Cluster Kernel For Learning Similarities Between Symmetric Positive Definite Matrix Time Series

Sara Akodad, Lionel Bombrun, Yannick Berthoumieu, Christian Germain

▶ To cite this version:

HAL Id: hal-02959479

https://hal.science/hal-02959479v1

Submitted on 6 Oct 2020

CLUSTER KERNEL FOR LEARNING SIMILARITIES BETWEEN SYMMETRIC POSITIVE DEFINITE MATRIX TIME SERIES

Sara Akodad, Lionel Bombrun, Yannick Berthoumieu and Christian Germain

Université de Bordeaux, CNRS, IMS, UMR 5218, Groupe Signal et Image, F-33405 Talence, France
e-mail: {sara.akodad, lionel.bombrun, yannick.berthoumieu, christian.germain}@ims-bordeaux.fr

ABSTRACT

The launch of the last generation of Earth observation satellites has yield to a great improvement in the capabilities of acquiring Earth surface images, providing series of multitemporal images. To process these time series images, many machine learning algorithms have been proposed in the literature such as warping based methods and recurrent neural networks (LSTM, . . .). Recently, based on an ensemble learning approach, the time series cluster kernel (TCK) has been proposed and has shown competitive results compared to the state-of-the-art. Unfortunately, it does not model the spectral/spatial dependencies. To overcome this problem, this paper aims at extending the TCK approach by modeling the time series of second-order statistical features (SO-TCK). Experimental results are conducted on different benchmark datasets, and land cover classification with remote sensing satellite time series over the Reunion Island.

Index Terms— Multivariate times series, classification, second-order statistics, log-Euclidean, remote sensing.

1. INTRODUCTION

A time series is a sequence of data points which has been acquired during an ordered time segment. Many real-world pattern recognition tasks deal with time-series analysis [1] such as biomedical signals (e.g. EEG and ECG), financial data (e.g. stock market and currency exchange rates), industrial devices (e.g. gas sensors and laser excitation), biometrics (e.g. voice, signature and gesture), video processing, music mining, forecasting, etc. Recently, the launch of the last generation of Earth observation satellites such as Sentinel-1 and 2 has yield more recurrent acquisition of Earth surface images. These sensors allow the acquisition of multivariate time series such as spectral surface reflectance in several wavelengths. The availability of these multivariate time series has raised the interest of the image processing community to develop novel machine learning strategies for supervised classification. In particular, many deep learning approaches have recently been proposed [2]. They include convolutional neural networks (CNN) [3] and deep recurrent neural networks (RNN) such as long short-term memory (LSTM) [4, 5] and gated recurrent units (GRU) [6]. In the following, these ap-

proaches will not be considered since the training process involves large data for avoiding overfitting and generalization purpose, while the aim of this paper is to work with relatively small datasets. Other non deep learning approaches can be decomposed into three categories: distance-based time series classifiers, feature-based time series classifiers and ensemble classifiers. Distance-based strategies rely on a point-to-point distance between time series which is then fed in a conventional classifier such as a k nearest neighbor or SVM. In this family, the most popular approach is certainly the dynamic time warping (DTW) which measures the similarity between two time series by aligning them [7, 8]. It has the ability to match time series that are distorted and shifted along the temporal axis. Inspired by the principle of DTW, some closely related approaches have been proposed such as derivative DTW (DDTW) [9] or weighted DTW (WDTW) [10]. Kernel based methods have also been introduced such as the global alignment kernel (GAK) [11] for the family of DTW distances. More recently, many feature-based methods have been proposed. It consists in extracting features such as wavelet coefficients or 1D SIFT descriptors from multivariate time series before the classification step. Among these feature-based methods, codebook based representations have raised an interest [12]. For example, the bag-of-words (BoW) model has been used in [13] to obtain an histogram representation of the time series by encoding SIFT features in a codebook. The third family concerns ensemble based classifier systems. The basic idea relies on the combination of multiple classifiers in order to obtain more accurate and robust decisions. Once again, many approaches can be considered. For example, a random forest classifier trains a single base classifier (*i.e.* decision tree) on different subsets of training data (sample, attributes and/or temporal subsets). Another strategy consists in using different classifiers on the same dataset such as in the collective of transformation-based ensembles (COTE) [14] and its extension based on a hierarchical vote (HIVE-COTE) [15] where 35 and 37 standalone classifiers are respectively considered. Even if this later has demonstrated successful results and is considered as the reference for time series classification [16], it suffers from an high computation cost since each classifier should be trained on the whole dataset. Since then, in order to get benefit of the ad-

vantages of kernel methods, codebook based representations and ensemble learning strategies, Mikalsen *et al.* have introduced the time series cluster kernel (TCK) method in [17] which has demonstrated competitive results for times series classification.

At the same time, second-order features have shown a great interest for many image processing applications including person re-identification, texture recognition, material categorization or EEG classification in brain-computer interfaces to cite a few of them [18, 19, 20]. For example, the use of covariance matrices has demonstrated to be successful in [21] for remote sensing scene classification. This strategy, Hybrid LE FV, achieves better results than its analog Hybrid FV [22] which exploits only first-order features. Since then, the main contribution of this paper is to present how TCK proposed in [17] for multivariate time series can be extended to work with second-order feature time series. The paper is structured as follows. Section 2 presents and discusses the principle of TCK. Then, Section 3 introduces the proposed extension of TCK for the modeling of the time series of second-order statistics (SO-TCK). An application on different datasets is next presented in Section 4. Finally, Section 5 concludes this paper and provides some perspectives of this work.

2. TIME SERIES CLUSTER KERNEL (TCK)

TCK has recently been introduced in [17] for the classification of multivariate time series. It exploits the power of kernel methods, codebook based representations and ensemble learning strategies. The global principle is explained in Fig. 1. The main idea behind this method is to compute a positive semi-definite similarity measure (*i.e.* a kernel) between two multivariate times series (MTS). For that, a GMM model is first trained on the training set, which is next used to encode each multivariate time series. Moreover, to ensure robustness, an ensemble learning strategy is considered. The next subsection presents the main steps during training and testing.

2.1. Training phase

An MTS $\mathbf{X}$ is represented as a matrix of dimension $V \times T$ where V is the number of attributes and T is the time length. It is a finite sequence of V univariate time series, *i.e.* $\mathbf{X} = \{\mathbf{x}_v \in \mathbb{R}^T\}$ for $v = 1, \ldots, V$. During this stage, N MTS are considered for training, and $\mathbf{X}^{(n)}$ represents the nth MTS sample. To cope with missing data, a second MTS is considered. $\mathbf{R}^{(n)}$ is a binary MTS defined by $\mathbf{r}_v^{(n)}(t) = 0$ if the value $\mathbf{x}_v^{(n)}(t)$ is missing and $\mathbf{r}_v^{(n)}(t) = 1$ otherwise.

As explained before, TCK is based on an ensemble learning approach. Practically, Q subsets are considered, where each subset is a subsample of N_q data, V_q attributes and T_q time segments extracted from the training set. Then, for each subset, a codebook is created by learning a GMM model which is next used to encode each MTS.

2.1.1. Codebook: GMM modeling

For the qth subset, the training set is composed of N_q MTS $\mathbf{X}_q^{(n)}$ of dimension $V_q \times T_q$ with their associated binary MTS $\mathbf{R}_q^{(n)}$. A GMM model with G components is considered to learn a codebook, where its probability density function for the incompletely observed MTS $(\mathbf{X}_q^{(n)}, \mathbf{R}_q^{(n)})$ is given by:

$$p(\mathbf{X}_q^{(n)}|\mathbf{R}_q^{(n)}, \Theta) = \sum_{g=1}^{G} \omega_g \prod_{v=1}^{V_q} \prod_{t=1}^{T_q} \mathcal{N}(\mathbf{x}_v^{(n)}(t)|\mu_{gv}(t), \sigma_{gv})^{\mathbf{r}_v^{(n)}(t)},$$

(1)

where $\Theta = \{\omega_g, \mu_{gv}, \sigma_{gv}\}$ for $g = 1, \ldots, G$. MTS are assumed to have time-dependent means, *i.e.* $\{\mu_{gv} \in \mathbb{R}^{T_q}\}$ for each attribute v. To enforce regularity, a kernel-based Gaussian prior is defined for the mean. In addition, the covariance matrix for each Gaussian component is assumed to be diagonal and constant, that is $\Sigma_g = \text{diag}\{\sigma_{g1}^2, \ldots, \sigma_{gV_q}^2\}$ and σ_{gv}^2 is the variance of attribute v for data belonging to the gth cluster. To estimate the GMM parameters Θ, a MAP-EM algorithm is considered. For more details, the interested reader is referred to Algorithm 1 in [17]. The set composed by the estimated GMM parameters represents the codebook. To ensure even more robustness, the number of GMM components G and the initialization of the MAP-EM algorithm are randomly selected for each subset.

2.1.2. Coding

The encoding of the nth MTS $\mathbf{X}_q^{(n)}$ for subset q is carried out by computing the posterior assignment $\Pi_q^{(n)}$, obtained by:

$$\Pi_q^{(n)}(g) = \frac{\omega_g \prod_{v=1}^{V_q} \prod_{t=1}^{T_q} \mathcal{N}(\mathbf{x}_v^{(n)}(t)|\mu_{gv}(t), \sigma_{gv})^{\mathbf{r}_v^{(n)}(t)}}{\sum_{g=1}^{G} \omega_g \prod_{v=1}^{V_q} \prod_{t=1}^{T_q} \mathcal{N}(\mathbf{x}_v^{(n)}(t)|\mu_{gv}(t), \sigma_{gv})^{\mathbf{r}_v^{(n)}(t)}}.$$

(2)

In the following, vector $\Pi_q^{(n)}$ containing the G posterior probabilities $\Pi_q^{(n)}(g)$ is considered to define the feature map used in the kernel.

2.2. Testing step

To compute a similarity measure between a training MTS $\mathbf{X}^{(n)}$ and a testing MTS $\mathbf{X}^{(m)}$, a kernel based approach is considered. It is built on the basis of an inner product between two posterior distributions to form a linear kernel in the space of posterior distributions as:

$$K\left(\mathbf{X}^{(n)}, \mathbf{X}^{(m)}\right) = K_{nm} = \sum_{q=1}^{Q} \Pi_q^{(n)T} \Pi_q^{(m)}, \qquad (3)$$

where $\Pi_q^{(n)}$ and $\Pi^{(m)}(q)$ are respectively the vector of posterior probabilities for the training and testing MTS obtained with (2). In the end, a nearest neighbor classifier is used with the induced distance d between $\mathbf{X}^{(n)}$ and $\mathbf{X}^{(m)}$
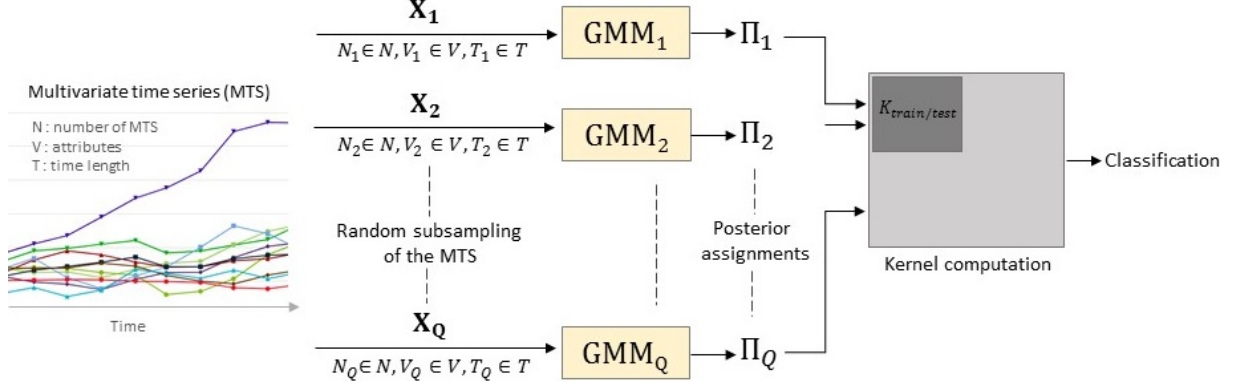
Fig. 1. General principle of the time series cluster kernel (TCK) [17].

given by:

$$d^2\left(\mathbf{X}^{(n)}, \mathbf{X}^{(m)}\right) = K\left(\mathbf{X}^{(n)}, \mathbf{X}^{(n)}\right) - 2K\left(\mathbf{X}^{(n)}, \mathbf{X}^{(m)}\right)$$
$$+ K\left(\mathbf{X}^{(m)}, \mathbf{X}^{(m)}\right). \tag{4}$$

As observed in (3) and (4), two MTS will be similar if their posterior probability vectors $\Pi_q^{(n)}$ and $\Pi_q^{(m)}$ are similar for each subset q. When a low number G of GMM components are considered, the kernel measures the similarity at large scale. While for large G, it focuses more locally. Hence, varying the number of components in the GMM model allows to capture both local and global structures in the data.

3. TIME SERIES CLUSTER KERNEL FOR SECOND-ORDER STATISTICAL FEATURES (SO-TCK)

In the literature, second-order features, in particular covariance matrices features, have proved to play an important role in different tasks related to visual recognition process [23]. Compared to first-order feature based classification algorithms, many authors have shown the interest of exploiting second-order statistics such as covariance matrix attributes [18, 19, 20]. This kind of data has a particular structure, they are symmetric positive-definite (SPD) matrices. This section's focus is twofold. First, we present how second-order features are extracted from a MTS to form SPD matrix time series (SPD-MTS). Secondly, we introduce the log-Euclidean representation of SPD-MTS in order to exploit them in TCK, yielding to the proposed SO-TCK method.

3.1. SPD matrix time series (SPD-MTS)

In this work, each considered time series is represented by a set of time-dependent second-order features which constitutes a SPD matrix time series (SPD-MTS) as shown in Fig. 2. Those latter are computed on a sliding temporal window of dimension Δt. Hence, for the qth subset, we associate for each MTS of dimension $V_q \times T_q$ a SPD-MTS of dimension $V_q \times V_q \times (T_q - \Delta t + 1)$.
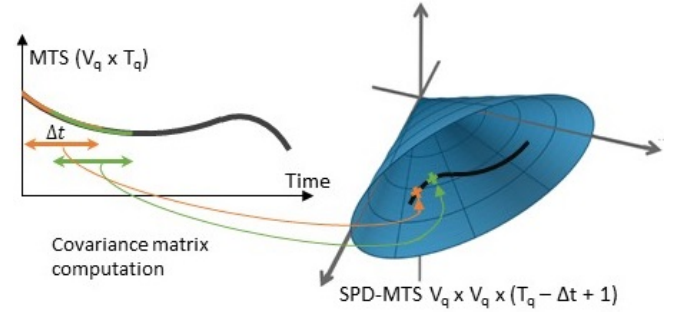

Fig. 2. Illustration of SPD-MTS computation.

In order to improve the feature representation and enhance classification performance, a full local Gaussian descriptor can be exploited. In this model, the local mean vector μ is jointly exploited with the $V_q \times V_q$ covariance matrix $\mathbf{M}$ which gives the augmented SPD matrix of dimension $(V_q + 1) \times (V_q + 1)$ as proposed in [24]:

$$\mathbf{M}_{augmented} = |\mathbf{M}|^{-\frac{1}{V_q+1}} \begin{bmatrix} \mathbf{M} + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix}. \tag{5}$$

3.2. Log-Euclidean representation of SPD-MTS

In order to adapt TCK to work with SPD-MTS, the geometry of the space $\mathcal{P}_d$ of $d \times d$ symmetric and positive definite (SPD) matrices should be considered. In case of using covariance matrices, $d = V_q$, while for the augmented SPD matrix: $d = V_q + 1$. As observed in Fig. 2, SPD matrices lie in a convex cone which is a Riemannian manifold. Tools developed in the context of Euclidean geometry are hence not adapted to manipulate these data points. A Riemannian metric is better suited such as the log-Euclidean (LE) one [25]. It consists in projecting the set of SPD matrices on a tangent space defined at a reference point, classically considered at the identity matrix. After being projected on the tangent space, tools of

the Euclidean geometry can used such as the MAP-EM algorithm defined in Section 2.1.1 to estimate GMM parameters and the Kernel method presented in Section 2.2. Practically, each SPD matrix $\mathbf{M}$ is mapped on the tangent space by applying the following operations:

$$\mathbf{m} = \text{Vec}(\text{logm}(\mathbf{M})), \qquad (6)$$

where logm() is the matrix logarithm and Vec() the vectorization operator:

$$\mathbf{x} = \text{Vec}(\mathbf{X}) = \left[X_{11}, \sqrt{2}X_{12}, \dots, \sqrt{2}X_{1d}, X_{22}, \sqrt{2}X_{23}, \dots, X_{dd} \right]. \quad (7)$$

To summarize, these operations lead to a transformation from a SPD matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ to a vector $\mathbf{m} \in \mathbb{R}^{\frac{d(d+1)}{2}}$.

4. EXPERIMENTS

In this section, we illustrate the potential of the proposed approach to multivariate time series classification on four benchmark datasets from UCI machine learning repository [26] and one on a remote sensing (RS) application. For this latter, the goal is to predict 9 land cover classes from a set composed by a series of $T = 23$ Landsat 8 images above the Reunion Island. It contains $V = 10$ attributes (7 surface reflectances and 3 vegetation/water indices) for each pixel at each timestamp. Table 1 gives the main characteristics of these datasets. In the following, performance are measured in terms of mean overall accuracy evaluated over 5 runs.

Datasets	V	T	Train	Test	Classes
Libras	2	23	180	180	15
Natops	3	51	180	180	6
Char.Traj.	3	23	300	2 558	20
RacketSports	6	30	151	152	4
RS	10	23	900	900	9

Table 1. Presentation of the five time series datasets.

Fig. 3 draws the evolution of the overall accuracy on the Libras dataset as a function of the temporal support Δt for the proposed SO-TCK approach. Three kind of second-order descriptors are considered: the covariance matrix (in red), the second order moment ($\mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right]$, in blue) and the full local Gaussian descriptor (Augmented SPD matrix in green). As observed, best results are obtained for this latter. In the following, SO-TCK will refer to the classification performance obtained with this full local Gaussian descriptor. Note also that the choice of Δt reflects a trade-off. It should be small enough to preserve a good temporal localisation, and at the same time sufficiently large for estimation purpose.

Secondly, a classification comparison is conducted over five datasets. Two state-of-the-art approaches are employed, namely DTW [7] and TCK [17]. We have adopted the same experimental setup as the one used in [17]. For reproducibility purpose, TCK has been launched with the authors Matlab implementation. As observed in Table 2, the best results are
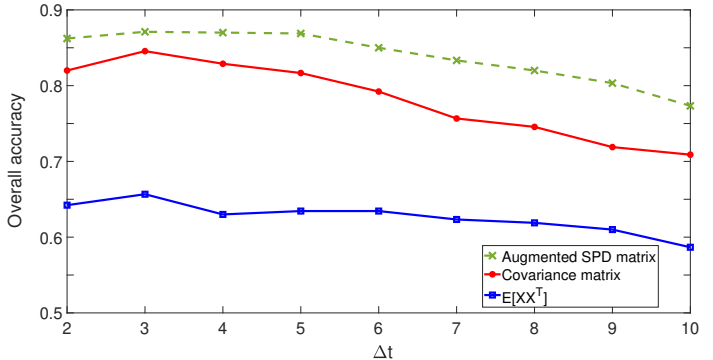

Fig. 3. Classification comparison between three considered second-order statistics for SPD-MTS computation ($V_q = 2$).

obtained for the proposed SO-TCK approach with a gain of 2 to 15 % compared to DTW and TCK, hence illustrating its potential for various application on time series classification.

Datasets	DTW [7]	TCK [17]	SO-TCK
Libras	85.6	72.6	**87.1**
Natops	57.2	61.4	**71.3**
Char. Traj	90.5	91.7	**93.9**
RacketSports	82.6	81.5	**85.2**
RS	64.7	63.7	**74.5**

Table 2. Comparison of classification performance on five benchmark datasets.

5. CONCLUSION

Inspired by the time series cluster kernel (TCK) and the potential of second-order statistical descriptors for many classification tasks, this paper has extended the formalism of TCK to second-order features: SPD matrix time series (SPD-MTS). For that, the log-Euclidean metric has been considered to represent a SPD-MTS as a multivariate time series where the principle of TCK can be employed. Experimental results on benchmark datasets and land cover classification with remote sensing data have shown the potential of the proposed method compared to state-of-the-art times series classification algorithm. Future works will include the evaluation of the proposed method for forest health monitoring using Sentinel-2 time series images.

6. ACKNOWLEDGMENT

7. REFERENCES

[1] G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. USA: Holden-Day, Inc., 1990.

[2] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Mining and Knowledge Discovery*, vol. 33, no. 4, pp. 917–963, Jul 2019. [Online]. Available: https://doi.org/10.1007/s10618-019-00619-1

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[5] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1685–1689, Oct 2017.

[6] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *CoRR*, vol. abs/1409.1259, 2014. [Online]. Available: http://arxiv.org/abs/1409.1259

[7] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, February 1978.

[8] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, ser. AAAIWS'94. AAAI Press, 1994, p. 359–370.

[9] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *In SIAM International Conference on Data Mining*, 2001.

[10] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," *Pattern Recognition*, vol. 44, no. 9, pp. 2231 – 2240, 2011, computer Analysis of Images and Patterns.

[11] M. Cuturi, "Fast global alignment kernels," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11. Madison, WI, USA: Omnipress, 2011, p. 929–936.

[12] J. Wang, P. Liu, M. F. She, S. Nahavandi, and A. Kouzani, "Bag-of-words representation for biomedical time series classification," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 634 – 644, 2013.

[13] A. Bailly, S. Malinowski, R. Tavenard, L. Chapel, and T. Guyet, "Dense bag-of-temporal-sift-words for time series classification," in *Advanced Analysis and Learning on Temporal Data*, A. Douzal-Chouakria, J. A. Vilar, and P.-F. Marteau, Eds. Cham: Springer International Publishing, 2016, pp. 17–30.

[14] A. Bagnall, J. Lines, J. Hills, and A. Bostrom, "Time-series classification with cote: The collective of transformation-based ensembles," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, May 2016, pp. 1548–1549.

[15] J. Lines, S. Taylor, and A. Bagnall, "Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Dec 2016, pp. 1041–1046.

[16] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery*, vol. 31, no. 3, pp. 606–660, May 2017. [Online]. Available: https://doi.org/10.1007/s10618-016-0483-9

[17] K. O. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, and R. Jenssen, "Time series cluster kernel for learning similarities between multivariate time series with missing data," *Pattern Recognition*, vol. 76, pp. 569 – 581, 2018.

[18] M. Faraki, M. T. Harandi, and F. Porikli, "More about VLAD: A leap from Euclidean to Riemannian manifolds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4951–4960.

[19] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications," *NeuroComputing*, vol. 112, pp. 172–178, 2013.

[20] S. Said, L. Bombrun, and Y. Berthoumieu, "Texture classification using Rao's distance on the space of covariance matrices," in *Geometric Science of Information*, vol. 9389, Oct 2015, pp. 371–378.

[21] S. Akodad, L. Bombrun, C. Yaacoub, Y. Berthoumieu, and C. Germain, "Image classification based on log-Euclidean Fisher vectors for covariance matrix descriptors," in *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Xi-an, China, Nov. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01930156

[22] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5653–5665, Oct 2017.

[23] B. Julesz, B, E. N. Gilbert, L. A. Shepp, and H. L. Frisch, "Perception," *Inability of humans to discriminate between visual textures that agree in second-order statistics-revisited*, vol. 2, no. 4, pp. 391–405, 1973.

[24] M. Lovric, M. Min-Oo, and E. A. Ruh, "Multivariate normal distributions parametrized as a riemannian symmetric space," *Journal of Multivariate Analysis*, vol. 74, no. 1, pp. 36 – 48, 2000.

[25] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-Euclidean metrics for fast and simple calculus on diffusion tensors," in *Magnetic Resonance in Medicine*, vol. 56, no. 2, Aug 2006, pp. 411–421.

[26] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml