Supplementary Material to
Kernelized Covariance for Action Recognition

Jacopo Cavazza*†, Andrea Zunino*†, Marco San Biagio* and Vittorio Murino*‡
* Pattern Analysis & Computer Vision – Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genova, Italy
† Università degli Studi di Genova – Dipartimento di Ingegneria Navale, Elettrica, Elettronica e delle Telecomunicazioni, Via All'Opera Pia, 11A, 16145, Genova, Italy
‡ Università di Verona – Dipartimento di Informatica, Strada le Grazie 15, 37134, Verona, Italy
{jacopo.cavazza,andrea.zunino,marco.sanbiagio,vittorio.murino}@iit.it

In this supplementary material, we present an extended exposition of the mathematical framework that supports the proposed kernelization pipeline for the sampling covariance estimator $\widehat{\mathbb{S}}$.

Let n the total number of joints of the considered motion capture (MoCap) system and let $\mathbf{x}_i(t) = [x_i(t), y_i(t), z_i(t)]^\top$ the recorded x, y, z coordinates of the i-th joint at time t, for every $i = 1, \dots, n$ and $t = 1, \dots, T$. Globally, for every $t = 1, \dots, T$,

$$\mathbf{x}(t) = [\mathbf{x}_1(t), \mathbf{x}_2(t) \dots, \mathbf{x}_n(t)]^\top = [x_1(t), y_1(t), z_1(t), x_2(t), y_2(t), z_2(t), \dots, x_n(t), y_n(t), z_n(t)]^\top$$

is the $3n$ column vector stacking all the temporal acquisitions of the n joints at time t.

Since resulting from an acquisition process, $\mathbf{x}(t)$ is certainly affected by a degree of uncertainty related, for instance, to the level of noisy corruption which perturbs an arbitrary coordinate of a generic joint. Thus, it is natural to think about $\mathbf{x}(t)$ as a random vector in $\mathbb{R}^{3n}$ and therefore assume that such level of uncertainty is modelled by a stationary probability distribution π from which $\mathbf{x}(1), \dots, \mathbf{x}(T)$ are independently sampled. Then, as a classical tool in probability theory and statistics, the *covariance matrix* $\mathbb{S}$ (also known as *dispersion matrix* or *variance-covariance matrix*) is used to measure how any pair of joint coordinates mutually change in time. Precisely, if we assume that π has finite second momentum, so that the integrals $\int_{t=0}^{\infty} \|\mathbf{x}(t)\| \pi(\mathbf{x}(t)) dt$ and $\int_{t=0}^{\infty} \|\mathbf{x}(t)\|^2 \pi(\mathbf{x}(t)) dt$ are both finite, then $\mathbb{S}(\pi)$ is the $3n \times 3n$ matrix defined as

$$\begin{aligned}\mathbb{S}(\pi) &= \int_{t=0}^{\infty} \left(\mathbf{x}(t) - \int_{s=0}^{\infty} \mathbf{x}(s)\pi(\mathbf{x}(s))ds \right) \left(\mathbf{x}(t) - \int_{s=0}^{\infty} \mathbf{x}(s)\pi(\mathbf{x}(s))ds \right)^\top dt \\ &= \int_{t=0}^{\infty} \mathbf{x}(t)\mathbf{x}(t)^\top \pi(\mathbf{x}(t))dt - \left(\int_{t=0}^{\infty} \mathbf{x}(t)\pi(\mathbf{x}(t))dt \right) \left(\int_{t=0}^{\infty} \mathbf{x}(t)\pi(\mathbf{x}(t))dt \right)^\top . \end{aligned} \tag{1}$$

Despite the formal correctness of equation (1), in real case applications, it is not actually applicable since requiring a continuous timestamp t of acquisition which is clearly unfeasible. Last but not least, we do not know, in general, the distribution π according to which the data $\mathbf{x}(1), \dots, \mathbf{x}(T)$ are sampled. Thus, as a natural estimator for $\mathbb{S}$, one typically exploits the *sampling covariance estimator*

$$\widehat{\mathbb{S}}(\mathbf{X}) = \frac{1}{T-1} \sum_{t=1}^{T} \left(\mathbf{x}(t) - \frac{1}{T} \sum_{s=1}^{T} \mathbf{x}(s) \right) \left(\mathbf{x}(t) - \frac{1}{T} \sum_{s=1}^{T} \mathbf{x}(s) \right)^\top , \tag{2}$$

where $\mathbf{X}$ represents the $3n \times T$ data matrix which encodes $\mathbf{x}(1), \dots, \mathbf{x}(T)$ in a way that $\mathbf{X}_{it}$ is the i-th component of $\mathbf{x}(t)$. The multiplicative factor $\frac{1}{T-1}$ represents the Bessel correction so that $\widehat{\mathbb{S}}(\mathbf{X})$ is an unbiased estimator of the original covariance matrix $\mathbb{S}(\pi)$. For the sake of simplicity, since in the rest of our discussion we will focus on $\widehat{\mathbb{S}}(\mathbf{X})$ only, we will simply refer to it as covariance, omitting both "sampling" and "estimator" attributes. In order to support what follows, we now rewrite $\widehat{\mathbb{S}}$ in a matrix expression.

Proposition 1. *Let* $\mathbf{P}$ *the* $T \times T$ *symmetric matrix whose generic* (s, t) *entry is*

$$P_{st} = \frac{\delta_{st}}{T-1} - \frac{1}{T(T-1)} = \begin{cases} \dfrac{1}{T} & \text{if } s = t \\ \dfrac{1}{T^2 - T} & \text{if } s \neq t, \end{cases} \tag{3}$$

where δ_{st} denotes the Kronecker symbol ($\delta_{st} = 1$ if $s = t$, vanishing otherwise). Then, we get

$$\widehat{\mathbb{S}}(\mathbf{X}) = \mathbf{X}\mathbf{P}\mathbf{X}^\top, \tag{4}$$

Proof. Let us define with s_{ij} the generic entry of $\widehat{\mathbb{S}}(\mathbf{X})$ of row i and column j. It results

$$\begin{aligned} s_{ij} &= \frac{1}{T-1}\sum_{t=1}^{T}\left(\mathbf{X}_{it} - \frac{1}{T}\sum_{s=1}^{T}\mathbf{X}_{is}\right)\left(\mathbf{X}_{jt} - \frac{1}{T}\sum_{r=1}^{T}\mathbf{X}_{jr}\right) \\ &= \frac{1}{T-1}\sum_{t=1}^{T}\mathbf{X}_{it}\mathbf{X}_{jt} - \frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{r=1}^{T}\mathbf{X}_{it}\mathbf{X}_{jr} - \frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbf{X}_{jt}\mathbf{X}_{is} + \frac{1}{T^2(T-1)}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{r=1}^{T}\mathbf{X}_{is}\mathbf{X}_{jr} \end{aligned} \tag{5}$$

In the last summation in the right side of (5) there is no addend which depends on t, thus

$$s_{ij} = \frac{1}{T-1}\sum_{t=1}^{T}\mathbf{X}_{it}\mathbf{X}_{jt} - \frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{r=1}^{T}\mathbf{X}_{it}\mathbf{X}_{jr} - \frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbf{X}_{jt}\mathbf{X}_{is} + \frac{1}{T(T-1)}\sum_{s=1}^{T}\sum_{r=1}^{T}\mathbf{X}_{is}\mathbf{X}_{jr} \tag{6}$$

since the summation over t counts T elements and we also simplified with the T in the denominator. In the right side of (6) the second and fourth addends are equal in magnitude and opposite in sign: this follows by modifying the summation index in the fourth addends according to the transformation $s \mapsto t$. Therefore

$$s_{ij} = \frac{1}{T-1}\sum_{t=1}^{T}\mathbf{X}_{it}\mathbf{X}_{jt} - \frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbf{X}_{jt}\mathbf{X}_{is}.$$

We can exploit the properties of Kronecker symbol, consequently obtaining

$$\begin{aligned} s_{ij} &= \frac{1}{T-1}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbf{X}_{is}\delta_{st}\mathbf{X}_{jt} - \frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbf{X}_{jt}\mathbf{X}_{is} \\ &= \sum_{s=1}^{T}\sum_{t=1}^{T}\mathbf{X}_{is}\left(\frac{\delta_{st}}{T-1} - \frac{1}{T(T-1)}\right)\mathbf{X}_{jt} \end{aligned} \tag{7}$$

From (7), for every $s, t = 1, \ldots, T$ the definition $\mathbf{P}_{st}$ according to the first equality of (3) ensures that the second one is immediately verified (this is easily checked with a few algebra). Thus, (7) rewrites

$$s_{ij} = \sum_{s=1}^{T}\sum_{t=1}^{T}\mathbf{X}_{is}\mathbf{P}_{st}\mathbf{X}_{jt} = \sum_{s=1}^{T}\sum_{t=1}^{T}\mathbf{X}_{is}\mathbf{P}_{st}\left(\mathbf{X}^\top\right)_{tj} \tag{8}$$

which produces the thesis thanks to the formal definition of the row-by-column matrix product and the arbitrary indexes i and j considered. □

As a classical property of covariance $\widehat{\mathbb{S}}$, it can well capture linear interdependencies between variables. Indeed, if the greater value of variable i corresponds with an increased variable j, $\widehat{\mathbb{S}}$ tends to show a similar trend and this reflects in $s_{ij} > 0$. In the opposite situation when variable i increases and variable j decreases, then $s_{ij} < 0$. Thus, the sign of s_{ij} is indicative of linear/anti-linear tendencies between i-th and j-th variables. Also, normalized versions of the covariance (e.g., McPerson's correlation coefficient) are also able to quantify how such tendencies are strong. Anyway, covariance is not able to understand more general relationships than the linear ones and, unfortunately, this can be insufficient when dealing with real data in applicative scenarios. Thus, as a naive approach to solve such an issue, one might apply a preliminary encoding of the raw data $\mathbf{x}(t)$ by computing a suitable feature transformation $\Phi(\mathbf{x}(t))$ for $t = 1, \ldots, T$. As a result, one gets

$$\widehat{\mathbb{S}}(\mathbf{\Phi}(\mathbf{X})) = \mathbf{\Phi}(\mathbf{X})\mathbf{P}\mathbf{\Phi}(\mathbf{X})^\top, \tag{9}$$

where we defined $\mathbf{\Phi}(\mathbf{X})$ the matrix whose (i,t)-th entry is the i-th component of $\Phi(\mathbf{x}(t))$. Formally, (9) looks for linear relationships in the augmented feature space by means of Φ and, if a suitable feature map is designed, this can be therefore equivalent to model arbitrary interdependencies in the original data space. However, as the main bottleneck with (9) is, generally, the dimension of the feature space is much higher than the original one: computationally, $\mathbf{\Phi}(\mathbf{X})$ can be extremely onerous to compute and storage. Also, it totally precludes the case of infinite dimensional feature spaces, although they are quite common in practice: indeed, in this case, Φ can not be exactly computed and, for instance, has to be approximated with a finite surrogate.

As a different perspective, to solve the aforementioned issue, many established algorithms (such as support vector machines or principal component analysis) observed that rather than the explicit computation of the feature map ϕ, the quantity that has

to be computed is instead the dot product between feature maps and this is classically done via a kernel function. Formally, a kernel function $k\colon \mathbb{R}^{3n} \times \mathbb{R}^{3n} \to \mathbb{R}$ is defined to be symmetric and positive definite and, thanks to Mercer's theorem [1],

$$k(\mathbf{x}, \mathbf{z}) = \sum_{n \in \mathbb{N}} \lambda_n \phi_n(\mathbf{x}) \phi_n(\mathbf{z}) \tag{10}$$

where $(\lambda_n, \phi_n)_{n \in \mathbb{N}}$ is the countable eigen-system of the Hilbert-Schmidt operator related with k, where the eigenvalues λ_n are non-negatives and $\lambda_n \geq \lambda_{n+1}$ for every n. Clearly, from (10), if one defines $\mathcal{H}$ the Hilbert space $\ell^2(\mathbb{N})$ of square-summable sequences and define $\Phi(\mathbf{x}) = (\sqrt{\lambda_n}\phi_n)_{n \in \mathbb{N}}$, we obtain

$$k(\mathbf{x}, \mathbf{z}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \rangle_{\mathcal{H}}, \tag{11}$$

which is exactly equation (5) in the paper. By jointly applying (11) and (9), we can rewrite s_{ij}, the (i,j)-th entry $\widehat{\mathbb{S}}_{ij}(\mathbf{\Phi}(\mathbf{X}))$, in the following manner

$$\widehat{\mathbb{S}}_{ij}(\mathbf{\Phi}(\mathbf{X})) = \sum_{s=1}^{T} \sum_{t=1}^{T} \Phi_i(\mathbf{x}(s)) \mathbf{P}_{st} \Phi_j(\mathbf{x}(t)) = \sum_{s=1}^{T} \sum_{t=1}^{T} \langle \Phi(\mathbf{x}(s)), \mathbf{e}_i \rangle_{\mathcal{H}} \mathbf{P}_{st} \langle \Phi(\mathbf{x}(s)), \mathbf{e}_j \rangle_{\mathcal{H}} \tag{12}$$

where $\mathbf{e}_j$ denotes the j-th element of the orthonormal basis of $\mathcal{H}$ which is clearly well defined since $\mathbf{e}_j \in \ell^2(\mathbb{N})$. If we now assume that, for any j, there exists $\mathbf{h}_j \in \mathbb{R}^{3n}$ such that $\Phi(\mathbf{h}_j) = \mathbf{e}_j$ we can rephrase (12) in

$$\widehat{\mathbb{S}}_{ij}(\mathbf{\Phi}(\mathbf{X})) = \sum_{s=1}^{T} \sum_{t=1}^{T} \langle \Phi(\mathbf{x}(s)), \Phi(\mathbf{h}_i) \rangle_{\mathcal{H}} \mathbf{P}_{st} \langle \Phi(\mathbf{x}(s)), \Phi(\mathbf{h}_j) \rangle_{\mathcal{H}} = \sum_{s=1}^{T} \sum_{t=1}^{T} k(\mathbf{x}(s), \mathbf{h}_i) \mathbf{P}_{st} k(\mathbf{x}(t), \mathbf{h}_j). \tag{13}$$

Hence, if we define $\mathbf{K}[\mathbf{X}, \mathbf{h}]$ as the matrix whose generic (i,t)-th entry is $k(\mathbf{x}(t), \mathbf{h}_i)$ we can finally conclude

$$\widehat{\mathbb{S}}(\mathbf{\Phi}(\mathbf{X})) = \mathbf{K}[\mathbf{X}, \mathbf{h}] \mathbf{P} \mathbf{K}[\mathbf{X}, \mathbf{h}]^{\top}, \tag{14}$$

which is precisely the claim of Lemma 1 in the paper. Such theoretical statement is relevant since, if defined $\widehat{\mathbb{S}}(k) = \mathbf{K}[\mathbf{X}, \mathbf{h}] \mathbf{P} \mathbf{K}[\mathbf{X}, \mathbf{h}]^{\top}$, equation (14) consists in an operative formula to compute the sampling covariance $\widehat{\mathbb{S}}(\mathbf{\Phi}(\mathbf{X}))$ by means of the equivalent expression $\widehat{\mathbb{S}}(k)$ which, additionally, involves the kernel function k only and not requires the explicit usage of Φ. Despite of this, such expression grounds on the assumption

$$\Phi(\mathbf{h}_i) = \mathbf{e}_i \qquad \text{for every } i = 1, \dots, \dim(\mathcal{H}) \tag{15}$$

which is actually restricting in general since it enforces the range of Φ to include all the elements $\{\mathbf{e}_i \colon i = 1, \dots, \dim(\mathcal{H})\}$. Indeed, since we want to perform the computation of $\widehat{\mathbb{S}}$ in terms of the kernel only, due to the fact the latter actual implicitly define the shape of Φ, in the paper, our proposed solution to recover the applicability of (15) is to fix the family of kernel function adopted. Thus, we assume that there exist non-negative coefficients $a_\ell \geq 0$ for $\ell \in \mathbb{N}$ such that

$$k(\mathbf{x}, \mathbf{z}) = \sum_{\ell=0}^{\infty} a_\ell \langle \mathbf{x}, \mathbf{z} \rangle^{\ell}, \tag{16}$$

where $\langle \mathbf{x}, \mathbf{z} \rangle$ is the usual inner product in $\mathbb{R}^{3n}$. It is notable that (16) includes both finite and infinite linear combinations, defines a proper kernel function in the sense of Mercer's theorem [1]. Also, equation (16) generalizes commonly used kernels: for instance, the (in)homogeneous polynomial kernel $k(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle^d + \beta$ of degree $d \in \mathbb{N}$ and bias $\beta \geq 0$ follows from the choices $a_0 = \beta$, $a_d = 1$ and $a_\ell = 0$ otherwise. Further, since playing a crucial role in the experimental part of the paper (Section 4.), we shall consider the exponential-dot product kernel

$$k(\mathbf{x}, \mathbf{z}) = \exp\left(\frac{\langle \mathbf{x}, \mathbf{z} \rangle}{\sigma}\right) = \sum_{\ell=0}^{\infty} \frac{\langle \mathbf{x}, \mathbf{z} \rangle^{\ell}}{\sigma^{\ell} \ell!} \tag{17}$$

is the case of $a_\ell = \frac{1}{\sigma^\ell \ell!} > 0$. for every $\sigma > 0$. In correspondence of the aforementioned kernel functions, a random class of feature maps $\mathbf{\Psi}$ is devised in the paper so that the linear kernels $\langle \mathbf{\Psi}(\mathbf{x}), \mathbf{\Psi}(\mathbf{z}) \rangle$ is an unbiased estimator and of $k(\mathbf{x}, \mathbf{z})$, also uniformly approximating with the additional assumption that the data belong to a compact set of $\mathbb{R}^{3n}$. Most importantly, it can be shown that $\mathbf{\Psi}$ fulfils the assumption (15). Precisely, $\mathbf{\Psi}\colon \mathbb{R}^{3n} \to \mathbb{R}^M$ is defined in a way that each component $\Psi_1, \dots, \Psi_M$ is an independent and identical distributed copy of the function which associate

$$\mathbb{R}^{3n} \ni \mathbf{x} \longmapsto \sqrt{a_N p^{N+1}} \prod_{j=1}^{N} \langle \boldsymbol{\omega}_j, \mathbf{x} \rangle$$

in correspondence of a random integer N where $N = n$ is sampled with probability $\frac{1}{p^{n+1}}$ for some hyper-parameter $p > 1$. If we thus compute the expectation of the linear kernel induced by $\mathbf{\Psi}$ over $\boldsymbol{\omega}$ and N we get

$$\mathbb{E}_{\boldsymbol{\omega},N}\left[\langle \mathbf{\Psi}(\mathbf{x}), \mathbf{\Psi}(\mathbf{z})\rangle\right] = \mathbb{E}_{\boldsymbol{\omega},N}\left[\frac{1}{M}\sum_{m=1}^{M}\Psi_m(\mathbf{x})\Psi_m(\mathbf{z})\right] = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\boldsymbol{\omega},N}\left[\Psi_m(\mathbf{x})\Psi_m(\mathbf{z})\right] \tag{18}$$

since the expectation is linear. Now, by using the definition of $\mathbf{\Psi}$ and Lemma 1 of the paper, we obtain

$$\mathbb{E}_{\boldsymbol{\omega},N}\left[\langle \mathbf{\Psi}(\mathbf{x}), \mathbf{\Psi}(\mathbf{z})\rangle\right] = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=0}^{\infty}\frac{1}{p^{n+1}}a_n p^{n+1}\langle \mathbf{x}, \mathbf{z}\rangle^n = \sum_{n=0}^{\infty}a_n\langle \mathbf{x}, \mathbf{z}\rangle^n = k(\mathbf{x}, \mathbf{z}), \tag{19}$$

ensuring that the bias of $\langle \mathbf{\Psi}(\mathbf{x}), \mathbf{\Psi}(\mathbf{z})\rangle$ as estimator of $k(\mathbf{x}, \mathbf{z})$ vanishes. Additionally, it easily checkable by the reader that all the results in [3] are still valid in our case, due to the fact that we can replace [3, Lemma 7] with Lemma 1 in our paper. Therefore, Lemma 8, 10 and 11 in [3] are applicable, giving the following statement of uniform approximation of $k(\mathbf{x}, \mathbf{z})$ by means of $\langle \mathbf{\Psi}(\mathbf{x}), \mathbf{\Psi}(\mathbf{z})\rangle$ and actually retrieving [3, Theorem 12].

Theorem 1. *Let assume that $k(\mathbf{x}, \mathbf{z})$ as in* (16) *is a dot product kernel over Ω where there exists $R > 0$ such that, for every $\mathbf{x} \in \Omega$, the Euclidean norm $\|\mathbf{x}\|$ of $\mathbf{x}$ satisfies $\|\mathbf{x}\| \leq R$. Then, for every small constant $\epsilon > 0$, the relationship*

$$\sup_{\mathbf{x},\mathbf{z}\in\Omega}\left|\langle \mathbf{\Psi}(\mathbf{x}), \mathbf{\Psi}(\mathbf{z})\rangle - k(\mathbf{x}, \mathbf{z})\right| \leq \epsilon \tag{20}$$

holds with probability $1 - 2\exp\left(-\frac{M\epsilon^2}{8C^2}\right)\left(\frac{32RL}{\epsilon}\right)^{6n}$, under the assumption that M is inferiorly bounded by $\frac{3n}{\epsilon^2}\log\left(\frac{1}{\epsilon}\right)$, while L and C are a quadratic and linear function of R, respectively.

In a few words, Theorem 1 certifies that with great probability, $\langle \mathbf{\Psi}(\mathbf{x}), \mathbf{\Psi}(\mathbf{z})\rangle$ is a uniform approximation for $k(\mathbf{x}, \mathbf{z})$. In addition, the map $\mathbf{\Psi}$ actually fulfils the assumption (15) as certified by Proposition 1 in the paper and, globally, this solves all theoretical issue. Additionally, the random feature map $\mathbf{\Psi}$ is much more convenient than Φ for the reason that the system $\Phi(\mathbf{h}_j) = \mathbf{e}_j$ has $\dim(\mathcal{H})$ equations and unknowns (thus can be infinite dimensional). Thus, on the contrary, for $\mathbf{\Psi}$ such dimension is just M: the aforementioned theorem enforces us to choose M in a way that $M \geq 3n$ and in the experimental part of the paper we precisely fixed $M = 3n$ in order gather a low dimension for the $\widehat{\mathbb{S}}(k)$. However, at the same time, such choice was able to score the strong performance registered for all the MoCap dataset considered in the paper.

REFERENCES

[1] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, ser. Adaptive Computation and Machine Learning, 2002.

[2] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, Ed., 1966.

[3] P. Kar and H. Karnick, "Random features maps for dot product kernels," in *JMLR*, 2012.