Low redundancy estimation of correlation matrices for time series using triangular bounds

Erik Scharwächter[1,2] ✉, Fabian Geier[2], Lukas Faber[2], and Emmanuel Müller[1,2]

[1] GFZ German Research Centre for Geosciences, Potsdam, Germany
[2] Hasso Plattner Institute, Potsdam, Germany
{erik.scharwaechter,fabian.geier,emmanuel.mueller}@hpi.de
lukas.faber@student.hpi.de

Abstract. The dramatic increase in the availability of large collections of time series requires new approaches for scalable time series analysis. Correlation analysis for all pairs of time series is a fundamental first step of analysis of such data but is particularly hard for large collections of time series due to its quadratic complexity. State-of-the-art approaches focus on efficiently approximating correlations larger than a hard threshold or compressing fully computed correlation matrices in hindsight. In contrast, we aim at estimates for the *full pairwise correlation structure without computing and storing all pairwise correlations*. We introduce the novel problem of low redundancy estimation for correlation matrices to capture the complete correlation structure with as few parameters and correlation computations as possible. We propose a novel estimation algorithm that is very efficient and comes with formal approximation guarantees. Our algorithm avoids the computation of redundant blocks in the correlation matrix to drastically reduce time and space complexity of estimation. We perform an extensive empirical evaluation of our approach and show that we obtain high-quality estimates with drastically reduced space requirements on a large variety of datasets.

1 Introduction

The monitoring of earth, society and personal life through various sensors has led to a ubiquity of large-scale collections of time series. Correlation analysis for all pairs of time series is often the first step of analysis of such data. In the past decade, many works have used estimates of the full pairwise correlation matrix among time series, e.g., to infer functional brain networks [17], for portfolio selection in empirical finance [9], to detect periods of financial crisis [19] and to better understand the climate system [20]. Since the time and space complexity for computing the full pairwise correlation matrix is quadratic in the number of time series, analyses that rely on exact computation of the full matrix do not scale with the increasing size of time series collections. For this reason, there is a need for approaches that estimate all pairwise correlations without computing and storing the entire matrix.

We introduce the novel problem of low redundancy estimation for correlation matrices. A low redundancy estimate describes the *complete correlation matrix*

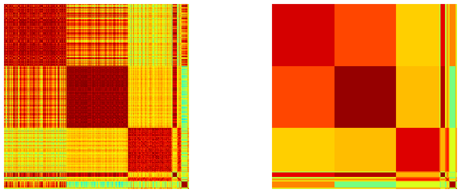

Fig. 1. Example correlation matrix R (left) and low redundancy estimate $\mathfrak{R}$ (right).

R of a time series collection using a *smaller representation* $\mathfrak{R}$ and *without computing all pairwise correlations*. Our estimation approach COREQ (CORrelation EQuivalence) is driven by the observation that many time series collections show inherent group structure that leads to blocks of redundant entries in the correlation matrix. We exploit this structure by computing equivalence classes of highly correlated time series and pooling the redundant correlation estimates into a single class estimate. The resulting estimate is visualized in Fig. 1. We describe an algorithm to obtain the estimate $\mathfrak{R}$ on the right directly from the data after computing only a small fraction of the actual correlations in R. The computational problem lies in finding—with as few correlation computations as possible—a suitable partition of the time series collection into equivalence classes that allows correlation estimation with bounded loss.

Our contributions are as follows. We formalize low redundancy estimation as an approximation problem and formally derive low redundancy estimates with error guarantees. Furthermore, we propose a greedy approximation algorithm and two powerful heuristics to obtain high-quality estimates with few correlation computations. We carefully evaluate our algorithm on 85 time series collections from the UCR Time Series Classification Archive [1] and a large satellite image time series dataset from the geoscientific domain as a real-life use case.

2 Related work

There are two challenges for efficient correlation estimation for large time series collections. The first challenge is the increasing *number* of time series that are jointly analyzed, while the second challenge is the increasing *velocity* of newly arriving observations in streaming time series.

COREQ addresses the first challenge. Most work in the field has been done on rapidly retrieving all pairs of highly correlated time series [25, 23, 16] and avoiding the computation of weak correlations. Conceptually, all these approaches discard information about weak correlations. In contrast, our COREQ algorithm provides estimates for the *complete correlation structure*, including weak correlations. Low-rank approximations to a correlation matrix remove redundancies for a more space efficient representation of the full correlation structure, but existing methods [24, 7] take fully estimated correlation matrices as inputs for their approximations. In contrast, we aim at low redundancy estimates *without computing all pairwise correlations* first. Mueen et al. [11] propose two algorithms to approximate all entries in the correlation matrix that are larger than some threshold τ. By design, they lose information about correlations below the hard threshold τ, while we provide accurate estimates for all correlations. We briefly describe their algorithms in Section 5 and evaluate COREQ against them.

Methodologically, COREQ exploits structure in time series collections by computing equivalence classes of time series that behave similarly under correlation. There is extensive literature on clustering time series with similar behavior for generic subsequent processing [10, 14, 18]. In contrast to these works, COREQ has *theoretical quality guarantees* for the resulting correlation estimates.

Orthogonal to our approach, works on streaming time series have focused on efficient updating schemes for correlation monitoring [25, 4, 12], robust correlation tracking [13], detection of lag correlations [15, 21, 22] and correlated windows [2, 5, 6] in streaming time series. We assume for now that our time series collections are static and defer streaming versions to future work.

3 Low redundancy estimation

3.1 Preliminaries

Let $\mathcal{X} = \{X_1, ..., X_N\}$ be a collection of N univariate time series of length T with $X_i = (X_{i1}, ..., X_{iT})$. We assume that the time series are equi-length and temporally aligned as in many use cases from the geosciences, neuroimaging, finance and other domains. The Pearson correlation coefficient between time series X_i and X_j (at lag 0) is given by $\rho_{ij} = \frac{1}{T}\sum_{t=1}^{T} \frac{X_{it}-\mu_i}{\sigma_i} \cdot \frac{X_{jt}-\mu_j}{\sigma_j}$, where μ_i and σ_i denote the mean and standard deviation of time series X_i, respectively. The correlation coefficient captures linear relationships and ranges from 1 (strong positive correlation) to -1 (strong negative correlation). A value of 0 means that time series are uncorrelated. The matrix $R \in [-1, 1]^{N \times N}$ denotes the symmetric correlation matrix that contains all pairwise correlations between the input time series, i.e. $R = (\rho_{ij})_{i,j \in \{1,...,N\}}$. A useful property of Pearson's correlation coefficient is that it comes with triangular bounds similar to the triangle inequality in metric spaces [8]. These bounds allow estimating the correlation between two time series X_i and X_j via their correlations with a third time series X_k:

Theorem 1 (Triangular bounds). *For time series X_i, X_j and X_k it holds that* $\rho_{ik}\rho_{kj} - \sqrt{(1-\rho_{ik}^2)(1-\rho_{kj}^2)} \leq \rho_{ij} \leq \rho_{ik}\rho_{kj} + \sqrt{(1-\rho_{ik}^2)(1-\rho_{kj}^2)}$.

3.2 Problem statement

Our goal is to obtain a small estimate $\mathfrak{R}$ that well approximates the full correlation matrix R without computing all pairwise correlations. Intuitively, the size of an estimate is the number of model parameters that need to be stored, and the quality is measured by the absolute deviation from the true correlation. Formally, let $\hat{\rho}(i, j \mid \mathfrak{R}) : \{1, ..., N\}^2 \longrightarrow [-1, 1]$ be an *estimator* for the correlation ρ_{ij} based on the representation $\mathfrak{R}$. The *loss* of an estimator is given by

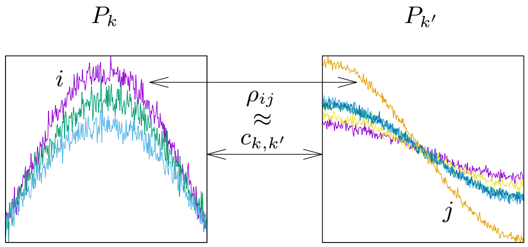

Fig. 2. Estimating pairwise time series correlations by inter-class correlations

the absolute deviation from the true correlation $\ell_{ij} = |\hat{\rho}(i, j \mid \mathfrak{R}) - \rho_{ij}|$. The traditional brute force estimator is the special case $\mathfrak{R} = R$ and $\hat{\rho}(i, j \mid \mathfrak{R}) = \rho_{ij}$. The brute force approach has $\frac{1}{2}N(N+1)$ model parameters and incurs a loss of zero. The other extreme is the special case $\mathfrak{R} = c \in [-1, 1]$ and $\hat{\rho}(i, j \mid \mathfrak{R}) = c$, which has only a single parameter to store, but potentially high loss. We aim at trade-offs between these two extremes. The general problem is thus to find a *low redundancy representation* $\mathfrak{R}$ with a small number of parameters and an estimator $\hat{\rho}(i, j \mid \mathfrak{R})$ that incurs a small loss.

We restrict ourselves to representations based on partitions of the dataset into classes of similar time series. The idea is illustrated in Fig. 2 for time series from two equivalence classes P_k and $P_{k'}$. All pairwise correlations between members of the two classes are redundant and can be collapsed to a single estimate for the inter-class correlation $c_{k,k'}$ with minor loss. Formally, we aim at representations of the form $\mathfrak{R} = (\mathcal{P}, C)$, where $\mathcal{P}$ is a partition of $\mathcal{X}$ into $K = |\mathcal{P}|$ equivalence classes and $C = \{c_{k,k'} \in [-1, 1] \mid 1 \leq k \leq k' \leq K\}$ is a set of inter-class correlations. The respective estimator is $\hat{\rho}(i, j \mid \mathcal{P}, C) = c_{k,k'}$ for $i \in P_k$ and j in $P_{k'}$. Such representations have $N + \frac{1}{2}K(K+1)$ parameters. The fewer classes K are necessary to capture all pairwise correlations with small loss, the lower the redundancy in the final estimate. We formalize our problem as the following approximation problem:

Problem 1. Given a collection of time series $\mathcal{X}$ and an error bound $\epsilon \geq 0$, find a partition $\mathcal{P}$ of $\mathcal{X}$ and a set of inter-class correlations C, such that the estimate $\mathfrak{R} = (\mathcal{P}, C)$ has a loss $\ell_{ij} = |c_{k,k'} - \rho_{ij}| \leq \epsilon$ for all $i \in P_k$ and $j \in P_{k'}$.

The challenge is to obtain such estimates with as few correlation computations as possible. In particular, without computing the full matrix R. A trivial solution for Problem 1 is the partition into N singleton classes $\mathcal{P} = \{\{X_1\}, ..., \{X_N\}\}$ such that the inter-class correlations are exactly the pairwise time series correlations. This solution collapses to the full correlation matrix R with zero loss but without reduction of redundancy or any computational efficiency improvements. In the following, we formally derive non-trivial approximations that guarantee a loss of at most ϵ with lower redundancy than R, and can be computed way more efficiently than the full matrix.

4 COREQ

The intuition behind our construction is that homogeneous equivalence classes with high *intra*-class correlations lend themselves to high-quality estimates for the *inter*-class correlations. Based on our formal analysis we propose the efficient greedy partitioning algorithm COREQ (CORrelation EQuivalence) and three estimators to obtain pairwise class correlations from the resulting partitions: an estimator with approximation guarantees and two powerful heuristics.

4.1 Approximations with quality guarantees

We start with the formal construction of a solution to Problem 1 with quality guarantees. The idea is to build homogeneous equivalence classes by a pivoting approach. Each class is identified with a unique pivot time series, and all other time series are assigned to classes such that the correlations to their respective pivot time series are at least $\alpha \in (0, 1]$. The parameter α controls the class homogeneity: the closer α to 1, the more homogeneous the equivalence classes, and the lower the estimation loss. Since we do not specify the number of classes K in advance, such partitions exist for any choice of α. The following theorem establishes how large α needs to be chosen to guarantee a loss of at most ϵ:

Theorem 2. *Let $\alpha \in (0, 1]$ and $\epsilon \geq 0$. Let $\mathcal{P} = \{P_k \mid k = 1, ..., K\}$ be a partition of $\mathcal{X}$ with associated pivot time series $X_{i_k} \in P_k$ such that $\forall X_i \in P_k : \rho_{i,i_k} \geq \alpha$. Furthermore, let the inter-class correlations C be the correlations between these pivot time series scaled by a correction factor that depends on α:*

$$c_{k,k'} = \frac{1}{2}\left(1 + \alpha^2\right)\rho_{i_k,i'_k}. \tag{1}$$

It holds that $\ell_{ij} \leq \epsilon$ for all $X_i, X_j \in \mathcal{X}$, if $\alpha \geq \sqrt{1 - \left(\frac{2\epsilon}{\sqrt{5}+2}\right)^2}$.

A proof based on the triangular bounds from Theorem 1 can be found in the Supplementary Material.[3] Section 4.2 provides an efficient greedy algorithm to compute such partitions. The scaling factor $\frac{1}{2}(1 + \alpha^2)$ in Equation 1 can be interpreted as the uncertainty about the representativeness of pivot correlations: the smaller α, the more heterogeneous the equivalence classes, and the less representative the pivots for their classes. Consequently, it is safer—in the general case—to estimate correlations close to zero instead of extremal values. Theorem 2 states that for any desired error bound ϵ we can find a (possibly) non-trivial solution $\mathfrak{R} = (\mathcal{P}, C)$ to Problem 1 that guarantees $\ell_{ij} < \epsilon$ for all pairs of time series. However, the quality guarantee is based on the worst-case bounds from Theorem 1 which do not make any assumptions on the distribution of correlations within a dataset. In particular, we do not assume that the time series cluster into homogeneous groups as motivated in Fig. 2 for many real-life time series collections. For any realistic choice of ϵ the theorem thus requires a threshold α

[3] available on the project website `https://hpi.de/mueller/coreq.html`

very close to 1 to guarantee the quality on any possible input dataset. For example, a loss $\ell_{ij} \leq 0.1$ can only be guaranteed for all pairs of time series on any input dataset if we set $\alpha \geq 0.9989$. The downside of choosing a value of α close to 1 is that we will most likely obtain the trivial solution with high redundancy and no computational efficiency improvements. As we see in Section 5, we can efficiently obtain estimates with low redundancy *and* low losses on *many real-life datasets* for much lower values of α.

4.2 A greedy estimation algorithm

We compute the pivot-based partitions formally defined in Theorem 2 as follows. We start by picking an arbitrary time series X_i from $\mathcal{X}$ as a pivot series and compute the correlations between X_i and all remaining time series. All time series with a correlation to X_i not smaller than α are stored in a new equivalence class P. The class P always contains X_i itself. All elements from P are removed from the original time series collection $\mathcal{X}$, and the procedure is repeated with a newly picked pivot series until all time series are processed. This procedure terminates with a partition as of Theorem 2 for any $\alpha \in (0, 1]$ with at most $\frac{1}{2}N(N+1)$ correlation computations. In the best case, if all correlations are larger than α, it terminates with only N correlation computations. Given such a partition, the question is how to best estimate the inter-class correlations C. We propose three alternatives to obtain a complete correlation estimate:

(i) **COREQ-P1**: scaled pivot correlations from Equation 1 in Theorem 2 which theoretically guarantee low errors on all datasets for $\alpha \longrightarrow 1$ but have a bias towards zero for smaller choices of α.
(ii) **COREQ-P2**: simplified estimate that uses unscaled pivot correlations $c_{k,k'} = \rho_{i_k, i'_k}$ to remove the bias for smaller choices of α.
(iii) **COREQ-A**: average estimate that samples a logarithmic number of correlations between pivot X_{i_k} and the class $P_{k'}$

$$c_{k,k'} = \frac{1}{\max\left(1, \lceil \log_2 N_{k'} \rceil\right)} \sum_{j'=1}^{\max(1, \lceil \log_2 N_{k'} \rceil)} \rho_{i_k, \mathrm{rand}(P_{k'})},$$

where $N_{k'} = |P_{k'}|$ and $\mathrm{rand}(P_{k'})$ returns a random time series from $P_{k'}$.

All of these estimates can be obtained from the correlations computed during class construction and do not require additional correlation computations. In COREQ-A we sample a logarithmic number of correlations to account for the heterogeneity in large equivalence classes. All three estimates converge to the pivot correlations for $\alpha \longrightarrow 1$ and differ only for $\alpha \ll 1$.

4.3 Formal relation to clustering algorithms

There is a clear relationship between our equivalence class-based correlation matrix approximations and the well-known optimization problem of time series clustering. We could relax the goal of strict approximation guarantees for all pairs of time series towards estimation with minimal *aggregated* loss. Let $X \in \mathbb{R}^{N \times T}$ be a matrix representation of $\mathcal{X}$ where all time series are standardized to have zero mean and unit variance over time. Furthermore, let $R = \frac{1}{T} X X^\top$ be the true correlation matrix, $Z = \{0,1\}^{N \times K}$ be an indicator matrix that encodes class memberships of a partition $\mathcal{P} = \{P_1, ..., P_K\}$, and $C \in [-1,1]^{K \times K}$ be a matrix of inter-class correlations. The error function $E = \|R - ZCZ^\top\|^2$ encodes the goal of finding an estimate $\mathfrak{R} = (\mathcal{P}, C)$ that well represents all correlations within R. We observe that this error function is a quadratic form of the sum of squared errors (SSE) that is used extensively for clustering, most prominently in K-Means. To see this relation, let $M \in \mathbb{R}^{K \times T}$ be the matrix of cluster centroids in K-Means. The sum of squared errors is defined as SSE $= \|X - ZM\|^2$. Using the pairwise centroid correlations as estimates for the inter-class correlations $C = \frac{1}{T} M M^\top$, we obtain $E = \|\frac{1}{T} X X^\top - Z \frac{1}{T} M M^\top Z\|^2$. Due to the structural similarity of E and SSE, we use K-Means clustering as a baseline in our experiments. However, to the best of our knowledge, there is no clustering algorithm that allows approximating correlations up to an error bound ϵ.

5 Empirical evaluation

Our empirical evaluation consists of two parts. In the first part, we extensively analyze the quality of the estimates obtained by COREQ in terms of average loss and model size on a large variety of datasets. In the second part, we compare the performance of COREQ against two state-of-the-art competitors and the K-Means baseline on a real-life dataset from the geoscientific domain. We implemented COREQ as a Python C module. All source codes necessary to reproduce our results are available on GitHub.[4] Additional information is provided on our project website.[5]

5.1 Experimental setup

Performance measures. The *average loss* for an estimate $\mathfrak{R}$ is given by $\bar{\ell} = \frac{1}{Z} \sum_{i=1}^{N} \sum_{j=i}^{N} \ell_{ij}$ with $Z = \frac{1}{2} N(N+1)$. The closer to 0, the better. The *model size* is given by the total number of model parameters that need to be stored by an algorithm, divided by the number of entries in the true correlation matrix. Model sizes close to 0 indicate a low redundancy, whereas values close to 1 indicate high redundancy. We also count the number of *correlation computations* necessary to obtain an estimate. All performance measures are averaged over ten independent runs to obtain stable results for each algorithm and dataset.

Data. To analyze the performance of COREQ over a large variety of time series collections, we run experiments on all 85 time series collections from the

[4] `https://github.com/KDD-OpenSource/coreq`

[5] `https://hpi.de/mueller/coreq.html`

well-known publicly available *UCR Time Series Classification Archive* [1]. For a real-life comparison with state-of-the-art algorithms, we use satellite image time series obtained from the NASA Terra MODIS satellite mission [3]. The dataset contains 236,197 *EVI time series* (Enhanced Vegetation Index) for South America, captured with a temporal resolution of 16 days between 2000 and 2015 (length 368). The EVI is computed from multi-spectral satellite images and captures the level of greenness at a given point in time as a proxy for vegetation cover.

Competitors. As a baseline, we perform one iteration of K-Means clustering with a fixed K to obtain a partition of the dataset and use the pairwise centroid correlations as class correlations. Using more iterations is infeasible since it drastically increases the number of correlation computations. We also compare against two state-of-the-art algorithms proposed by Mueen et al. [11] to compute an Approximate Threshold Correlation Matrix (ApproxThresh) and a Threshold Boolean Correlation Matrix (ThreshBoolean). ApproxThresh approximates (up to an error ϵ) all correlations larger than a threshold τ by exploiting a Discrete Fourier Transform-based early-abortion criterion for individual correlation computations; all correlations below τ are set to 0 without error guarantee. ApproxThresh is designed to reduce the number of operations for individual correlation computations. To compare the total costs of correlation estimation with our approach, we scale the number of correlation computations with the speedup factor per correlation computation. ThreshBoolean uses a dynamic programming-based pruning strategy to reduce the number of pairwise comparisons. It estimates all (absolute) correlations above τ as ±1 and all other correlations as 0, without any quality guarantees.

5.2 Quality of estimates

We first analyze the performance of COREQ in terms of average loss and resulting model size on all 85 UCR datasets for various values of α. Fig. 3 visualizes the distribution of average loss over all UCR datasets as boxplots along with the mean model size. We provide separate boxplots for COREQ-P1/P2 and COREQ-A; mean model sizes are identical. As expected, increasing α pushes the average loss on all datasets towards zero since equivalence classes become more homogeneous. At the same time, it increases the model size. COREQ-A outperforms COREQ-P1/P2 over the full parameter space, with the margin of improvement largest for low values of α. Lower values of α typically come with larger and more heterogeneous equivalence classes, such that the pivot correlations are not representative. The scaled pivot correlations from COREQ-P1 perform worse than the unscaled variant COREQ-P2 on many datasets. The datasets where COREQ-P2 outperforms COREQ-P1 contain time series that are all very strongly correlated. In these cases, the theoretically justified bias towards zero correlations is harmful. With $\alpha = 0.9$, all three estimation variants achieve high-quality estimates with average losses below 0.1 and a mean model size below 0.35.

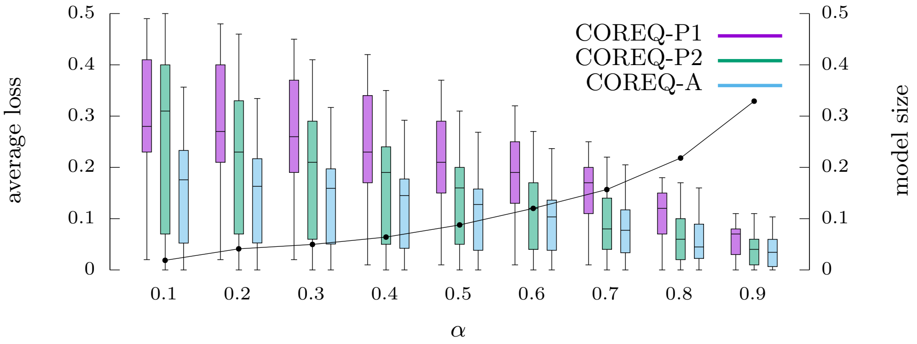

Fig. 3. Distribution of average loss (boxplots) and mean model size (line) across all UCR datasets for $\alpha \in [0, 1]$.

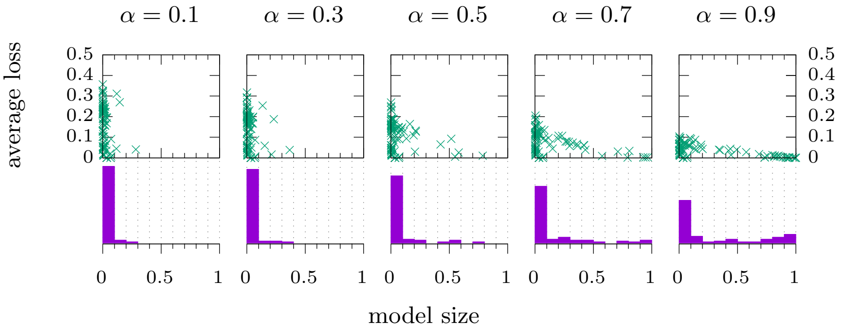

Fig. 4. Average loss against model size achieved by COREQ-A on all UCR datasets for $\alpha \in [0, 1]$, along with histograms over model size.

Detailed scatter plots of the results of COREQ-A can be found in Fig. 4. Each point in a plot shows the model size and average loss achieved on a single dataset. The histograms below show the corresponding distributions of model sizes. We observe that even for $\alpha = 0.9$ the large majority of datasets can well be estimated with model sizes below 0.1. Only a few datasets appear on the far right with model sizes close to 1. Manual inspection of these datasets revealed that they contain purely uncorrelated time series or ambiguous group structures. These instances cannot be estimated more efficiently with our approach. COREQ provides low redundancy estimates with low average losses on all datasets with strong group structures.

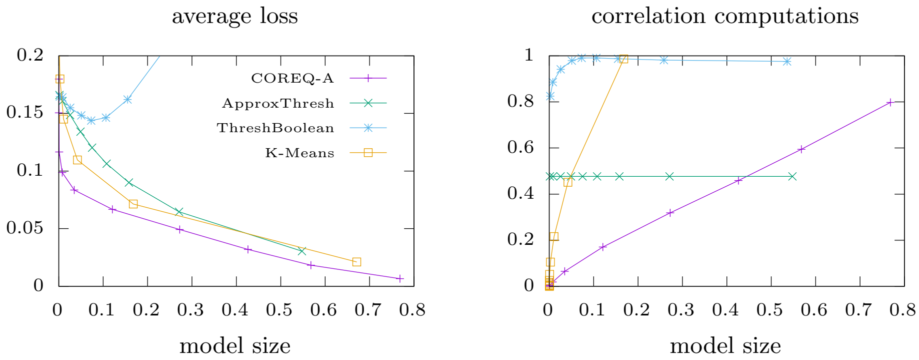

Fig. 5. Performance on EVI data over the full parameter space of each algorithm.

5.3 Comparison with existing methods

We now compare COREQ-A with the state-of-the-art algorithms introduced by Mueen et al. [11] and our K-Means baseline. We address two questions in our analysis: (1) How much loss does an algorithm incur at a given model size? (2) How many correlation computations are necessary to obtain an estimate with that model size? All algorithms in our evaluation depend on different input parameters that affect the estimation performance. These input parameters directly control the model size: the larger α in COREQ and K in K-Means, the more pairwise class correlations have to be estimated and stored, while a smaller threshold τ in ApproxThresh and ThreshBoolean means that more pairwise time series correlations have to be stored. To compare these approaches in a meaningful and fair way, we run all algorithms over a wide range of parameterizations ($\alpha \in \{0.1, 0.2, ..., 0.9\}$, $K \in \{1, 2, 4, ..., 8192\}$, $\tau \in \{0.9, 0.8, ..., 0.1\}$) and use the resulting model size as the unified scale. The error bound for ApproxThresh is set to $\epsilon = 0.05$. We use the EVI dataset[6] as a real-life example from the geoscientific domain.[7]

To answer the first question, Figure 5 (left) shows the average loss of the resulting correlation estimates against the model size. If a curve is close to the origin, it means that small estimates obtained with that algorithm capture most of the information from the correlation matrix. COREQ-A clearly outperforms K-Means, ApproxThresh and ThreshBoolean over the full parameter space: our algorithm has lower losses at the same model sizes. The improvement is largest for very small estimates. The ThreshBoolean approach behaves un-

[6] subsamples of 10,000 time series for COREQ/K-Means/ApproxThresh and 1,000 time series for ThreshBoolean due to performance reasons

[7] We also ran experiments on the chlorine concentration data used in the original publication by Mueen et al. [11]; the results are consistent with the results presented in this paper and reported for completeness in the Supplementary Material.

usually: since it can only estimate correlations as either 0 or ±1, lowering the threshold τ means that more and more weak correlations are stored and estimated as ±1. The algorithm is not designed to capture weak correlations accurately. Overall, COREQ-A provides the highest quality estimates for the full correlation structure, with improvements being largest for very small estimates.

For the second question, Figure 5 (right) shows the number of correlation computations required to obtain the final estimates (normalized by the total number of pairs) against model size. Our approach scales linearly with the model size: the number of correlations that we compute is roughly the same as the number of model parameters we output. The K-Means baseline performs worst, even though we run only one iteration. More iterations or more sophisticated clustering algorithms could improve the quality of the estimates, but come with an even higher computational cost. ApproxThresh requires a constant number of correlation computations for all threshold values τ. The early abortion criterion yields an average speed-up of only 2 per correlation computation, meaning that the EVI time series are uncooperative [2]. ApproxThresh outperforms our approach in terms of correlation computations only in the large model size region on the right. The pruning strategy employed in ThreshBoolean is effective at the far left of the plot, where the threshold τ is close to 1. For lower threshold values almost all pairwise correlations are computed. COREQ is the fastest algorithm in terms of correlation computations in the small model size region of the parameter space—with a large margin to all competitors. In the same region, we obtain the lowest average loss values.

6 Conclusion and future work

We provide a novel way to estimate correlation matrices for large time series collections that exploits redundancies in the input data to drastically reduce the number of parameters to estimate. We show that the partitions we obtain for estimation have theoretical approximation guarantees, allow for very small high-quality estimates on a large variety of real-life datasets, and outperform state-of-the-art approaches. There is still need for a robust way to select the parameter α optimally for any input dataset as to obtain the best trade-off between model size and average loss. Algorithmically, dynamically adapting α during estimation to process datasets with weak and strong group structures could be beneficial. We defer this challenge to future work. Furthermore, combining our estimation approach with a probabilistic model for time series collections would allow us to devise more concise probabilistic error guarantees on top of the worst-case bounds we used in Theorem 2. At last, an extension of COREQ for streaming time series would allow efficient monitoring of correlations for anomaly detection.

References

1. Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G.: The UCR Time Series Classification Archive (July 2015), http://www.cs.ucr.edu/~eamonn/time_series_data/

2. Cole, R., Shasha, D., Zhao, X.: Fast window correlations over uncooperative time series. In: KDD (2005)
3. Didan, K.: MOD13C1 MODIS/Terra Vegetation Indices 16-Day L3 Global 0.05Deg CMG V006 (2015). https://doi.org/10.5067/modis/mod13c1.006
4. Guha, S., Gunopulos, D., Koudas, N.: Corrrelating Synchronous And Asynchronous Data Streams. In: KDD (2003)
5. Guo, T., Sathe, S., Aberer, K.: Fast Distributed Correlation Discovery Over Streaming Time-Series Data. In: CIKM (2015)
6. Keller, F., Müller, E., Böhm, K.: Estimating Mutual Information on Data Streams. In: SSDBM (2015)
7. Kulis, B., Sustik, M.A., Dhillon, I.S.: Low-Rank Kernel Learning with Bregman Matrix Divergences. Journal of Machine Learning Research **10** (2009)
8. Langford, E., Schwertman, N., Owens, M.: Is the Property of Being Positively Correlated Transitive? The American Statistician **55**(4) (2001)
9. Ledoit, O., Wolf, M.: Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance **10**(5) (2003)
10. Liao, T.W.: Clustering of time series data: A survey. Pattern Recognition **38**(11) (2005)
11. Mueen, A., Nath, S., Liu, J.: Fast approximate correlation for massive time-series data. In: SIGMOD (2010)
12. Papadimitriou, S., Sun, J., Faloutsos, C.: Streaming Pattern Discovery in Multiple Time-Series. In: VLDB (2005)
13. Papadimitriou, S., Sun, J., Yu, P.S.: Local Correlation Tracking in Time Series. In: ICDM (2006)
14. Paparrizos, J., Gravano, L.: k-Shape: Efficient and Accurate Clustering of Time Series. In: SIGMOD (2015)
15. Sakurai, Y., Papadimitriou, S., Faloutsos, C.: BRAID: Stream Mining through Group Lag Correlations. In: SIGMOD (2005)
16. Sathe, S., Aberer, K.: AFFINITY: Efficiently querying statistical measures on time-series data. In: ICDE (2013)
17. Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Ramsey, J.D., Woolrich, M.W.: Network modelling methods for FMRI. NeuroImage **54**(2) (2011)
18. Ulanova, L., Begum, N., Keogh, E.: Scalable Clustering of Time Series with U-Shapelets. In: SIAM SDM (2015)
19. Wied, D., Galeano, P.: Monitoring correlation change in a sequence of random variables. Journal of Statistical Planning and Inference **143**(1) (2013)
20. Wiedermann, M., Radebach, A., Donges, J.F., Kurths, J., Donner, R.V.: A climate network-based index to discriminate different types of El Niño and La Niña. Geophysical Research Letters **43**(13) (2016)
21. Wu, D., Ke, Y., Yu, J.X., Yu, P.S., Chen, L.: Detecting Leaders from Correlated Time Series. DASFAA (2010)
22. Xie, Q., Shang, S., Yuan, B., Pang, C., Zhang, X.: Local Correlation Detection with Linearity Enhancement in Streaming Data. In: CIKM (2013)
23. Xiong, H., Shekhar, S., Tan, P.n., Kumar, V.: TAPER: A Two-Step Approach for All-Strong-Pairs Correlation Query in Large Databases. TKDE **18**(4) (2006)
24. Zhang, Z., Wu, L.: Optimal low-rank approximation to a correlation matrix. Linear Algebra and its Applications **364** (2003)
25. Zhu, Y., Shasha, D.: StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time. In: VLDB (2002)