Testing equality of two covariance operators

Daniela JARUŠKOVÁ

CzechTechnical University of Prague, Dept. of Mathematics

We observe two independent sequences of i.i.d. zero mean Gaussian processes $X_1(t), \ldots, X_{N_1}(t)$ and $Y_1(t), \ldots, Y_{N_2}(t)$ defined for $t \in [0,1]$ such that $E\,\|X_1(t)\|^2 = E\int_0^1 X_1^2(t)\,dt < \infty$ and $E\,||Y_1(t)||^2 = E\int_0^1 Y_1^2(t)\,dt < \infty$.

($L^2[0,1]$ is a Hilbert space of square integrable functions on $[0,1]$ with a scalar product $< f, g >= \int_0^1 f(t)\,g(t)\,dt$ and $\|\,\|$ the corresponding L_2 norm.)

The covariance functions $A(t,s) = E\,X_1(t)X_1(s)$ and $B(t,s) = E\,Y_1(t)Y_1(s)$ are continuous functions on $[0,1]^2$. $\mathcal{A}$ is the corresponding covariance operator of X_1 defined by the kernel $A(t,s)$ and $\mathcal{B}$ is the covariance operator of Y_1 defined by the kernel $B(t,s)$:

$$(\mathcal{A}v)(t) = \int_0^1 A(t,s)\,v(s)\,ds, \quad (\mathcal{B}v)(t) = \int_0^1 B(t,s)\,v(s)\,ds.$$

According to the Mercer lemma there exist expansions:

$$A(t,s) = \sum_{i=1}^{\infty} \lambda_i \, u_i(t) \, u_i(s), \quad B(t,s) = \sum_{i=1}^{\infty} \mu_i \, v_i(t) \, v_i(s),$$

where $\{\lambda_i\}$, $\{u_i(t),\ t \in [0,1]\}$ are eigenelements of $\mathcal{A}$ and $\{\mu_i\}$, $\{v_i(t),\ t \in [0,1]\}$ are eigenelements of $\mathcal{B}$. Clearly

$$\lambda_k = \int_0^1 \int_0^1 u_k(t) A(t,s) u_k(s) \, dt \, ds =< u_k, \mathcal{A} u_k >,$$

$$\mu_k = \int_0^1 \int_0^1 v_k(t) B(t,s) v_k(s) \, dt \, ds =< v_k, \mathcal{B} v_k > .$$

$\lambda_1 > \lambda_2 > \cdots > \lambda_K > \ldots,$

$\mu_1 > \mu_2 > \cdots > \mu_K > \ldots.$

We estimate the covariance function $A(t,s)$ and $B(t,s)$ by

$$\widehat{A}(t,s) = \frac{1}{N_1}\sum_{i=1}^{N_1} X_i(t)X_i(s), \quad \text{resp.} \quad \widehat{B}(t,s) = \frac{1}{N_2}\sum_{i=1}^{N_2} Y_i(t)Y_i(s).$$

The corresponding operators are $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$:

$$(\widehat{\mathcal{A}}v)(t) = \int_0^1 \widehat{A}(t,s)\, v(s)\, ds, \quad (\widehat{\mathcal{B}}v)(t) = \int_0^1 \widehat{B}(t,s)\, v(s)\, ds.$$

We denote $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots$ eigenvalues and $\widehat{u}_1, \widehat{u}_2, \dots$ eigenfunctions of $\widehat{\mathcal{A}}$ while $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \dots$ eigenvalues and $\widehat{v}_1, \widehat{v}_2, \dots$ eigenfunctions of $\widehat{\mathcal{B}}$. ($\int \widehat{u}_k u_k \geq 0, \int \widehat{v}_k v_k \geq 0, k = 1, \dots$)
We introduce an operator $\widehat{\mathcal{C}}$ with the kernel

$$\widehat{C}(t,s) = \frac{1}{N_1+N_2}\Big(\sum_{i=1}^{N_1} X_i(t)X_i(s) + \sum_{i=1}^{N_2} Y_i(t)Y_i(s)\Big).$$

Two-sample problem

$$H_0 : \mathcal{A} = \mathcal{B} \qquad A : \mathcal{A} \neq \mathcal{B}$$

We define for $i = 1, \ldots, N_1$ and $k = 1, 2, \ldots$:

$$\beta^u_{X_i}(k) = < u_k, X_i > .$$

They are uncorrelated and because of normality also independent.

We define for $i = 1, \ldots, N_1$ and $k = 1, 2, \ldots,\ k' = k, k+1, \ldots$:

$$\eta^u_{X_i}(k, k') = < u_k, X_i >< u_{k'}, X_i > .$$

It holds

$$\begin{aligned} E\,\eta^u_{X_i}(k,k) &= \lambda_k \quad var\,\eta^u_{X_i}(k,k) = 2\lambda_k^2 \\ E\,\eta^u_{X_i}(k,k') &= 0 \quad var\,\eta^u_{X_i}(k,k') = \lambda_k\lambda_{k'} \end{aligned}$$

The variables $\left\{\eta_{X_i}(k,k'), k = 1, \ldots,\ k' = k, k+1, \ldots,\ i = 1, \ldots\right\}$ are uncorrelated.

For $k = 1, 2, \ldots$ it holds

$$|\widehat{\lambda}_k - \lambda_k| \leq |||\widehat{\mathcal{A}} - \mathcal{A}|||,$$
$$||\widehat{u}_k - u_k|| \leq const(k)|||\widehat{\mathcal{A}} - \mathcal{A}|||.$$

For $N_1 \to \infty$:

$$|||\widehat{\mathcal{A}} - \mathcal{A}||| = O_P(1/\sqrt{N_1}),$$
$$|\widehat{\lambda}_k - \lambda_k| = O_P(1/\sqrt{N_1}),$$
$$||\widehat{u}_k - u_K|| = O_P(1/\sqrt{N_1}).$$

Moreover

$$\sqrt{N_1}\,\frac{\widehat{\lambda}_k - \lambda_k}{\sqrt{2}\lambda_k} \sim N(0,1).$$

$$|||\mathcal{K}||| = \left(\int_0^1 \int_0^1 K(t,s)^2\,dt\,ds\right)^{1/2}$$

Panaratos et al. (2010):

$$\sum_{k=1}^{K}\sum_{k'=1}^{K}\left(\widehat{J}_N(k,k')\right)^2 = \sum_{k=1}^{K}\left(\widehat{J}_N(k,k)\right)^2 + 2\sum_{k=1}^{K}\sum_{k'=1}^{k-1}\left(\widehat{J}_N(k,k')\right)^2,$$

where

$$\widehat{J}_N(k,k') = \sqrt{\frac{N_1 N_2}{2N}}\frac{<\widehat{\phi}_k, (\widehat{\mathcal{A}}-\widehat{\mathcal{B}})\widehat{\phi}_{k'}>}{\sqrt{<\widehat{\phi}_k, \widehat{\mathcal{C}}\widehat{\phi}_k><\widehat{\phi}_{k'}, \widehat{\mathcal{C}}\widehat{\phi}_{k'}>}},$$

where the functions $\{\widehat{\phi}_k(t)\}$ are eigenfunctions of the operator $\widehat{\mathcal{C}}$.

Under H_0 we can replace $\widehat{\phi}_k$ by the right eigenfunctions u_k.

Under H_0 supposing that $N_1/N \to \alpha \in (0,1)$ the variables $(k \neq k')$

$$J_N(k,k) = \sqrt{\frac{N_1 N_2}{N}} \frac{< u_k, (\widehat{\mathcal{A}} - \widehat{\mathcal{B}}) u_k >}{\sqrt{2\,\lambda_k^2}} = \frac{\overline{\eta^u_{X_i}(k,k)} - \overline{\eta^u_{Y_i}(k,k)}}{\sqrt{2}\lambda_k \sqrt{1/N_1 + 1/N_2}}$$

and

$$J_N(k,k') = \sqrt{\frac{N_1 N_2}{N}} \frac{< u_k, (\widehat{\mathcal{A}} - \widehat{\mathcal{B}}) u_{k'} >}{\sqrt{\lambda_k\,\lambda_{k'}}} = \frac{\overline{\eta^u_{X_i}(k,k')} - \overline{\eta^u_{Y_i}(k,k')}}{\sqrt{\lambda_k \lambda_{k'}} \sqrt{1/N_1 + 1/N_2}}$$

are asymptotically $N(0,1)$ distributed. It follows that the suggested test statistic has asymptotically χ^2 distribution with $K(K+1)/2$ degrees of freedom.

What is K?

Two-sample problem

$$H_0 : \mathcal{A}_K = \mathcal{B}_K \qquad A : \mathcal{A}_K \neq \mathcal{B}_K,$$

where $\mathcal{A}_K$ corresponds to $A_K(t,s)$ and $\mathcal{B}_K$ corresponds to $B_K(t,s)$:

$$A_K(t,s) = \sum_{k=1}^{K} \lambda_k \, u_k(t) \, u_k(s), \quad B_K(t,s) = \sum_{k=1}^{K} \mu_k \, v_k(t) \, v_k(s).$$

$$\sum_{k=1}^{K} \left(\left(\widehat{T}_u(k) \right)^2 + \left(\widehat{T}_v(k) \right)^2 \right) / 2,$$

where

$$\widehat{T}_u(k) = \sqrt{\frac{N_1 N_2}{2\,N}} \frac{< \widehat{u}_k, (\widehat{\mathcal{A}} - \widehat{\mathcal{B}}) \widehat{u}_k >}{< \widehat{u}_k, \widehat{\mathcal{C}}\, \widehat{u}_k >} = \sqrt{\frac{N_1 N_2}{2\,N}} \frac{\widehat{\lambda}_k - \widetilde{\lambda}_k}{< \widehat{u}_k, \widehat{\mathcal{C}}\, \widehat{u}_k >}$$

and

$$\widehat{T}_v(k) = \sqrt{\frac{N_1 N_2}{2\,N}} \frac{< \widehat{v}_k, (\widehat{\mathcal{A}} - \widehat{\mathcal{B}}) \widehat{v}_k >}{< \widehat{v}_k, \widehat{\mathcal{C}}\, \widehat{v}_k >} = \sqrt{\frac{N_1 N_2}{2\,N}} \frac{\widetilde{\mu}_k - \widehat{\mu}_k}{< \widehat{v}_k, \widehat{\mathcal{C}}\, \widehat{v}_k >}.$$

$$\sum_{k=1}^{K} \left(\left(\widehat{T}_u(k) \right)^2 + \left(\widehat{T}_v(k) \right)^2 \right) / 2,$$

Under H_0 the test has a χ^2 distribution with K. degrees of freedom.

Let $\mathcal{A}_K \neq \mathcal{B}_K$ then there exists $k \leq K$ such that $< u_k, (\mathcal{A} - \mathcal{B})u_k > \neq 0$ or $< v_k, (\mathcal{A} - \mathcal{B})v_k > \neq 0$.

Assume that $N_1/N \to \alpha > 0$ as $N \to \infty$. Then under A the test based on my test statistic is consistent. More specifically, it holds

$$\left(\widehat{T}_u(k) - \sqrt{N}\sqrt{\alpha(1-\alpha)} \frac{< u_k, (\mathcal{A} - \mathcal{B})u_k >}{\sqrt{2}\big(\alpha\lambda_k + (1-\alpha)\kappa_k\big)} \right) = O_P(1).$$

$$\left(\widehat{T}_v(k) - \sqrt{N}\sqrt{\alpha(1-\alpha)} \frac{< v_k, (\mathcal{A} - \mathcal{B})v_k >}{\sqrt{2}\big(\alpha\nu_k + (1-\alpha)\mu_k\big)} \right) = O_P(1).$$

Let the operators $\mathcal{A}$ and $\mathcal{B}$ be represented by the matrix A, resp. B:

$$A = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 10 & 0 \\ 0 & 90 \end{pmatrix}.$$

The first principal component explains 90% of total variability of A and the same is true for B. Therefore, according to the described rule $K = 1$. Supposing $N_1 = 0.9\,N$ and $N_2 = 0.1\,N$, the covariance matrix of the pooled sample

$$C = 0.9A + 0.1B = \begin{pmatrix} 10 & 0 \\ 0 & 9.9 \end{pmatrix}$$

has the largest eigenvalue 10 and the corresponding eigenvector $(1,0)^T$. As $(1,0)\,(A-B)\,(1,0)^T = 0$, the procedure with $K = 1$ does not detect that $A_1 \neq B_1$ (and $A \neq B$) even if N is very large.

APPLICATIONS

Our original data were daily mean temperatures measured in two stations, namely in Milan in years 1763-1998 ($N_1 = 236$) and in Padua in years 1766 - 1982 ($N_2 = 217$). The data were organized into vectors of 365 components representing annual cycles that were smoothed by a kernel smoothing technique using an Epanechnikov window with a bandwidth of $h = 25$. In this way the analyzed data are two samples of 236 and 217 random functions.

k	$\widehat{\lambda}_k$	$\widehat{\lambda}_k / \sum \widehat{\lambda}_i$	$\frac{\sum_{i=1}^{k} \widehat{\lambda}_i}{\sum \widehat{\lambda}_i}$	$\widehat{\mu}_k$	$\widehat{\mu}_k / \sum \widehat{\mu}_i$	$\frac{\sum_{i=1}^{k} \widehat{\mu}_i}{\sum \widehat{\mu}_i}$
1	153.4	37 %	37 %	210.2	41 %	41 %
2	103.4	25 %	62 %	131.1	26 %	67 %
3	57.4	14 %	76 %	70.8	14 %	81 %
4	35.0	8 %	84 %	37.1	7 %	87 %
5	25.7	6 %	91 %	25.5	5 %	93 %

Table 1. The largest eigenvalues and corresponding proportions of total variability.

K	1	2	3	4	5
test stat.	6.32	9.46	13.98	14.19	14.28
p-values	0.012	0.009	0.003	0.007	0.014

Table 2. Values of test statistic and corresponding $p-$values for $K = 1, \ldots, 5$.

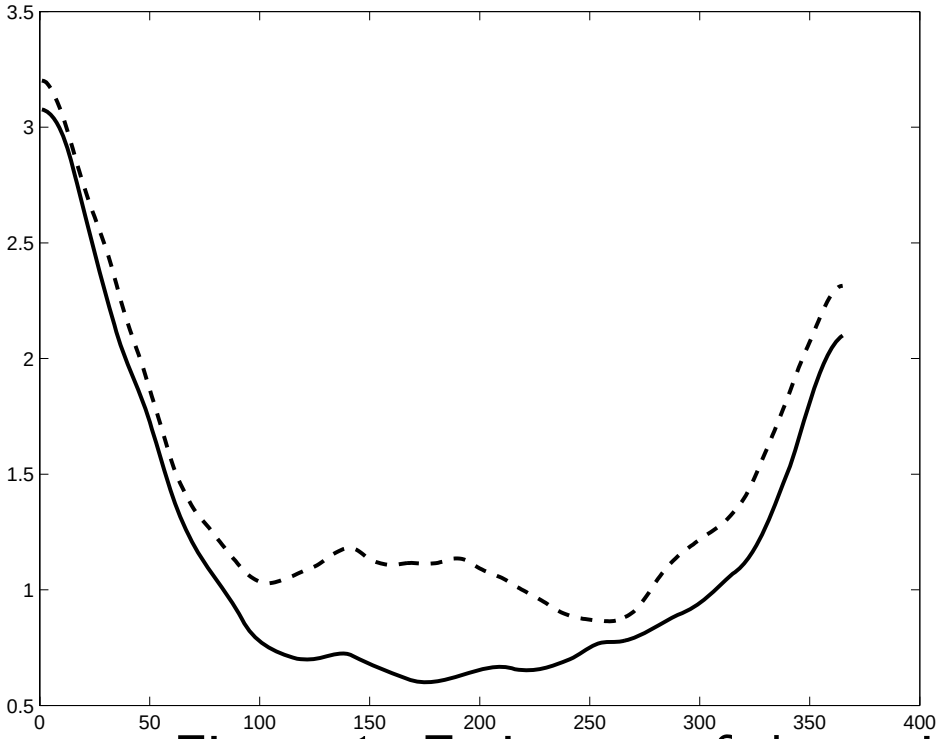

Figure 1. Estimates of the variance function for Milan (solid line) and for Padua (dashed line).

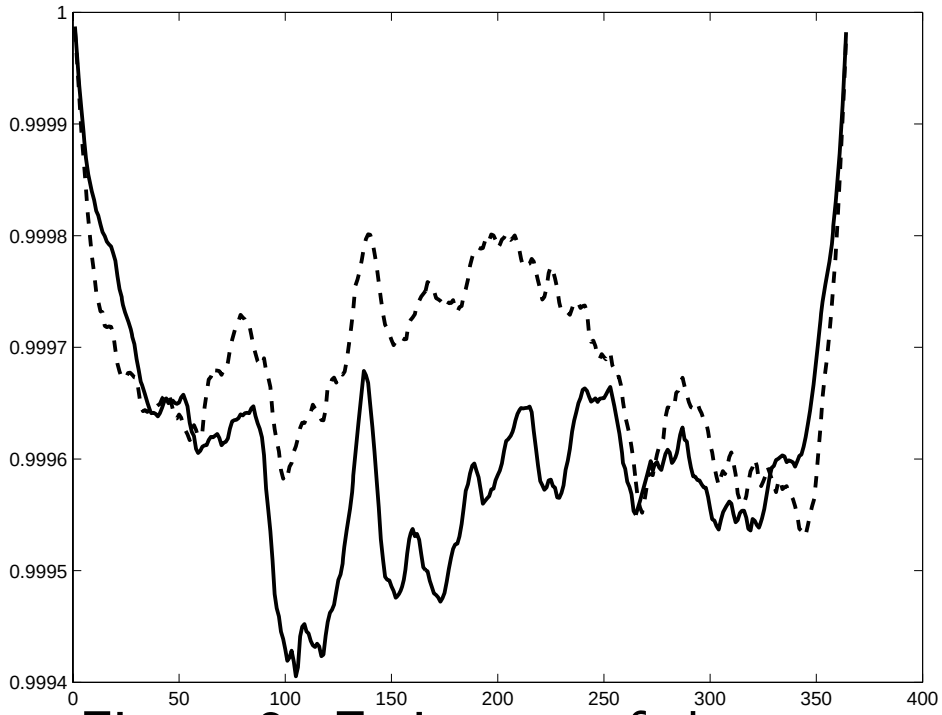

Figure 2. Estimates of the correlation function with the lag 1 for Milan (solid line) and for Padua (dashed line).

We observe that years with cold winters are followed by years with warm winters. This is true for Milan as well as for Padua. In spite of the fact that in winter the estimated correlation functions with the lag 1 for Milan and Padua data do not differ substantially, the correlation function with the lag 1 for Padua data attains larger values everywhere except in winter. In the other words the Padua temperature oscillates around its mean annual cycle more slowly but with slightly larger amplitudes. The sample variance of Milan annual averages is 0.36, while the sample variance of Padua annual averages is 0.56.

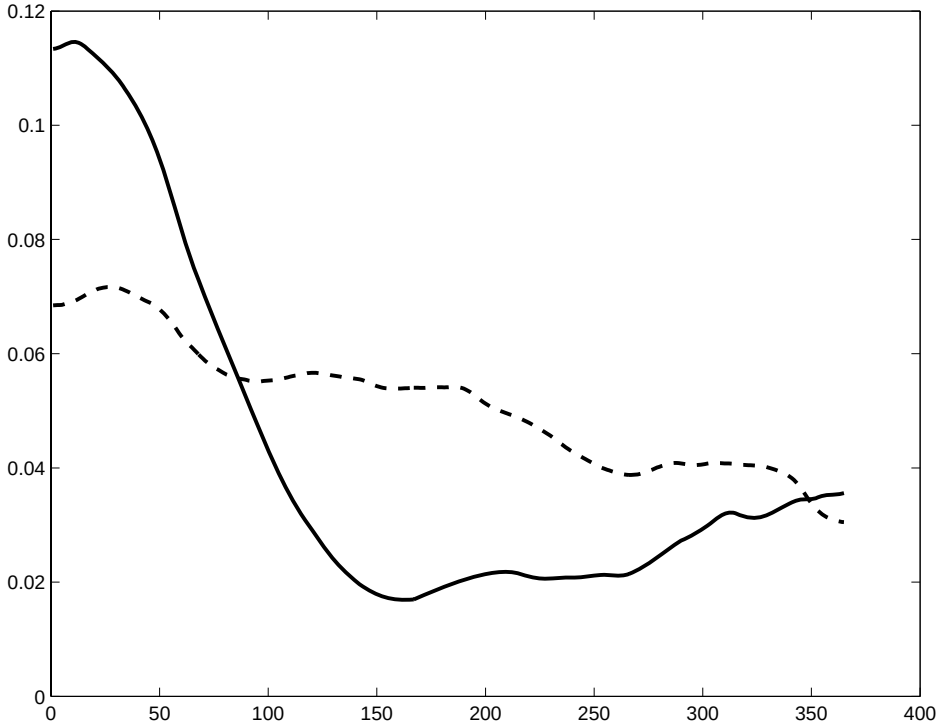

Figure 3. Eigenfunctions $\widehat{u}_1$ (solid line) and $\widehat{v}_1$ (dashed line).

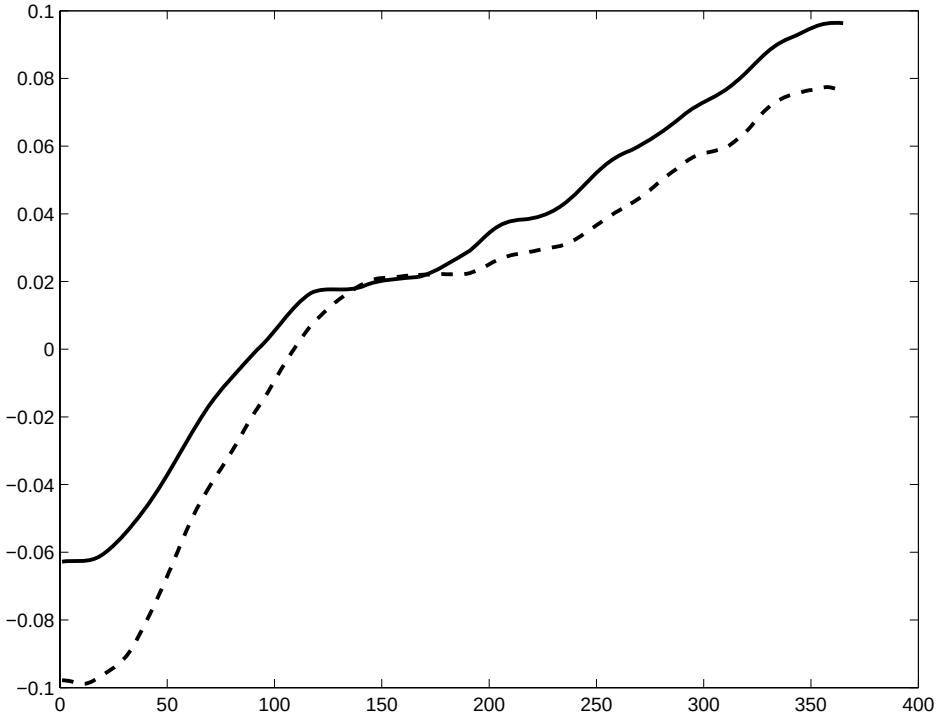

Figure 4. Eigenfunctions $\widehat{u}_2$ (solid line) and $\widehat{v}_2$ (dashed line).

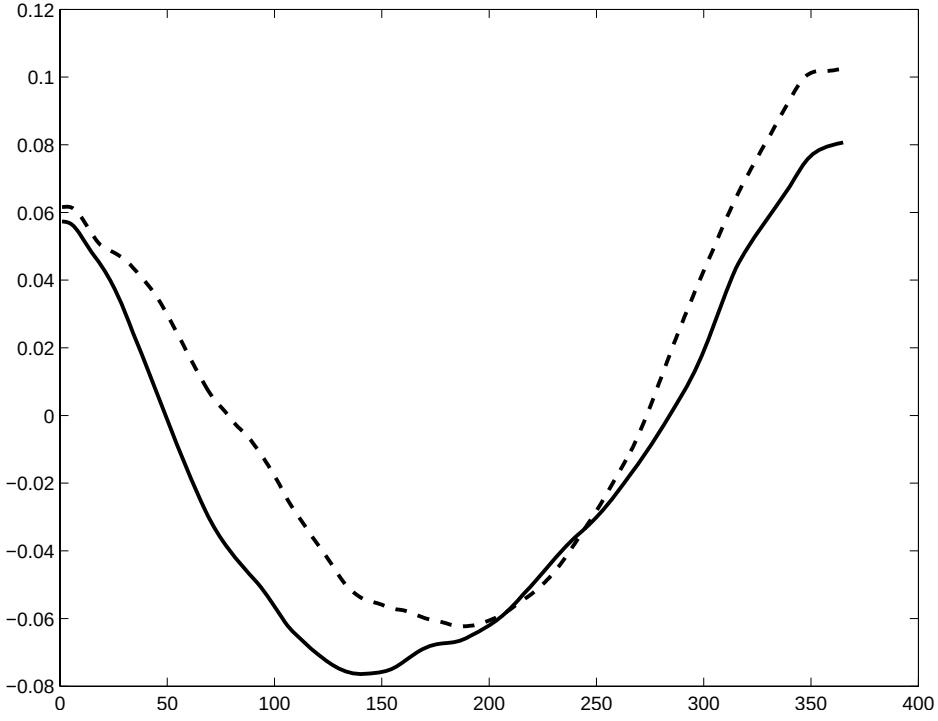

Figure 5. Eigenfunctions $\widehat{u}_3$ (solid line) and $\widehat{v}_3$ (dashed line).

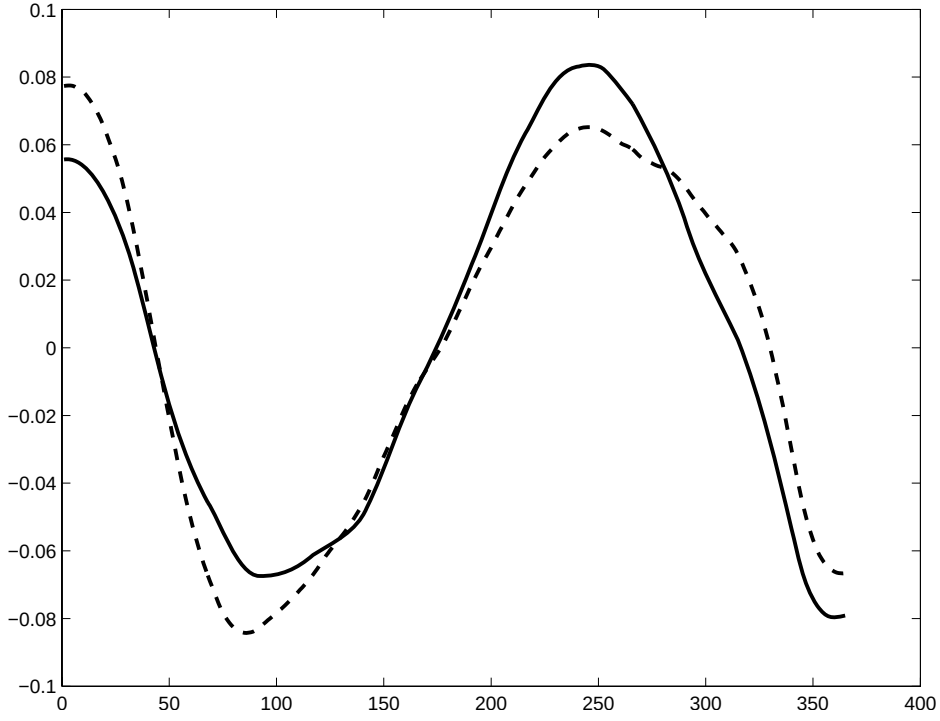

Figure 6. Eigenfunctions $\widehat{u}_4$ (solid line) and $\widehat{v}_4$ (dashed line).

Among all pairs $\{\widehat{u}_i, \widehat{v}_i\}$, $i = 1, 2, 3, 4$ the difference between $\widehat{u}_1$ and $\widehat{v}_1$ is the most striking. Both functions attain its maximum in the beginning of the year that corresponds to the fact that the main source of variability comes from the large year-to-year differences in winter temperatures. However, the eigenfunction $\widehat{v}_1$ decreases relatively slowly and is more similar to a constant function. That means that "weights" assigned to "daily" values in one calendar year are more equal. It means that the values of $\langle \widehat{v}_1, X_i \rangle$ and $\langle \widehat{v}_1, Y_i \rangle$ will attain values close to the corresponding annual averages. This is a reason why the function $\widehat{v}_1$ has a large ability to detect difference in covariance structure of Milan and Padua series ($\widehat{T}^v(1) = -3.24$.)