Supplementary materials for "Curve correlation"

October 29, 2025

A A third type of centering

While we do not advocate centering each function by its expectation as in Dubin & Müller (2005) and Liu et al. (2016), there is a third type of centering that may be warranted in some cases, such as the Canadian temperature data of Ramsay & Silverman (2005) analyzed in Paul et al. (2025) and in section 4: when there is a strong common trend among the p functions, this should be removed to avoid an uninteresting matrix of very high positive correlations among all of them (see section 4 of Paul et al. (2025)). To summarize, then, considering for simplicity an $n \times p \times M$ data array comprising M instances of p functions each observed at n time points, dynamic correlation estimation Dubin & Müller (2005) entails centering along the first and third dimensions. (A disadvantage of this, as noted by Opgen-Rhein & Strimmer (2006), is that the dynamic correlation cannot be estimated when $M = 1$.) Curve correlation, on the other hand, entails centering only along the first dimension, and possibly along the second, in the case of a strong common trend. Different approaches to centering may be appropriate for different applications; one should, of course, always be clear about what was done and why.

B Separable multivariate Gaussian processes

Before deriving formula (13) in supplementary appendix C below, we briefly explain the notion of a separable multivariate Gaussian process. Given a mean function $\boldsymbol{\mu} : \mathcal{I} \to \mathbb{R}^p$, temporal (within-curve) covariance function $\Gamma : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$ and $p \times p$ between-curve covariance matrix $\boldsymbol{\Sigma}$, we say that $\boldsymbol{x} : \mathcal{I} \to \mathbb{R}^p$ arises from the separable multivariate Gaussian process $\mathrm{MGP}(\boldsymbol{\mu}, \Gamma, \boldsymbol{\Sigma})$ (Morris & Carroll 2006, Chen et al. 2020) if, for any $t_1, \ldots, t_n \in \mathcal{I}$, the $n \times p$ matrix $\boldsymbol{X}_{t_1,\ldots,t_n} \equiv [x_u(t_i)]_{1 \le i \le n, 1 \le u \le p}$ has the matrix-variate normal distribution (Dawid 1981, Gupta & Nagar 1999) with $n \times p$ mean matrix $\boldsymbol{M}_{t_1,\ldots,t_n} \equiv [\mu_u(t_i)]_{1 \le i \le n, 1 \le u \le p}$, between-row covariance matrix $\boldsymbol{\Gamma}_{t_1,\ldots,t_n} \equiv [\Gamma(t_i, t_j)]_{1 \le i,j \le n}$ and between-column covariance matrix $\boldsymbol{\Sigma}$; or equivalently,

$$\mathrm{vec}(\boldsymbol{X}^T_{t_1,\ldots,t_n}) \sim \mathcal{N}_{np}[\mathrm{vec}(\boldsymbol{M}^T_{t_1,\ldots,t_n}), \boldsymbol{\Gamma}_{t_1,\ldots,t_n} \otimes \boldsymbol{\Sigma}].$$

The Kronecker product form of the above covariance matrix makes the process "separable," in the sense for each i, j,

$$\mathrm{Cov}[\boldsymbol{x}(t_i), \boldsymbol{x}(t_j)] = \Gamma(t_i, t_j)\boldsymbol{\Sigma}, \tag{B1}$$

i.e., the cross-covariance matrix for times t_i, t_j can be separated into (expressed as the product of) the temporal covariance $\Gamma(t_i, t_j)$ and the between-variable covariance $\boldsymbol{\Sigma}$. (Multivariate Gaussian processes *not* satisfying this separability property are commonly studied in geostatistics (Gelfand 2021) and in the kernel literature (Alvarez et al. 2012).)

The above process is unidentifiable in the sense that it is equal to $\mathrm{MGP}(\boldsymbol{\mu}, h\Gamma, h^{-1}\boldsymbol{\Sigma})$ for any $h > 0$. But if the process is stationary, identifiability can be established by letting Γ be an autocorrelation function

$$\Gamma(s, t) = \varphi(s - t) \tag{B2}$$

where φ is an even function satisfying $\varphi(0) = 1$.

C Derivation of (13)

Our derivation of (13) is based on a related formula for a stationary bivariate Gaussian time series (x_{1t}, x_{2t}), $t = 1, \ldots, n$ with autocorrelations and lagged cross-correlations

$$\rho_{11,k} = \mathrm{Cor}(x_{1t}, x_{1,t+k}), \quad \rho_{22,k} = \mathrm{Cor}(x_{2t}, x_{2,t+k}), \quad \rho_{12,k} = \mathrm{Cor}(x_{1t}, x_{2,t+k})$$

for integer-valued lags k. Note that $\rho_{12,0}$ is the (non-lagged) cross-correlation ρ. By eq. (2) of Afyouni et al. (2019), the ordinary sample correlation r_n between x_{1t} and x_{2t} has approximate variance[1]

$$\begin{aligned}
\mathrm{Var}(r_n) \;\approx\; & n^{-2}\Big[n(1-\rho^2)^2 \\
& + \rho^2 \sum_{k=1}^{n-1}(n-k)(\rho_{11,k}^2 + \rho_{22,k}^2 + \rho_{12,k}^2 + \rho_{12,-k}^2) \\
& - 2\rho \sum_{k=1}^{n-1}(n-k)(\rho_{11,k} + \rho_{22,k})(\rho_{12,k} + \rho_{12,-k}) \\
& + 2\sum_{k=1}^{n-1}(n-k)(\rho_{11,k}\rho_{22,k} + \rho_{12,k}\rho_{12,-k})\Big].
\end{aligned} \tag{C3}$$

Expression (13) is an analogous variance formula for r^* given $[x_1(t), x_2(t)]$ arising from a bivariate Gaussian process on $\mathcal{I}$, with lag-τ auto- and cross-correlations $\varrho_1(\tau), \varrho_2(\tau), \varrho_{12}(\tau)$ defined by (12). For simplicity we take $\mathcal{I} = [0, 1]$. Moreover, since (C3) is valid for bivariate observations measured at n arbitrary equally spaced time points, we can take these time points to be $\frac{1}{n}, \frac{2}{n}, \ldots, 1$. The derivation of (13) proceeds in four steps:

[1]The authors of Afyouni et al. (2019) wrote to the publisher to report typographical errors in equation (2) and two other formulas, but the corrections appear not to have been published.

1. "Translate" the discrete time series values and their auto- and cross-correlations to appropriate function values.

2. Show that under mild conditions,

$$\mathrm{Var}(r_n) \to \mathrm{Var}(r^*). \tag{C4}$$

3. Compute the limit of the approximate variance (C3) of r_n as $n \to \infty$, and conclude from (C4) that this is an approximate variance formula for r^*.

4. Show that if, in addition to being stationary, the bivariate process is separable, then the limit from step 3 equals (13).

Step 1. Given the time points $\frac{1}{n}, \frac{2}{n}, \ldots, 1$, for $u = 1, 2$, the discrete observations $x_{u1}, \ldots, x_{un}$ are replaced by $x_u(\frac{1}{n}), \ldots, x_u(1)$, and $\rho_{11,k}, \rho_{22,k}, \rho_{12,k}$ in (C3) are equal to

$$\varrho_1(k/n), \varrho_2(k/n), \varrho_{12}(k/n),$$

respectively.

Step 2. Aside from stationarity and separability, the only assumptions required for the bivariate Gaussian process are as follows:

(i) For $u = 1, 2$, $x_u : [0, 1] \to \mathbb{R}$ is almost surely (a.s.) Riemann integrable.

(ii) For $u = 1, 2$, $\int_0^1 x_u^c(t)^2 dt \neq 0$ a.s., where the superscript ^{c} denotes time-centering as in (2).

(iii) The autocorrelation functions $\varrho_1(\tau), \varrho_2(\tau)$ and the cross-correlation function $\varrho_{12}(\tau)$ are a.s. Riemann integrable.

By Assumption (i), the sample variances of $x_1(\frac{1}{n}), \ldots, x_1(1)$ and $x_2(\frac{1}{n}), \ldots, x_2(1)$, and the sample covariance between them, converge a.s. to

$$\int_0^1 x_1^c(t)^2 dt, \quad \int_0^1 x_2^c(t)^2 dt, \quad \int_0^1 x_1^c(t) x_2^c(t) dt,$$

respectively. It follows by Assumption (ii), definition (3) of r^*, and the continuous mapping theorem that $r_n \to r^*$, a.s. and therefore in probability. This, together with Theorem 5.12 of Kallenberg (2021) and the boundedness of r_n, implies that $r_n \to r^*$ in mean square, from which (C4) follows as required.

Step 3. Using the "translations" from step 1, (C3) can be rewritten as

$$\begin{aligned}\mathrm{Var}(r_n) \;\approx\; & n^{-2}\Big[n(1-\rho^2)^2 \\ & +\rho^2\sum_{k=1}^{n-1}(n-k)\left\{\varrho_1^2(k/n)+\varrho_2^2(k/n)+\varrho_{12}^2(k/n)+\varrho_{12}^2(-k/n)\right\} \\ & -2\rho\sum_{k=1}^{n-1}(n-k)\left\{\varrho_1(k/n)+\varrho_2(k/n)\right\}\left\{\varrho_{12}(k/n)+\varrho_{12}(-k/n)\right\} \\ & +2\sum_{k=1}^{n-1}(n-k)\left\{\varrho_1(k/n)\varrho_2(k/n)+\varrho_{12}(k/n)\varrho_{12}(-k/n)\right\}\Big], \end{aligned} \tag{C5}$$

As $n\to\infty$, the first term in (C5) vanishes and the remaining terms (the three sums) converge to integrals. For example, the second term in (C5) can be written as

$$\rho^2\sum_{k=1}^{n-1}\left(\frac{1}{n}\right)\left(1-\frac{k}{n}\right)\left[\varrho_1^2(k/n)+\varrho_2^2(k/n)+\varrho_{12}^2(k/n)+\varrho_{12}^2(-k/n)\right],$$

which, by Assumption (iii), is a Riemann sum converging to

$$\rho^2\int_0^1(1-\tau)\left[\varrho_1^2(\tau)+\varrho_2^2(\tau)+\varrho_{12}^2(\tau)+\varrho_{12}^2(-\tau)\right]d\tau$$

as $n\to\infty$. Applying the same argument to the last two terms of (C5) yields the following approximate variance formula for r^*:

$$\begin{aligned}\mathrm{Var}(r^*) \;\approx\; & \rho^2\int_0^1(1-\tau)\left[\varrho_1^2(\tau)+\varrho_2^2(\tau)+\varrho_{12}^2(\tau)+\varrho_{12}^2(-\tau)\right]d\tau \\ & -2\rho\int_0^1(1-\tau)\left[\varrho_1(\tau)+\varrho_2(\tau)\right]\left[\varrho_{12}(\tau)+\varrho_{12}(-\tau)\right]d\tau \\ & +2\int_0^1(1-\tau)\left[\varrho_1(\tau)\varrho_2(\tau)+\varrho_{12}(\tau)\varrho_{12}(-\tau)\right]d\tau. \end{aligned} \tag{C6}$$

Step 4. Assuming the stationary Gaussian bivariate process is separable (see (B1), (B2)) with between-curve covariance matrix $\mathbf{\Sigma}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$ and lag-τ autocorrelation $\varphi(\tau)$, we have

$$\varrho_1(\tau)=\frac{\mathrm{Cov}[x_1(t),x_1(t+\tau)]}{\sqrt{\mathrm{Var}[x_1(t)]\mathrm{Var}[x_1(t+\tau)]}}=\frac{\varphi(\tau)\sigma_1^2}{\sqrt{[\varphi(0)\sigma_1^2][\varphi(0)\sigma_1^2]}}=\varphi(\tau),$$

and similarly $\varrho_2(\tau) = \varphi(\tau)$ and $\varrho_{12}(\tau) = \varrho_{12}(-\tau) = \rho\varphi(\tau)$. Thus (C6) becomes

$$\begin{aligned}\mathrm{Var}(r^*) &\approx \rho^2(2+2\rho^2)\int_0^1 (1-\tau)\varphi(\tau)^2 d\tau \\ &\quad -8\rho^2\int_0^1 (1-\tau)\varphi(\tau)^2 d\tau + 2(1+\rho^2)\int_0^1 (1-\tau)\varphi(\tau)^2 d\tau \\ &= 2(1-\rho^2)^2\int_0^1 (1-\tau)\varphi(\tau)^2 d\tau,\end{aligned}$$

verifying (13).

References

Afyouni, S., Smith, S. M. & Nichols, T. E. (2019), 'Effective degrees of freedom of the Pearson's correlation coefficient under autocorrelation', *NeuroImage* **199**, 609–625.

Alvarez, M. A., Rosasco, L. & Lawrence, N. D. (2012), 'Kernels for vector-valued functions: A review', *Foundations and Trends in Machine Learning* **4**(3), 195–266.

Chen, Z., Wang, B. & Gorban, A. N. (2020), 'Multivariate Gaussian and Student-t process regression for multi-output prediction', *Neural Computing and Applications* **32**(8), 3005–3028.

Dawid, A. P. (1981), 'Some matrix-variate distribution theory: Notational considerations and a Bayesian application', *Biometrika* **68**(1), 265–274.

Dubin, J. A. & Müller, H.-G. (2005), 'Dynamical correlation for multivariate longitudinal data', *Journal of the American Statistical Association* **100**, 872–881.

Gelfand, A. E. (2021), Multivariate spatial process models, *in* M. M. Fischer & P. Nijkamp, eds, 'Handbook of Regional Science', Springer, Berlin and Heidelberg, pp. 1985–2016.

Gupta, A. K. & Nagar, D. K. (1999), *Matrix Variate Distributions*, Chapman and Hall/CRC.

Kallenberg, O. (2021), *Foundations of Modern Probability*, 3rd edn, Springer Nature, Cham, Switzerland.

Liu, S., Zhou, Y., Palumbo, R. & Wang, J.-L. (2016), 'Dynamical correlation: A new method for quantifying synchrony with multivariate intensive longitudinal data', *Psychological Methods* **21**(3), 291.

Morris, J. S. & Carroll, R. J. (2006), 'Wavelet-based functional mixed models', *Journal of the Royal Statistical Society: Series B* **68**(2), 179–199.

Opgen-Rhein, R. & Strimmer, K. (2006), 'Inferring gene dependency networks from genomic longitudinal data: a functional data approach', *REVSTAT - Statistical Journal* **4**(1), 53–65.

Paul, B., Reiss, P. T., Cui, E. & Foà, N. (2025), 'Continuous-time multivariate analysis', *Journal of Computational and Graphical Statistics* **34**(1), 384–394.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional Data Analysis*, 2nd edn, Springer, New York.