven and are well suited for the analysis of multivariate autocorrelated data (time series). No parametric assumptions about the data are made and the only required input is the observed time series data.

We formalize the problem of finding a confidence band as a problem of finding an envelope of minimum width. Given a set of N time series the *minimum width envelope* (MWE) problem is to find a subset of $N - K$ time series such that their envelope has the smallest total width. The problem turns out to be **NP**-hard and, in addition, we can show that the complement objective function is hard to approximate. However, we provide an efficient greedy algorithm for finding an approximate solution. We also describe a method that can be used to guarantee FWER control.

A motivating example is shown in Fig. 1. It shows a set of real normal heart beats along with three possible 90 % confidence bands. The narrowest band (solid red) is

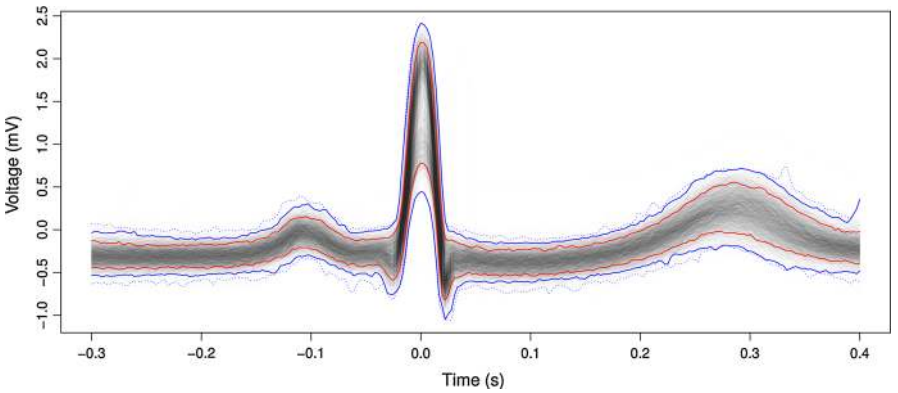


Fig. 1 A set of $N = 1{,}507$ normal heart beats along with different 90 % confidence bands. Shown are the simple quantile (*solid red*), minimum width envelope with $K = \lfloor 0.1 \times 1507 \rfloor = 150$ observations removed (*solid blue*) and minimum width envelope with guaranteed FWER control ($K_{\text{eff}} = 27$, *dotted blue*) confidence bands. Details about the data can be found in Sect. 5.1

computed using simple quantile at each time point without any multiplicity correction. Although this seems crude in light of the discussion above, this method is often used in practice. The confidence band produced by our MWE method (solid blue) is considerably wider. The widest band is computed using MWE with FWER control (dotted blue) and it is guaranteed to keep FWER under 10 %. Hence Fig. 1 clearly demonstrates that the width of the simple quantile confidence band is very far from a width that would control FWER.

From a set of confidence bands that control the FWER at some desired level α, a narrow band has the highest power of detecting an extreme observation. Hence, it is justified to use the confidence band width as a cost function in the two dimensional representation shown above.

In addition to the MWE method, we introduce for comparison novel methods based on ideas from multivariate outlier detection literature.
To summarize, the contributions of this work are the following:

1. Definition and characterization of the MWE problem for time series data,
2. Efficient algorithm to solve the MWE problem,
3. A procedure to guarantee FWER control,
4. Comparison to other applicable methods, and
5. Publicly available code (algorithms and experiments).[1]

In the following we first define the MWE problem. We prove that the problem is **NP**-hard and that the complement objective function is hard to approximate. To confront with the problem we devise a greedy heuristic algorithm. We also provide definitions of quantile and distance-based confidence bands, followed by a general procedure that can be used to guarantee that the FWER remains controlled. Finally, we use synthetic and real datasets to compare the methods and discuss the related literature.

[1] https://bitbucket.org/jtkorpel/mwe_2014.

2 Problem definition

2.1 The minimum width envelope (MWE) problem

We consider N time series, each one having M time points. We organize this data in an $N \times M$ matrix $\mathbf{X}$. We write $I \subseteq \{1, \ldots, N\}$ to denote a subset of the N time series, or equivalently, a subset of rows of $\mathbf{X}$. We define the *upper envelope* of $\mathbf{X}[I, \cdot]$ to be $E_{\text{up}}(m, I) = \max_{n \in I} \mathbf{X}[n, m]$ and the *lower envelope* $E_{\text{low}}(m, I) = \min_{n \in I} \mathbf{X}[n, m]$. We define $U(I) = \sum_{m=1}^{M} \left(E_{\text{up}}(m, I) - E_{\text{low}}(m, I)\right)$ to be the size of the envelope of the sub-matrix $\mathbf{X}[I, \cdot]$. The *confidence band* is defined as the minimal area that bounds the envelope.

The task is to remove K observations from $\mathbf{X}$ such that the envelope of the remaining dataset is minimized. More formally, we have:

Problem 1 *Minimum width envelope (*MWE*) problem.* Given a dataset $\mathbf{X}$, representing N time series of dimension M, and given an integer K, find a subset of time series $I_{\text{opt}} \subseteq \{1, \ldots, N\}$ of size $|I| = N - K$, such that the size of the envelope $U(I_{\text{opt}})$ is minimized.

The solution to the MWE problem can be used to construct confidence bands. Suppose we have obtained a set of N time series with M samples from an unknown distribution F. An approximate empirical $1 - \alpha$ confidence band for F is constructed by setting $K = \lfloor \alpha N \rfloor$.

2.2 Complexity of the MWE problem

In this section we show that the MWE problem is **NP**-hard and that the complement objective function is hard to approximate. Our proof uses a reduction from the Maximum k-Subset Intersection (MSI) problem to a special case of the MWE problem.

The MSI problem is be defined as follows: given a collection $\mathcal{C} = \{S_1, \ldots, S_m\}$ of m subsets over a finite set of elements $\mathcal{E} = \{e_1, \ldots, e_n\}$, and a positive integer k, the objective is to select exactly k subsets $J \subseteq \{1, \ldots, m\}$, $|J| = k$, whose intersection size $\left|\cap_{j \in J} S_j\right|$ is maximum.

Theorem 1 (Xavier 2012) *The* MSI *problem is* **NP**-*hard.*

Theorem 2 (Xavier 2012) *Let $\epsilon > 0$ be an arbitrary small constant. Assume that SAT does not have a probabilistic algorithm that decides whether a given instance of size n is satisfiable in time $2^{n^{\epsilon}}$. Then there is no polynomial time algorithm for the* MSI *problem that achieves an approximation ratio of $1/N^{\epsilon'}$ where N is the size of the instance, and ϵ' depends only on ϵ.*

Theorem 3 *The* MWE *problem is* **NP**-*hard, and the complement objective function is hard to approximate.*

Proof We reduce the MSI problem to a special case of the MWE problem by constructing a 0–1 time series dataset $\mathbf{X} \in \{0, 1\}^{N \times M}$, where $N = 2m + 1$ and $M = n$, as follows,

$$\mathbf{X}[i, j] = \begin{cases} 1, & i \leq m \text{ and } e_j \notin S_i, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We finally set $K = m - k$, hence defining an instance of the MWE problem. Because $K \leq m$ and because at each time instance $j \in \{1, \ldots, M\}$ there are at least $m + 1$ zeroes in $X[\cdot, j]$, the lower envelope of any subset of $N - K$ rows is always zero. Therefore, the value of the solution of the MWE problem is given only by the upper envelope as

$$U(I) = \sum_{j=1}^{M} \max_{i \in I} \mathbf{X}[i, j]. \tag{2}$$

Furthermore, the time series $i \in \{m + 1, \ldots, 2m + 1\}$ always belong to I since, by definition, they are composed of only zeros and hence do not add to the envelope width. For a given j, the maximum in Eq. (2) is equal to 1 if and only if there exists some $i \in I$ such that $e_j \notin S_i$. Equivalently, the maximum term is zero only if such term does not exists, i.e., all of S_i with $i \in I$ contain e_j or $|\cap_{i \in I} \left(e_j \cap S_i\right)| = 1$. The cost function can therefore equivalently be written in terms of sets of the MSI problem as

$$U(I) = \sum_{j=1}^{M} \left(1 - \left|\cap_{i \in I \cap \{1,\ldots,m\}} \left(e_j \cap S_i\right)\right|\right) = M - \left|\cap_{i \in I \cap \{1,\ldots,m\}} S_i\right|, \tag{3}$$

where the latter term is M—the size of the original envelope of the time series data—minus the MSI cost function. Minimizing the MWE cost function is therefore equivalent to maximizing the MSI cost function with the MSI solution set given by $J = I \cap \{1, \ldots, m\}$, i.e., $U(I) = M - |\cap_{i \in J} S_i|$. It follows from Theorem 1 that the MWE problem is **NP**-hard and from Theorem 2 that the complement of the MWE cost function is hard to approximate. □

Because inapproximability results do not necessarily apply to the complement of cost functions, the inapproximability claim in Theorem 3 does not directly apply to the MWE problem, when the cost function is defined as a minimization of the envelope. However, the inapproximability results holds for the equivalent complement problem, i.e., maximizing the area outside the envelope.

3 Algorithms

3.1 Greedy minimum width envelope algorithm

We now describe a greedy algorithm for solving the MWE problem. The idea is to sequentially select K observations and remove them from the envelope. At each iteration, the observation to select for removal is the one whose removal yields the largest reduction ΔU in envelope size.

Algorithm 1 Greedy MWE algorithm

1: **input**: (1) dataset $\mathbf{X}$ of size $N \times M$ with N observations of length M, (2) number of observations to remove K
2: **output**: the set of central observations I
3: R $\leftarrow$ ordering structure N observations in $\mathbf{X}$ ▷ See Appendix 8.1
4: $I \leftarrow \{1, \ldots, N\}$ ▷ The central observations
5: **for** $k = 1 \ldots K$ **do**
6: Create hash table ΔU such that a query with a previously unused key returns a value of zero
7: $A \leftarrow \emptyset$
8: **for** $j = 1 \ldots M$ **do**
9: $\Delta U[small(R, j)] \leftarrow \Delta U[\text{small}(R, j)] + |\mathbf{X}[\text{small}(R, j), j] - \mathbf{X}[\text{2ndsmall}(R, j), j]|$
10: $\Delta U[\text{big}(R, j)] \leftarrow \Delta U[\text{big}(R, j)] + |\mathbf{X}[\text{big}(R, j), j] - \mathbf{X}[\text{2ndbig}(R, j), j]|$
11: $A \leftarrow A \cup \{\text{small}(R, j)\} \cup \{\text{big}(R, j)\}$
12: **end for**
13: $i_{\text{opt}} \leftarrow \arg\max_{i \in A} \Delta U[i]$
14: $R \leftarrow \text{remove}(R, i_{\text{opt}})$
15: $I \leftarrow I \backslash \{i_{\text{opt}}\}$
16: **end for**
17: **return** I

The details of the approach are shown in Algorithm 1. First, each column is sorted according to the values of its entries; this is shown in line 3. Information about the ordering of the columns is maintained in a data structure R, which consists of a doubly-linked list and an index structure. The data structure R and the respective methods are described in the Appendix. In brief, the functions $\text{small}(R, j)$ and $\text{2ndsmall}(R, j)$ return the indices of the smallest and second smallest non-removed observations within column j of $\mathbf{X}$. The functions $\text{big}(R, j)$ and $\text{2ndbig}(R, j)$ return the indices of the largest and 2nd largest values, respectively. The function $\text{remove}(R, i)$ removes the i-th time series and updates the data structure R.

The main part of the greedy algorithm, shown on lines 5–16, is an iteration over $k = 1 \ldots K$ to select the time series to remove. During the k-th iteration, the task is to find an observation i_{opt} that, when removed, reduces the size of the envelope of the remaining time series most. By the construction of the data structure R, only the time series that have not yet been removed are considered for removal. By extreme observation we mean an observation $\mathbf{X}[i, \cdot]$ that contains the largest/smallest value of some column m and has not yet been removed. To find out the reductions in envelope width (ΔU) corresponding to the extreme observations, it suffices to iterate over all M dimensions. Since the data structure R contains information of the ordering of each column, each iteration requires constant time. On line 13 the optimal time series to remove is selected and on line 14 the data structure R is updated.

To establish an approximate $1 - \alpha$ confidence band for the dataset $\mathbf{X}$, the number of observations to remove is set to $K = \lfloor \alpha N \rfloor$. As the dataset size N increases the true coverage of the band converges towards the target $1 - \alpha$.

3.2 Properties of the greedy algorithm

As we will see in our experimental evaluation, the greedy algorithm performs well in practice. However, it can be shown that it does not provide any approximation guarantee. An adversarial example is given by a setup with $N = 5$, $M = 1$, $K = 2$,

and a data matrix given by $\mathbf{x}^T = (1, 1-\epsilon/2, 2\epsilon, \epsilon, 0)$, with ϵ arbitrarily small. The optimal solution is given by $I_{\text{opt}} = \{3, 4, 5\}$ with cost $U(I_{\text{opt}}) = 2\epsilon$, while the greedy gives a solution $I_{\text{alg}} = \{1, 2, 3\}$ with cost $U(I_{\text{alg}}) = 1 - 2\epsilon$.

Constructing the data structure R requires time $O(MN \log N)$, as discussed in Appendix 8.1. The operations small(R, j), 2ndsmall(R, j), big(R, j), and 2ndbig(R, j) on lines 9–11 can be performed in $O(1)$ time, and the operation remove(R, i) on line 14 in $O(M)$ time. Lines 5–16 of the algorithm, after the construction of the ordering data structure R, can therefore be computed in $O(MK)$ time, resulting to a total time complexity of $O(MN \log N + MK)$ and memory requirement of $O(MN)$ for the whole algorithm.

3.3 Confidence bands based on quantiles

A different approach to construct confidence bands is by considering quantiles for each time instance separately. In this approach the $\alpha/2$ and $1-\alpha/2$ quantiles of $\mathbf{X}[\cdot, m]$ are computed, separately for each time instance m. We will refer to this approach as the QUANTILE method. A major drawback of the QUANTILE method is that the $1-\alpha$ quantiles are formed independently for each m, so that the overall FWER is not controlled. An approximate FWER control can be added using a Bonferroni correction. The correction makes the quantiles smaller by defining them as $\alpha/(2M)$ and $1-\alpha/(2M)$ (instead of $\alpha/2$ and $1-\alpha/2$). We refer to the resulting method as BONFERRONI. The FWER control of BONFERRONI is approximate because the quantile estimation is sufficiently accurate only for large N.

One major problem with quantiles is that the smallest quantile that can be estimated from a dataset of size N is $1/N$. For the quantile methods to be applicable, it should hold $N > 2/\alpha$ for QUANTILE and $N > 2M/\alpha$ for BONFERRONI. For example, for the values of $\alpha = 0.1$ and $M = 25$ we need $N > (2 \cdot 25)/0.1 = 500$ observations to apply the BONFERRONI method. For many real datasets, the number of dimensions M can easily reach hundreds and thus the required N becomes thousands of observations. To actually reach the FWER control even more observations are needed, as we will see in our experimental evaluation (Fig. 2a).

3.4 Confidence bands based on distance measures

Time series are conventionally visualized and interpreted by plotting them against time. However, a time series of length M can also be interpreted as a point in an M-dimensional space. Accordingly, a collection of N time series becomes a cloud of data points. The distance of a data point from the center of this data cloud can be used as a measure of extremeness, as often done in the context of outlier detection. A natural assumption is to define the set of central observations I to be the $N-K$ time series with the smallest distance to the mean of all time series. This idea is formalized in Algorithm 2.

Algorithm 2 can be applied for any distance function dist$(\cdot, \cdot)$. In the case of multivariate normal data, a commonly-used distance measure is the *Mahalanobis distance*. Given a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ the ellipsoids

Algorithm 2 Distance measure based confidence bands

1: **input**: (1) dataset $\mathbf{X}$ of size $N \times M$ with N observations of length M, (2) number of observations to remove K, (3) a distance function $\text{dist}(\cdot, \cdot)$ between two time series of length M
2: **output**: the set of central observations I
3: Let $\mathbf{x}$ be a M-dimensional vector of means of columns of $\mathbf{X}$
4: Create a N-dimensional vector of distances $\mathbf{d}$ such that $d_i = \text{dist}(\mathbf{X}[i, \cdot], \mathbf{x})$
5: Let I be the set indices of the $N - K$ smallest values in $\mathbf{d}$.
6: **return** I

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \chi^2_M (1 - \alpha) \tag{4}$$

define surfaces in $\mathbb{R}^M$ that mark the boundary of the $1-\alpha$ confidence region, where χ^2_ν is the chi-squared distribution with ν degrees of freedom. The left hand side of Eq. (4) is the Mahalanobis distance, which is often used to detect outliers from multinormal datasets.

The Mahalanobis distance works well for approximately normal data as it takes into account the covariance structure. The confidence regions defined by Eq. (4) are M-dimensional ellipsoids. The MWE confidence bands correspond to M-dimensional rectangles with the cost function corresponding to minimizing the sum of the lengths of the sides of the rectangle. A compromise between the two approaches (rectangles vs. ellipsoids) would be to define regions based on M-dimensional spheres. This can be achieved by employing Algorithm 2, and using the L_2 norm as the underlying distance measure. We refer to the two instantiations of Algorithm 2, with the Mahalanobis distance and with the L_2 distance, as MAHA and L2, respectively. Because confidence bands in two dimensions correspond to M-dimensional rectangles in $\mathbb{R}^M$, the rectangle approximation is used for the MAHA and L2 methods as well. In other words, the confidence bands, not the confidence regions in $\mathbb{R}^M$, are used when testing if an observation is extreme or not.

The time complexity of Algorithm 2 is composed of the preprocessing time that may be needed to perform more efficiently the pair-wise distance computations for a predefined set of N time series of length M, $O(F(N, M)$, computing the mean on line 3, $O(MN)$, and the distance vector on line 4, $O(Nf(M))$, where $f(M)$ is the time needed to compute a single distance. The sorting on line 5 requires $O(N \log N)$ time. The time complexity of the full algorithm is therefore $O(F(N, M) + Nf(M) + NM + N \log N)$. If $F(N, M) \leq O(NM)$ and $f(M) \leq O(M)$ the computation of the distance measure causes no computational overhead, and in this case the total complexity is $O(NM + N \log N)$. Computing the Mahalanobis distance involves a costly inversion of a $M \times M$ matrix resulting to $F(N, M) = NM + M^\omega$, where ω is the exponent in the best-known algorithm for matrix multiplication and whose current value is $\omega = 2.373$ (Williams 2011). On the other hand, the use of L_2 norm causes no computational overhead.

4 Controlling the FWER

Due to the finite sample size N, the confidence bands BONFERRONI, MAHA, L2, and MWE control the FWER only approximately. In order to circumvent this problem, we

Algorithm 3 FWER control algorithm

1: **input**: (1) dataset $\mathbf{X}$ of size $N \times M$, (2) number of cross validation folds L, (3) α desired level of FWER control, (4) CB.ALGORITHM function to find a confidence band that outputs the set of central observations
2: **output**: the set of central observations I
3: $F \leftarrow$ Random partition of rows of $\mathbf{X}$ into L folds
4: Initialize $\boldsymbol{\Phi}$ to a matrix of size $L \times \lceil \alpha N_L \rceil$, where N_L is the maximum number of rows in a fold
5: **for** $l = 1 \ldots L$ **do**
6: $\quad \boldsymbol{\Phi}[l, \cdot] \leftarrow$ FWER.PROFILE($\cup_{i \neq l} F_i$, F_l, $\lceil \alpha N_L \rceil$, CB.ALGORITHM)
7: **end for**
8: $\boldsymbol{\phi} \leftarrow$ col.sums($\boldsymbol{\Phi}$)$/N$
9: $K_{\text{eff}} \leftarrow$ largest k such that $\boldsymbol{\phi}[k] \leq \alpha$
10: $I \leftarrow$ CB.ALGORITHM($\mathbf{X}$, K_{eff})
11: **return** I

Algorithm 4 FWER.PROFILE($\mathbf{X}$, $\mathbf{X}_{\text{test}}$, K, CB.ALGORITHM)

1: **input**: (1) dataset $\mathbf{X}$ of size $N \times M$, (2) test dataset $\mathbf{X}_{\text{test}}$ of size $N_{\text{test}} \times M$, (3) number of observations to remove K and (4) a confidence band algorithm to use
2: Initialize $\boldsymbol{\phi}$ to a vector of length K
3: **for** $k = 1 \ldots K$ **do**
4: $\quad I \leftarrow$ CB.ALGORITHM($\mathbf{X}$, k)
5: $\quad \boldsymbol{\phi}[k] \leftarrow$ n.obs.outside.cb($\mathbf{X}_{\text{test}}$, $\mathbf{X}$, I)
6: **end for**
7: **return** $\boldsymbol{\phi}$

use L-fold cross validation to estimate the true level of FWER control, as a function of the number of removed observations (K). We thus obtain a vector $\boldsymbol{\phi}$ of FWER values, where $\boldsymbol{\phi}[K]$ indicates the fraction of observations outside the confidence band (see Fig. 8 for an example). By finding the largest K for which $\boldsymbol{\phi}[K] \leq \alpha$, we get an effective number of observations K_{eff} to remove, such that FWER stays controlled at level α. The respective effective confidence level α_{eff} is defined as $\alpha_{\text{eff}} = K_{\text{eff}}/N$.

Algorithms 3 and 4 describe how the FWER profile, K_{eff} and respective confidence bands are computed. The FWER control algorithm (Algorithm 3) randomly partitions the dataset into L folds (line 3) and uses FWER.PROFILE (Algorithm 4) to compute FWER profile for each fold (line 6). The FWER profile for the whole dataset is an average of the fold profiles (line 8) and it is used to compute K_{eff} (line 9). The set of extreme rows is computed using any of the algorithms defined above.

The FWER.PROFILE algorithm computes the FWER profile for a given partition of the data and its implementation is shown in Algorithm 4. CB.ALGORITHM is applied repeatedly (line 4) with increasing K and the respective number of test observations outside confidence bands is stored to the vector $\boldsymbol{\phi}$ (line 5). The function n.obs.outside.cb returns a vector indicating how many observations are outside the confidence band. Note that once an observation falls outside the confidence band it will stay outside, as the band always gets narrower.

The simplest approach is to compute the FWER profile for all values of K up to $K = N$, but this is expensive for large datasets. Since one typically needs to estimate only the beginning of the FWER profile, say up to $K = \lceil \alpha N_L \rceil$, we may interrupt the estimation as the desired level has been reached. The result will be correct up to the truncation point. Notice that with the MWE, MAHA, and L2, the Algorithm 4 can

be executed within the loops of Algorithms 1 and 3 without increasing the overall time complexity. Quantile-based methods can be incorporated using $\alpha = k/N$ inside CB.ALGORITHM.

A larger number of folds, i.e., larger L, yield more accurate results as more data are used for the confidence band estimation in FWER.PROFILE. In the extreme case $L = N$ we get a leave-one-out type of situation where the test dataset consists of a single observation, but the running time of our algorithm is multiplied by the factor of L.

4.1 Effective values of the parameters K, α and M

The effective number of observations to remove, K_{eff}, might be smaller or larger than the naïve target $K = \lfloor \alpha N \rfloor$. For methods, such as MWE, that make the confidence band aggressively narrower as more observations are marked extreme, K_{eff} is smaller than $\lfloor \alpha N \rfloor$. For others, such as MAHA, where an observation can be extreme without making the confidence band narrower, K_{eff} is usually larger than $\lfloor \alpha N \rfloor$. For simplicity, from now on all statements about K_{eff} are for MWE if not stated otherwise.

The covariance structure of the data and the number of observations available (N) dictate how much K_{eff} differs from $\lfloor \alpha N \rfloor$. To obtain a better insight we introduce a parameter M_{eff}, which captures the effective dimensionality of the data. One extreme is reached when the time series differ from each other only by a baseline shift, making any pair of time time series perfectly correlated. As a result, there is effectively only one variable since all M time instances convey the same information. We say that the effective M, M_{eff}, is one. In this extreme case, only little data are required to get a reasonably accurate estimate of the confidence band. At the other extreme, all time series in the dataset are uncorrelated. In this case all M time instances carry information and $M_{\text{eff}} = M$. To get an accurate estimate of the confidence band we need a large value of N.

5 Experiments

We first briefly describe some synthetic and real datasets. Using these, we then provide an extensive empirical evaluation of the different methods to obtain confidence bands.

5.1 Datasets

5.1.1 Synthetic data

Synthetic data were created by adding noise to a base signal. By changing the degree of autocorrelation of the noise, datasets with different covariance structures were generated. This process allowed us to create datasets with varying effective dimension M_{eff} needed to illustrate the concepts described above.

A noise vector of length M was generated by applying a moving average filter of length w to a $M + w - 1$ vector of normal random variables $\mathcal{N}(0, 1)$. The larger the value of w, the higher the degree of autocorrelation that was added to the noise vector. Finally the noise vector was scaled to control the size of noise compared to the actual

signal. The scaling remained the same for all applied w. The case $w = M$ was treated separately. In this case the whole signal was offset by a value drawn from $\mathcal{N}(0, 1)$.

Additionally, synthetic ECG datasets were created by adding noise to the average normal heart beat from `heartbeat-normal`. These datasets differ both in length (M) as well as by the amount of smoothing, measured using the relative width w/M of the smoothing window.

5.1.2 ECG data

The dataset consists of electrocardiographic (ECG) raw signal. PhysioNet offers several kinds of heart beat related data (Goldberger et al. 2000).[2] We chose the MIT-BIH arrhythmia database,[3] which contains annotated 30 min records of normal and abnormal heart beats, originally used for the evaluation of arrhythmia detectors (Moody and Mark 2001). We selected test subject 106, whose record contains 1,507 normal beats and 520 abnormal beats with premature ventricular contraction (PVC). We used the beat type annotation locations to align the beats and selected a time window of $[-300, 400]$ ms around these locations as the window of interest. This creates two datasets with $M = 253$ time points: `heartbeat-normal` (N = 1,507) and `heartbeat-pvc` ($N = 520$).

5.1.3 Temperature data

We used the publicly available Global Historical Climatology Network (GHCN) daily dataset,[4] from US National Oceanic and Atmospheric Administration's National Climatic Data Center (NOAA NCDC).[5] Using the information found in ghcnd-inventory.txt, we verified that the station ITE00100554 in Milan, Italy, contains the longest range of measurement years from 1763 to 2008. Several meteorological variables are available of which we select the maximum temperature ("tmax"). We then aggregated the daily temperatures to average monthly temperatures. Only December 2008 was missing completely, so the final dataset `max-temp-milan` contained years 1763–2007 corresponding to $N = 245$ observations.

5.1.4 Power-consumption data

The `UCI-power` dataset is the individual household electric power consumption data from the UCI machine learning repository (Bache and Lichman 2013).[6] It consists of hourly averages of the variable "active.power".

Table 1 shows statistics of our datasets and some related properties. The synthetic datasets demonstrate that the smallest achievable FWER decreases as the dataset

[2] http://physionet.org/.

[3] http://physionet.org/physiobank/database/mitdb/.

[4] ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily.

[5] http://www.ncdc.noaa.gov/.

[6] http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption.

Table 1 Datasets and their properties computed using MWEC, $L = 4$ and target confidence level $\alpha = 0.1$

	N	**M**	MIN FWER	α_{eff}	K_{eff}	$\boldsymbol{\Delta U}$
`synth-ECG-0prc`	1,000	25	$.068 \pm .002$	$.012 \pm .002$	12 ± 1.8	$.960 \pm .006$
`synth-ECG-75prc`	1,000	25	$.064 \pm .002$	$.014 \pm .002$	14 ± 1.9	$.958 \pm .006$
`synth-ECG-100prc`	1,000	25	$.002 \pm .000$	$.072 \pm .001$	72 ± 1.2	$.569 \pm .022$
`heartbeat-normal`	1,507	253	.038	.017	26	.639
`heartbeat-pvc`	520	253	.104	–	–	–
`max-temp-milan`	245	12	.069	.004	1	.941
`UCI-power`	1,417	24	.029	.037	52	.830

MIN FWER is the smallest reachable FWER, reached when the training data envelope is directly used as the confidence band in Algorithm 4. ΔU shows how much the original data envelope has to be narrowed down to achieve the $1 - \alpha$ confidence level, i.e., $\Delta U = U(I)/\{whole\ data\ envelope\}$. For synthetic data the mean of five iterations is given together with the standard error. Dashes indicate values that cannot be computed

becomes more correlated. This corresponds to smaller values of effective dimension M_{eff} allowing α_{eff} and K_{eff} to increase.

5.2 Properties of the confidence bands

FWER *control.* Confidence bands are first computed for synthetic ECG datasets with $M = 25$ and variable N. Consequently the FWER control of the bands is tested using a test dataset of size $N = 10^4$.

Figure 2a shows the observed FWER of the methods when no FWER control procedure is applied. Clearly the QUANTILE method does not control the FWER at all. As the number of observations increases, MWE and BONFERRONI converge to the target confidence level but distance based measures converge to zero, i.e., become overly conservative.

The observed FWER when FWER control procedure has been applied is shown in Fig. 2b. The FWER is controlled for all methods for $N > 500$ and remains controlled as dataset size increases. All methods yield similar results at $N = 500$ because at that point even the training data envelope cannot always provide full FWER control at level $\alpha = 0.1$. The brief upward notch for BONFERRONI at $N = 750$ is a rounding artifact: the quantiles can be estimated at $1/N$ intervals and for small N the grid is too sparse. As N increases the FWER converges toward the desired level, except for distance based methods.

5.2.1 Confidence band width

Figure 3a illustrates how the confidence band width grows as N increases. The MWE algorithm produces the narrowest confidence bands, as designed. The BONFERRONI bands are narrower than MWE bands for the largest N because BONFERRONI can mark more time series as extreme than MWE. The confidence band width is upper bounded by the true distribution of the data, which becomes more accurately represented as N increases. Also in Fig. 4 the FWER is lower bounded by the target rate α.

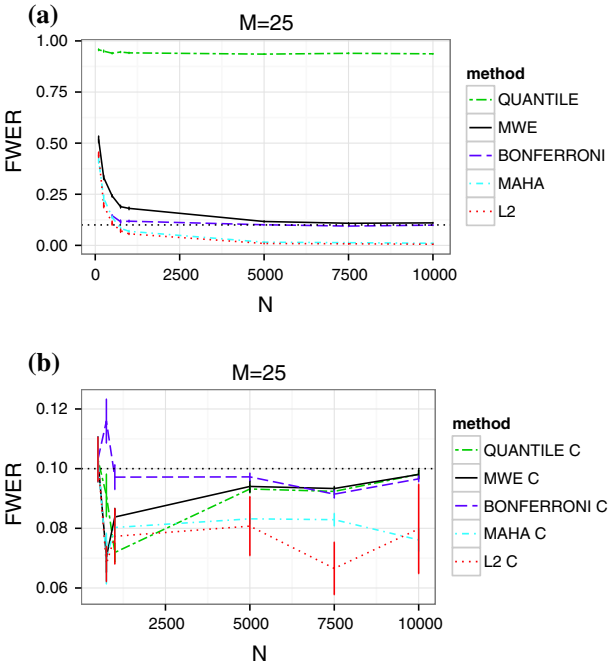


Fig. 2 Observed mean FWER along with standard error **a** without and **b** with FWER control procedure. The dataset is synthetic heart beat data with $M = 25$, $N = \{100, 250, 500, 750, 1000\}$ and $w/M = 0.75$ smoothing window. The test dataset is of size $N_{\text{test}} = 10^4$. The *horizontal dotted line* shows the target confidence level $\alpha = 0.1$. For the BONFERRONI method quantiles cannot be estimated for $N = \{100, 250\}$ due to insufficient data for quantile estimation. When FWER control procedure is applied, all methods keep the FWER under control after a sufficient dataset size of $N > 500$ has been reached

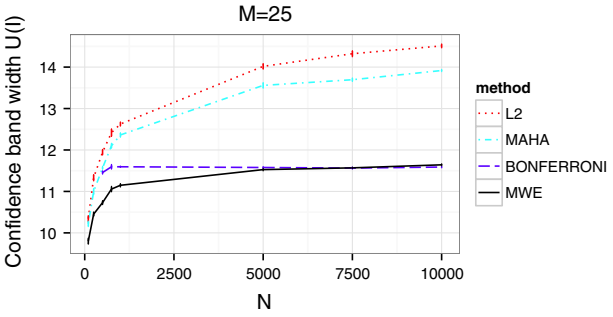
Figure 3b shows the situation with FWER control applied. BONFERRONI and QUANTILE bands are slightly narrower than MWE because more data are removed. All three provide roughly the same level of FWER control but MWE does it with fewest observations removed.

More FWER control and confidence band results for a larger pool of datasets and confidence band methods can be found in an additional material distributed together with the source code.[7]

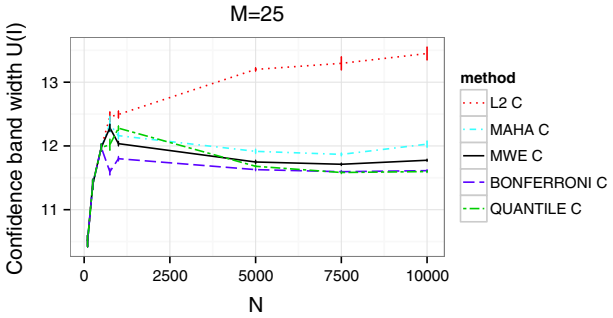
5.2.2 *Effective data dimension* M_{eff}

Figure 4 shows the effect of varying the covariance structure of the data. The FWER is better controlled for signals with higher autocorrelation, i.e., smaller M_{eff}. Also, with increasing N, FWER tends faster towards the target level $\alpha = 0.1$ if M_{eff} is small.

[7] https://bitbucket.org/jtkorpel/mwe_2014.



(a) no FWER correction



(b) with FWER correction

Fig. 3 Mean and standard error of the $1-\alpha$ confidence band width for the same datasets and confidence level as in Fig. 2a. Of methods that mark $K = \lfloor \alpha N \rfloor$ observations as extreme, MWE produces the smallest envelopes

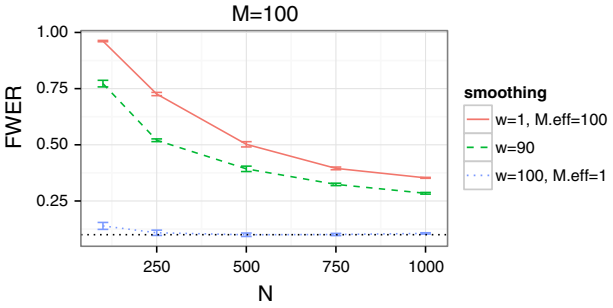


Fig. 4 Mean (FWER) control using MWE on synthetic heart beat data with different covariance structures. The synthetic time series were $M = 100$ points long with added moving average (MA) Gaussian noise. The MA lengths w were $\{1, 90, 100\}$, where $w = 1$ corresponds to $M_{\text{eff}} \approx M$ and $w = 100$ to $M_{\text{eff}} = 1$. The *horizontal dotted line* marks the target confidence level $\alpha = 0.1$. *Error bars* show the standard error of mean

5.2.3 Scalability of the algorithms

The running times of the algorithms are shown in Fig. 5. The overall scalability follows the theoretical estimates. Small deviations are due to R programming environment,

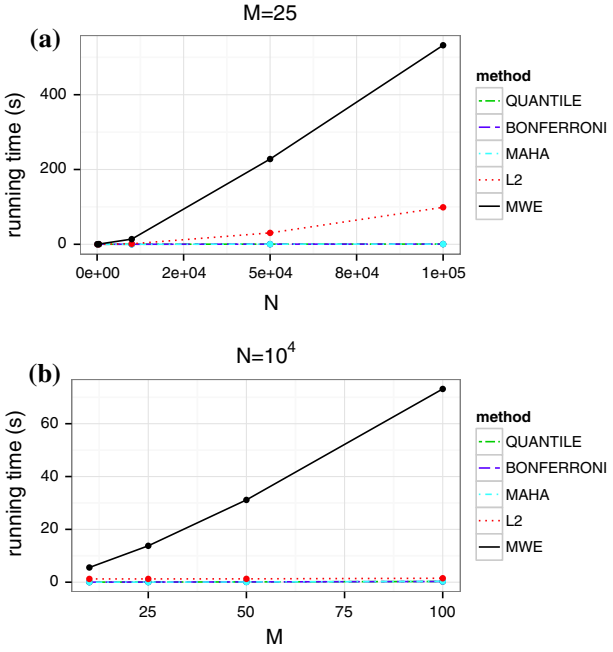


Fig. 5 Running times of the algorithms as a function of M and N

which has been used to implement the algorithm and experiments. For a dataset of size $N = 10^4$ and $M = 100$ the actual running times are approximately 73 s for MWE, 0.2 s for MAHA and 1.4 s for L2.

5.3 Consistency of extreme values between methods

The methods use different criteria to define extremeness and hence they label different time series as extreme. A fictive example of this is shown in Fig. 6, where especially MAHA differs considerably from the rest of the methods. QUANTILE, L2 and MWE select mainly observations whose removal makes the confidence band narrower, but differ in how many and which observations they pick.

Similar effects can be observed for more realistic datasets as well. An example is shown in Table 2, which lists the number of common extreme observations between pairs of methods. The number of overlaps depends on the distance measures and the dataset. Here overlaps are used to quantify the differences between methods.

QUANTILE marks almost all of the 1,000 observations as extreme, which indicates that 95 % of the observations belong to the 10 % tail in at least one time point. For BONFERRONI the number drops down to 69. This underlines the need for some kind of error rate control.

By design, the other methods mark $\lfloor \alpha N \rfloor = 100$ observations as extreme. However, they do this very differently: for example MWE has 29 common labelings with MAHA but only one with L2. In addition, MAHA and L2 do not share a single common observation.

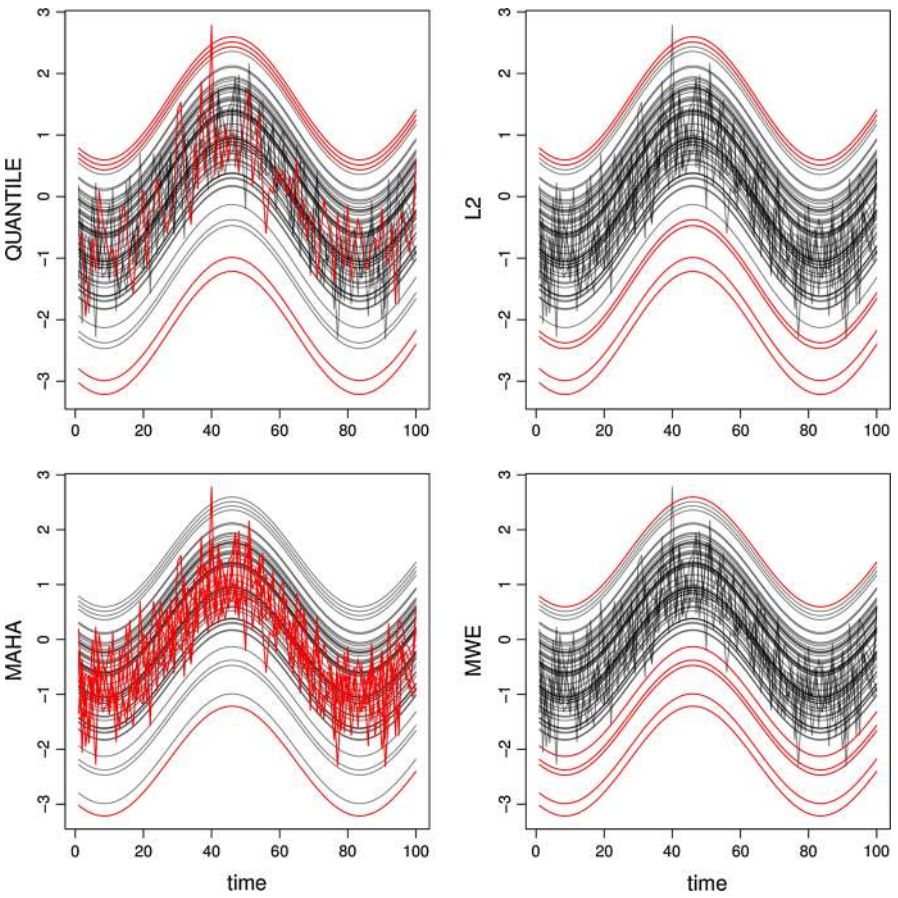


Fig. 6 Example of how methods consider different observations to be extreme. The dataset is a simple toy dataset consisting of a set of 55 perfectly autocorrelated time series and a set of 5 time series with added uncorrelated normal noise. Extreme time series are plotted in *red*. Notice how in this example the QUANTILE and MAHA mark extreme observations that do very little to reduce the width of the confidence band, whereas MWE consistently chooses to remove observations that make the confidence band narrower

Table 2 Total number of extreme observations (diagonal) and the number of common extreme observations between methods, when no FWER control procedure has been applied

	QUANTILE	BONFERRONI	MAHA	L2	MWE
QUANTILE	945	69	100	81	100
BONFERRONI	–	69	22	3	57
MAHA	–	–	100	0	29
L2	–	–	–	100	1
MWE	–	–	–	–	100

The dataset is synthetic heart beat data with $N = 1{,}000$, $M = 25$, and $w = 0.75 \times M$, confidence level $\alpha = 0.1$

Table 3 Number of common extreme observations between methods when FWER control procedure has been applied

	QUANTILE C	BONFERRONI C	MAHA C	L2 C	MWEC
QUANTILE C	24	24	11	5	11
BONFERRONI C	–	47	19	7	20
MAHA	–	–	112	5	9
L2 C	–	–	–	232	2
MWEC	–	–	–	–	22

The dataset is the same as in Table 2

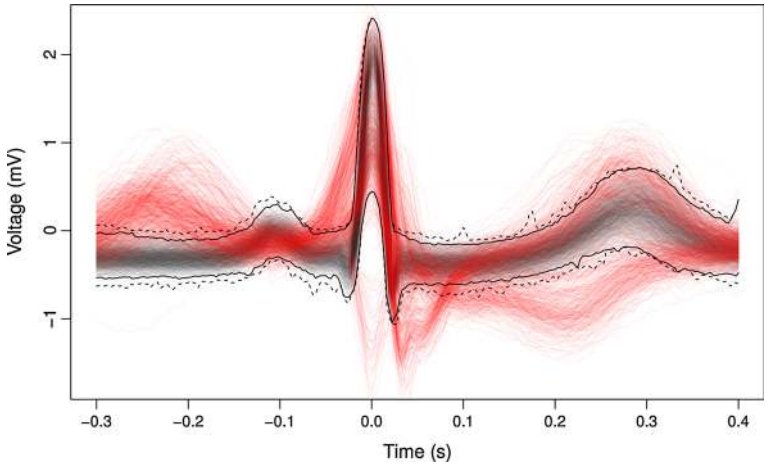


Fig. 7 Heart beats from MIT arrhythmia database, subject 106. A total of 1,507 normal heart beats are shown in *black* and 520 beats with premature ventricular contraction in *red*. The 90 % confidence band for the normal beats is shown in *bold black* and the fold method MWE 90 % confidence band ($K_{\text{eff}} = 25$) in *thick dashed black*. All beats are centered around the beat annotation location. The premature contraction is clearly visible around −0.25 s

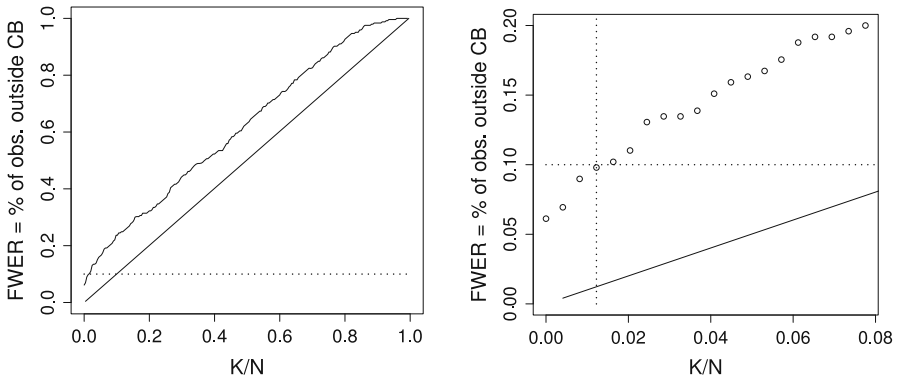
This emphasizes the fact that even a small change in the distance measure can have a huge impact on the outcome.

Table 3 contains the same experiment but with FWER control applied. The difference between distance based methods and MWE is clearly visible: whereas MAHA and L2 mark over 100 observations as extreme MWE marks only 22. This is possible because in the MWE method each removed observation makes the confidence band narrower, whereas the same is not true for distance based measures. Mostly there is no specific condition where overlap would be good or bad. The amount of overlap depends on the data and Fig. 6 illustrates a situation where the different definitions of extremeness lead to very different results.

5.4 Examples using real data

5.4.1 ECG data

We applied the MWE fold algorithm to the normal beats data using $L = 4$ folds. The resulting profile shows that the smallest achievable FWER is approximately 0.051,



(a) Whole profile **(b)** Zoom in (*lower left*)

Fig. 8 FWER profile $\boldsymbol{\phi}$ against proportion of removed observations (K/N) for temperature data using $L = 245$. The whole profile is shown on the *left* and a zoom-in on the *right*. The *solid line* represents $y = x$ and the *dashed line* shows the position of the desired confidence level. The first observation in **b** corresponds to $k = 0$, i.e., the data envelope. Note that the desired confidence level $\alpha = 0.1$ is reached already at approximately $K/N = 3/245 \approx 0.012$. The envelope of the whole original dataset controls FWER at rate 0.06

indicating that the dataset envelope would approximately provide 95 % confidence band. A conservative estimate for the 90 % confidence band is achieved using $K_{\text{eff}} = 25$.

Figure 7 contains an illustration of the data along with 90 % confidence bands. The fold method confidence bands are slightly wider but seem to coincide with the regular MWE around signal peaks. Notice how most of the PVC anomalies lie outside the confidence bands around -0.25 s.

5.4.2 Temperature data

Applying the MWE fold algorithm with $L = 245$, i.e., using leave-one-out folding scheme produces the profile in Fig. 8. The lowest achievable FWER turns out to be 0.069 meaning that there is not enough data to achieve a 95 % confidence band. A conservative estimate of the 90 % confidence band is found using $K_{\text{eff}} = 3$. Note that in Table 1 using $L = 4$ the same result was $K_{\text{eff}} = 1$. Thus the partitioning in the FWER control procedure has an effect, but the leave-one-out scheme used here provides the most accurate result because of maximal size of the training dataset.

The Milan dataset along with 90 % confidence bands is shown in Fig. 9. There are three years, namely 2003, 2006, and 2007, which lie outside the confidence bands; all of these are ones with a documented heat wave.[8] The number of observations limits the accuracy of the analysis. With the current $N = 245$ observations we can afford to mark three observations as extreme, before the data envelope becomes too narrow for

[8] http://en.wikipedia.org/wiki/2003_European_heat_wave.
http://en.wikipedia.org/wiki/2006_European_heat_wave.
http://en.wikipedia.org/wiki/2007_European_heat_wave.

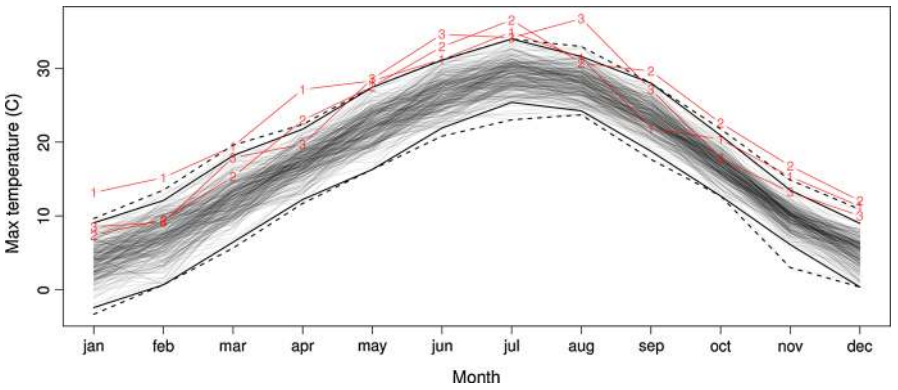


Fig. 9 Monthly temperatures in Milan, Italy from 1763 to 2007 (GHCN station id: ITE00100554, ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/all/ITE00100554.dly). The whole dataset is shown in *fine black*, regular MWE 90 % confidence band ($k = 24$) in *thick black* and the fold method MWE 90 % confidence band ($K_{\text{eff}} = 3$) in *thick dashed black*. The number of observations is 245. The extreme observations correspond to years 2003 (#1), 2006 (#2), and 2007 (#3) and are shown in *red*

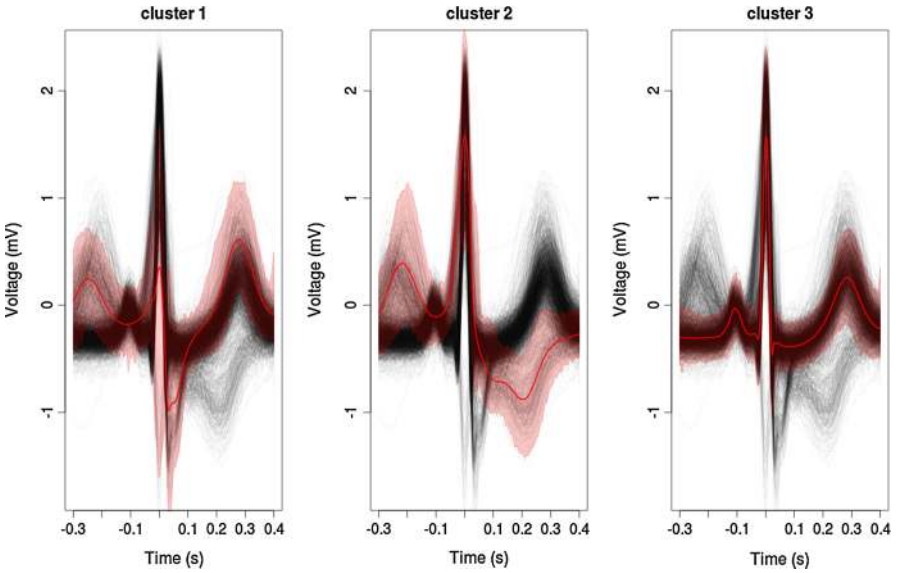


Fig. 10 Heart beats from MIT arrhythmia database, subject 106 (same data as in Fig. 7). Normal and PVC beats have been pooled (N =2,027) and are plotted in *faint black*. Cluster means are shown in *bold red* and 97 % MWE confidence bands are shown as *red bands* around the means

the desired FWER control. If we had more data, the confidence band could be estimated with more detail and some additional years would become extreme as well.

5.5 Disjoint confidence bands

If the dataset $\mathbf{X}$ has multiple modes, the confidence bands should be split accordingly. The straightforward solution is to cluster the dataset in c clusters and compute the

confidence bands separately for each cluster using $1 - \alpha/c$ as the confidence level. This is primarily a clustering problem with the main question being that of choosing a "correct" number of clusters. This is an open problem in data mining (Xu and Wunsch 2005) and therefore we demonstrate the approach using a predefined number of clusters. Dataset size is another limiting factor that comes into play, as clusters usually end up having too few observations to guarantee FWER control at the desired level α.

An example of this is shown in Fig. 10, where the dataset is a union of `heartbeat-normal` and `heartbeat-pvc`. As an example we have identified $c = 3$ clusters from the data using a standard k-means algorithm. Confidence bands are then constructed separately for each cluster using $1 - \alpha/c$ as the confidence level, in this case $1 - 0.1/3 \approx 97\ \%$. Comparison to Fig. 7 reveals that clusters 1 and 2 correspond to the PVC beats and cluster 3 to normal beats. Cluster sizes are 248, 251, and 1528 which matches to the numbers of normal beats (1,507) and PVC (520), respectively.

6 Related work

Confidence bands presented in this work focus on *describing* a multivariate target distribution consisting of time series, where the series are assumed to have a fixed length M. The estimated quantity is the location of a given percentile of the distribution. A related approach is *prediction*, where the interest is in predicting the future samples from the same population using e.g., prediction intervals (Hahn and Meeker 1991). For time series this means predicting the value of the series k time steps ahead at $t + k$, when observations up to t are available. In prediction it is assumed that the series gets longer as time passes, in contrast to our approach.

In some applications, such as with temperature data or when analyzing some repeated bio-signal waveform, new instances of a complete time series are constantly generated. In these applications it would make sense to speak of prediction bands that would describe the estimated location of a future realization of the M-point time series. Mahalanobis distance is an example of such a region, as it can be used to form the prediction region for a multivariate normal distribution (see Eq. (4)). However, in general, we do not consider prediction bands in this work.

We were surprised to find only very few references to principled, non-parametric approaches to the problem of finding simultaneous multivariate confidence intervals for autocorrelated data. In their book Davison and Hinkley present an idea of a graphical test involving confidence bands and multiplicity correction (see Davison and Hinkley 1997, p. 154). Mandel and Betensky (2008) extend the idea by providing exact algorithms to solve the problem, but do not apply the algorithms to time series data. Both approaches use the rank of the most extreme coordinate value within an observation as the ranking criterion. This leads to problems as the dimension of the data grows, because many observations start assuming equal ranks and finding quantiles in the rank distribution becomes difficult. The methods applied in the present work use cost functions that do not suffer from this phenomenon.

Another related field of research is that of outlier detection, either in a traditional multivariate setting (Aggarwal 2013) or for time series data (Gupta et al. 2013). The

distance functions defined by the outlier detection methods can be directly used as ranking criteria for extremeness. As outlier detection methods have not yet been applied to the construction of confidence bands, we included Mahalanobis (MAHA) and the Euclidean distance (L2) based confidence bands as comparison conditions. One should, however, bear in mind that MAHA and L2 confidence bands are based on a different definition of extremeness than MWE. Depending on the application the researcher might find one or the other more justified.

Interestingly, when the extremeness of some items with respect to others is defined in terms of dissimilarity rather than strict distance metrics, an outlier/deviance detection algorithm such as the one by Arning et al. (1996) may closely resemble the MWE approach adopted here. The database deviation detection approach by Arning et al. (1996) proceeds by adding elements to an "exception set" using dissimilarity and cardinality functions. The same happens in the MWE approach, where the cost function (envelope width) plays the role of a dissimilarity function but the cardinalities are not considered. The largest differences between the two approaches are the size of the "exception set" (fixed in MWE, variable in Arning's approach) and the application domain (time series for MWE, text data for Arning).

In the field of information visualization, the time series have been studied a lot, the confidence bands being one of the visual components used; see Aigner et al. (2011) for a review.

In statistics a related problem is the one of finding the confidence region for the estimate of the mean of a random vector. Examples of the different approaches proposed can be found for example in Owen (1990), Efron (2006), and Arlot et al. (2010). However, confidence bands are not mentioned in any of them.

Lastly, confidence bands can often be formed by inverting a statistical test. The standard multiplicity correction procedures (see Dudoit et al. 2003) define such tests, but the inversion is straightforward only for the Bonferroni method. However, recent developments by Guilbaud (2008) suggest that also step-wise correction procedures can be used to form simultaneous confidence regions.

7 Conclusions

The focus of this work is on the analysis of datasets with correlated variables such as time series. We introduce a minimum width envelope (MWE) method that can be used to compute confidence bands when several observations of a time series are available. The method is intuitive, non-parametric and adjusts automatically to varying degrees of correlation within the data. We also provide a procedure to ensure that the confidence bands are such that the FWER remains controlled.

In this work we define the MWE problem and show that, when the area of the confidence band is used as the cost function, the problem is **NP**-hard and the complement objective function is hard to approximate. We also provide a greedy algorithm to solve the problem along with several notes on how to implement the algorithm efficiently. The algorithm turns out to have time complexity $\mathcal{O}(MN \log N + MN)$, where N is the number of observed time series and M is the dimension of the data.

To ensure that the MWE confidence bands control the FWER as intended, we provide a cross validation type of approach. The approach can be applied to other confidence band computation methods as well.

We also compare MWE to other methods that can be used to form confidence bands. We show that due to the lack of multiplicity correction, naïve quantiles are not an option to use for robust time series analysis. The quantile method can be improved using the Bonferroni correction, but the resulting confidence bands cannot be reliably estimated for small N due to problems in estimating the tails of the empirical probability density function. Also, the Bonferroni correction does not take into account the correlation structure of the data and is known to be very conservative thus lacking statistical power.

We tested multivariate distance based confidence bands as well. As such these methods create overly conservative confidence bands. The FWER control procedure makes the bands less conservative but the achieved FWER control varies a lot. The bands are also wider than those of the other methods (especially for L2).

Summarizing, we have introduced an intuitive and fast method of computing confidence bands for time series data. The fact that MWE adapts automatically to the covariance structure of the data is an important feature, because time series data usually contain significant auto-correlations. Using the fold approach approximate FWER control is achieved also for small datasets.

With MWE, as with all data-driven confidence band computation methods, lack of data (small N) leads to problems. High dimensional datasets simply need lot of data for reliable estimation. Another problematic situation arises in special cases where the time series lie parallel to each other, i.e., the effective dimension is close to one. If two or more time series now lie very close to one another (or the exact same series is repeated), the greedy algorithm cannot shrink the confidence band from the respective direction as there is no reduction in confidence band width to be gained. In these cases the distance based methods produce narrower bands.

It should be kept in mind that the confidence bands in two dimensions correspond to a hypercube in the M dimensional data space. The cube is a crude approximation even to the multivariate normal distribution, for which an ellipsoid would be more appropriate. However, as time series are conventionally presented by plotting their values against time and this visualization unveils a lot of useful information, confidence bands are needed especially if visual inspection of data is required.

Distributions with multiple modes pose a problem for any confidence region/band estimation method in both univariate and multivariate domains. The main difficulty lies in deciding which of the modes are to be treated as separate clusters and which are just coincidental collections of outliers. Finding the "correct" number of clusters is an open problem in data mining community with approaches such as *minimum-description length principle* (MDL) and *Bayesian information criterion* (BIC) providing reasonable answers. Assuming that the desired number of clusters c is known, a set of disjoint confidence bands can be constructed by applying the MWE separately to each cluster using $1-\alpha/c$ as the confidence level. The main problem in this straightforward approach is that in many cases the clusters have so few observations that the desired level of FWER control (α/c) cannot be reached. This problem could be circumvented by relaxing the requirement that each cluster should contribute an equal number of outliers. Applying, e.g., the k-means minus-minus by Chawla and Gionis (2013) the clus-

ters and outliers can be estimated simultaneously. Also the cross-validation approach to guarantee FWER control can be easily included. Changing the distance measure from the Euclidean distance to a one that focuses on envelope widths could then provide a basis for a MWE method for distributions with multiple modes. To keep the scope of this work compact and as there still is research to be done in the unimodal problem, we decided to leave the extension of MWE to multimodal distributions as future work.

Other future research topics could include the study of approximability if the type of data is restricted in some way. As the MWE is in all practical applications a geometric problem, our intuition is that an approximation ratio might exist for some special data types. Another line of research would be to think of ways to control other error rates than FWER.

Acknowledgments The authors would like to thank Andreas Henelius for helpful discussions and suggestions. The work of J. Korpela and K. Puolamäki was supported in part by the Revolution of Knowledge Work Project, funded by Tekes (The Finnish Funding Agency for Innovation).

8 Appendix

8.1 Efficient implementation of order data structure R

This section describes the data structure R, referred to in Algorithm 1, that allows the MWE algorithm to be efficient. R stores the ordering information for columns j of the

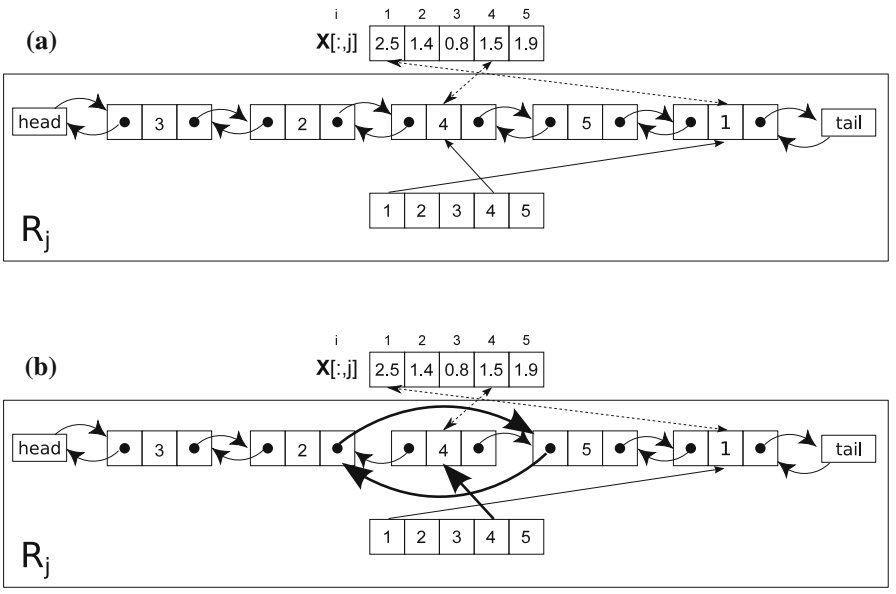


Fig. 11 **a** An example data structure R_j that combines a doubly linked list and an index vector to make the retrieval of largest/2nd largest ranks and associated observation indices a constant time operation. This structure allows the efficient implementation of rows 9–10 and 13 in Algorithm 1. **b** Same data structure with observation $i = 4$ removed showing the update of links within the list

data matrix $\mathbf{X}[i, j]$. A substructure $\mathrm{R_j}$ for a single column j with $N = 5$ observations is shown in Fig. 11a. The rank order of the values in column j are stored in a doubly linked list, with the first element corresponding to the index i of the smallest element in $\mathbf{X}[\cdot, j]$. The second element contains the index of the second largest value etc. The indices of the (second) largest and (second) smallest values can be extracted in $O(1)$ time for a single column j, or in time $O(M)$ for all columns (all values of j).

The substructure $\mathrm{R_j}$ additionally contains a vector of length N, where the ith item is a pointer to the node of the doubly linked list with a value of i. With the help of this additional vector, it is possible to delete (bypass) a node corresponding to any time series i from the doubly linked list as shown in Fig. 11b. This takes $O(1)$ time for single column j and $O(M)$ time for the whole time series. The data structure can be initialized in $O(MN \log N)$ time with the memory requirement of $O(MN)$.

References

Aggarwal CC (2013) Outlier analysis. Springer, New York

Aigner W, Miksch S, Schumann H, Tominski C (2011) Visualization of time-oriented data. Human–computer interaction series. Springer, New York

Arlot S, Blanchard G, Roquain E (2010) Some nonasymptotic results on resampling in high dimension, I: confidence regions. Ann Stat 38(1):51–82. doi:10.1214/08-AOS667

Arning A, Agrawal R, Raghavan P (1996) A linear method for deviation detection in large databases. In: KDD, pp 164–169

Bache K, Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml

Chawla S, Gionis A (2013) k -means: a unified approach to clustering and outlier detection. In: Proceedings of SIAM international conference data mining (SDM)

Davison A, Hinkley D (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. Stat Sci 18(1):71–103

Efron B (2006) Minimum volume confidence regions for a multivariate normal mean vector. J R Stat Soc Ser B Stat Methodol 68(4):655–670. doi:10.1111/j.1467-9868.2006.00560.x

Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23):E215–20

Guilbaud O (2008) Simultaneous confidence regions corresponding to Holm's step-down procedure and other closed-testing procedures. Biom J 50(5):678–92. doi:10.1002/bimj.200710449

Gupta M, Gao J, Aggarwal CC (2013) Outlier detection for temporal data: a survey. IEEE Trans Knowl Data Eng 25(1):1–20

Hahn GJ, Meeker WQ (1991) Statistical intervals: a guide for practitioners. Wiley, New York

Mandel M, Betensky R (2008) Simultaneous confidence intervals based on the percentile bootstrap approach. Comput Stat Data Anal 52(4):2158–2165. doi:10.1016/j.csda.2007.07.005

Moody GB, Mark RG (2001) The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 20(3):45–50

Owen A (1990) Empirical likelihood ratio confidence regions. Ann Stat 18(1):90–120. doi:10.1214/aos/1176347494

Williams VV (2011) Breaking the coppersmith-winograd barrier, manuscript

Xavier EC (2012) A note on a maximum k-subset intersection problem. Inf Process Lett 112(12):471–472. doi:10.1016/j.ipl.2012.03.007

Xu R, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678