Supplementary document: Components and completion of partially observed functional data

David Kraus

Institute of Social and Preventive Medicine, University Hospital Lausanne, Switzerland

Summary. This supplementary document describes computational details of the proposed methods and provides proofs of Propositions 1, 2, 3 and 4.

1. Computation

1.1. Preliminary steps

In most applications, functional data are observed at discrete time points and are possibly subject to measurement error, so it is necessary to preprocess the raw data using smoothing techniques to obtain functions or their derivatives. In the context of partially observed functional data, the measurement time points are located only in observation periods O_i, while there are no measurements in missing periods M_i. We assume that the measurement points are dense in the observation periods, so that it is possible to apply smoothing techniques to obtain the functional values of the ith curve from the measured values of this curve. We use spline smoothing with a roughness penalty, as described in Ramsay and Silverman (2005, Chapter 5), but other methods like kernel smoothing can be used as well. In our experience, a simple approach works well: we apply the smoothing procedure to all values measured for the ith curve but use the computed smooth curve only for $t \in O_i$ (ignoring it on M_i where measurements are not available to make it reliable).

In practice, the observation and missing periods are typically not given (because they are not designed) and one needs to define them. For instance, one can define M_i to consist of the periods before the first and after the last measurement time and of all gaps between two consecutive measurement times that are larger than a certain threshold g. The value of g is the largest length of intervals without measurements over which we are willing to smooth. The choice of g depends on the particular setting; in general, if, for example, one considers K equidistant points in $[0, 1]$ (e.g., $K = 10$) as the minimum reliable design for smoothing on the whole domain $[0, 1]$, then $g = 1/K$ seems reasonable.

Sometimes, registration of functional data is needed. Shift registration (Ramsay and Silverman, 2005, Section 7.2) is easy to implement for incomplete functions: in the registration criterion the sample mean of partially observed functions is computed by the method described in the next subsection and the distance of each shifted curve from the sample mean is computed by numerical integration over the observed period of the curve; the criterion is minimised by the Procrustes method as usual. Methods based on warping can be modified similarly but further investigation of their performance is needed.

1.2. Principal component analysis, functional reconstruction

For practical computation we must use finite dimensional representations of functions and operators. Two traditional approaches exist: we can use either basis expansions or evaluation on a grid

of points. It is difficult to use the basis approach in our situation because incompletely observed functions are available on different subsets of the time domain. The grid approach is more suited for this type of data since it works directly with time arguments. Let $t_k = (k-0.5)/d$, $k = 1, \dots, d$ be a fine grid of equidistant points on which all functions and kernels of integral operators will be evaluated. Denote by $\mathbf{x}_i$ the d-dimensional vector of values of X_i at points t_k; this vector contains missing values on components corresponding to $t_k \in M_i$ while for $t_k \in O_i$, its values are obtained by evaluation of the spline representation of X_i. Denote by $\mathbf{X}$ the $(n \times d)$-dimensional data matrix with $\mathbf{x}_i$, $i = 1, \dots, n$ in rows.

The vector $\mathbf{m}$ of values of the mean function μ on the grid is estimated by $\hat{\mathbf{m}}$ equal to the vector of column means of $\mathbf{X}$ computed from available (not missing) data in each column. The covariance kernel ρ of the operator $\mathscr{R}$ evaluated on the grid corresponds to the $(d \times d)$-matrix $\mathbf{R}$ with entries $R_{kl} = \rho(t_k, t_l)$ and is estimated by the sample covariance matrix $\hat{\mathbf{R}}$ with entry $\hat{R}_{kl}$ computed from the data matrix $\mathbf{X}$ using all complete pairs of observations in columns k, l.

To estimate the eigenvalues and eigenfunctions, one performs eigen-decomposition of the matrix $\hat{\mathbf{R}}$. Denote $\Delta = 1/d$, the distance between the points of the grid. If the eigenvalues and eigenvectors of $\hat{\mathbf{R}}$ are $\hat{\kappa}_j$ and $\hat{\mathbf{u}}_j$, $j = 1, \dots, d$, then the eigenvalues of the operator $\hat{\mathscr{R}}$ are $\hat{\lambda}_j = \hat{\kappa}_j \Delta$ and the corresponding eigenfunctions $\hat{\varphi}_j$ evaluated on the grid are $\hat{\mathbf{f}}_j = \hat{\mathbf{u}}_j \Delta^{-1/2}$. The observed part $\hat{\beta}_{ijO_i} = \langle X_{iO_i} - \hat{\mu}_{O_i}, \hat{\varphi}_{jO_i} \rangle$ of the jth principal score of the ith curve is computed by numerical quadrature as $\hat{\beta}_{ijO_i} = \langle \mathbf{x}_{iO_i} - \hat{\mathbf{m}}_{O_i}, \hat{\mathbf{f}}_{jO_i} \rangle \Delta$, where the latter inner product is the usual inner product of vectors and the vectors with subscript O_i are subvectors of the original vectors consisting of elements with indices k such that $t_k \in O_i$.

Within the grid representation, the evaluation of an integral operator $\mathscr{B}$ in the sense of numerical integration corresponds to matrix multiplication: for a function h, $\mathscr{B}h$ is computed as $\mathbf{Bh}\Delta$, where the vector $\mathbf{h}$ and the matrix $\mathbf{B}$ are the values of h and of the kernel of $\mathscr{B}$ on the grid. From a purely computational point of view, even linear operators that have no integral representation may be represented by matrices. In particular, the identity operator $\mathscr{I}$ used in ridge regularisation is represented by the matrix $\mathbf{I}$ equal to the identity matrix divided by Δ; indeed, its value at $\mathbf{h}$ is $\mathbf{Ih}\Delta = \mathbf{h}$, thus it maps the argument on itself. The regularised operator $\hat{\mathscr{R}}^{(\alpha)}_{O_iO_i}$ is represented by the matrix $\hat{\mathbf{R}}^{(\alpha)}_{O_iO_i} = \hat{\mathbf{R}}_{O_iO_i} + \alpha \mathbf{I}_{O_i}$, where the subscript O_i denotes the submatrix corresponding to grid points in O_i. Analogously, the operators $\hat{\mathscr{R}}_{M_iM_i}$, $\hat{\mathscr{R}}_{M_iO_i}$ etc. are given by the corresponding submatrices of $\hat{\mathbf{R}}$. Then the matrix representation of the prediction operator $\hat{\mathscr{A}}^{(\alpha)}_i$ is computed as $\hat{\mathbf{A}}^{(\alpha)}_i = \hat{\mathbf{R}}_{O_iM_i} \hat{\mathbf{R}}^{(\alpha)-1}_{O_iO_i} \Delta^{-1}$. The regularised prediction of the missing part of the principal score and of the missing part of the trajectory can be computed as

$$\hat{\beta}^{(\alpha)}_{ij} = \langle \hat{\mathbf{A}}^{(\alpha)}_i (\mathbf{x}_{iO_i} - \hat{\mathbf{m}}_{O_i}) \Delta, \hat{\mathbf{f}}_{jM_i} \rangle \Delta, \quad \hat{\mathbf{x}}^{(\alpha)}_{iM_i} = \hat{\mathbf{A}}^{(\alpha)}_i (\mathbf{x}_{iO_i} - \hat{\mathbf{m}}_{O_i}) \Delta + \hat{\mathbf{m}}_{M_i}.$$

The covariance operator $\hat{\mathscr{V}}_i$ for the missing trajectory is obtained as

$$\hat{\mathbf{V}}_i = \hat{\mathbf{R}}_{M_iM_i} - \hat{\mathbf{A}}^{(\alpha)}_i \hat{\mathbf{R}}_{O_iO_i} \hat{\mathbf{A}}^{(\alpha)\mathrm{T}}_i \Delta^2$$

and the variance for the score is $\hat{v}^2_{ij} = \langle \hat{\mathbf{f}}_{jM_i}, \hat{\mathbf{V}}_i \hat{\mathbf{f}}_{jM_i} \rangle \Delta^2$.

The effective degrees of freedom can be computed directly using the series in (9) truncated at d terms, with the eigenvalues $\hat{\lambda}_{O_iO_ik}$ of $\hat{\mathscr{R}}_{O_iO_i}$ obtained from the eigenvalues of the matrix $\hat{\mathbf{R}}_{O_iO_i}$ like in the case of those of $\hat{\mathscr{R}}$ discussed above. Alternatively, one can use the matrix trace formula $\mathrm{trace}(\hat{\mathbf{R}}^{(\alpha)-1}_{O_iO_i} \hat{\mathbf{R}}_{O_iO_i} \Delta^{-1}) \Delta$. The computation of the residual sum of squares for scores

is straightforward; in the case of trajectories, the squared norms of functions are computed as the squared norms of vectors, multiplied by Δ.

The generalised cross-validation score can be minimised numerically by a Newton-type iterative procedure. In particular, we use the method "L-BFGS-B" available in the function *optim* in the R package (R Core Team, 2013). For the reliability of the optimisation procedure, we found it useful to scale the input parameters: the minimisation is run with $(\mathbf{x}_i - \mathbf{m})/s$ in place of $\mathbf{x}_i$ (and, consequently, with $\hat{\mathbf{R}}/s^2$ in place of $\hat{\mathbf{R}}$, $\hat{\lambda}_{O_iO_ij}/s^2$ in place of $\hat{\lambda}_{O_iO_ij}$ etc.); once the optimal value of α is found, it is multiplied by s^2 to return to the original scale and perform other computations with original data. The value $s^2 = \hat{\lambda}_{O_iO_i1}$ works well. The evaluation of the generalised cross-validation score can be unstable for very small values of α. Therefore, we run the minimisation routine with a lower limit for α, namely with $\alpha_0 = \max(\varepsilon^{1/2}, \alpha_*)$, where ε is the value of machine epsilon and α_* is such that the effective degrees of freedom equal $n/4$ (which is a reasonable upper bound for the number of free parameters). We initialise the iterative procedure with α equal to $\max(\bar{\lambda}_{O_iO_i}, \alpha_0)$ where $\bar{\lambda}_{O_iO_i}$ is the average of the eigenvalues $\hat{\lambda}_{O_iO_ij}$.

2. Proofs

2.1. Proof of Proposition 1

We use the notation $Z_i = X_i - \mu$.

For part (a), denote $\bar{\mu}(t) = J(t)\mu(t)$ and write

$$\mathsf{E}\,\|\hat{\mu} - \mu\|^2 \le \mathsf{E}(\|\hat{\mu} - \bar{\mu}\| + \|\bar{\mu} - \mu\|)^2 = \mathsf{E}\,\|\hat{\mu} - \bar{\mu}\|^2 + 2\,\mathsf{E}(\|\hat{\mu} - \bar{\mu}\|\|\bar{\mu} - \mu\|) + \mathsf{E}\,\|\bar{\mu} - \mu\|^2. \quad (1)$$

The first term on the right-hand side of (1) equals

$$\begin{aligned}\mathsf{E}\left\|\frac{J}{\sum_{i=1}^n O_i}\sum_{i=1}^n O_i Z_i\right\|^2 &= n^{-2}\int_0^1 \sum_{j=1}^n\sum_{k=1}^n \mathsf{E}\left(\frac{n^2 J(t)}{(\sum_{i=1}^n O_i(t))^2} O_j(t)Z_j(t)O_k(t)Z_k(t)\right)dt \\ &= n^{-2}\int_0^1 \sum_{j=1}^n \mathsf{E}\left(\frac{n^2 J(t)O_j(t)}{(\sum_{i=1}^n O_i(t))^2}\right)\mathsf{E}\,Z_j(t)^2 dt,\end{aligned}$$

where the last equality follows from the independence of $(O_1, \ldots, O_n)$ and $(Z_1, \ldots, Z_n)$, and from the independence of Z_j and Z_k for $j \ne k$. Rewrite the first expectation in the integrand as

$$\mathsf{E}\left(\frac{n^2 J(t)O_j(t)}{(\sum_{i=1}^n O_i(t))^2} 1_{[n^{-1}\sum_{i=1}^n O_i(t) > \delta_1]}\right) + \mathsf{E}\left(\frac{n^2 J(t)O_j(t)}{(\sum_{i=1}^n O_i(t))^2} 1_{[n^{-1}\sum_{i=1}^n O_i(t) \le \delta_1]}\right).$$

For all $t \in [0,1]$, the first summand is bounded from above by δ_1^{-2} while the second summand is dominated by $n^2 \sup_{t\in[0,1]} P(n^{-1}\sum_{i=1}^n O_i(t) \le \delta_1)$. Hence we see that

$$\mathsf{E}\,\|\hat{\mu} - \bar{\mu}\|^2 \le n^{-1}\left\{\delta_1^{-2} + n^2 \sup_{t\in[0,1]} P\left(n^{-1}\sum_{i=1}^n O_i(t) \le \delta_1\right)\right\}\mathsf{E}\,\|Z_1\|^2 = O(n^{-1}).$$

For the last term in (1), we obtain

$$\int_0^1 \mathsf{E}(J(t) - 1)\mu(t)^2 dt = \int_0^1 P\left(\sum_{i=1}^n O_i(t) = 0\right)\mu(t)^2 dt$$

$$\leq \sup_{t\in[0,1]} P\biggl(n^{-1}\sum_{i=1}^{n} O_i(t) \leq \delta_1\biggr)\|\mu\|^2$$
$$= O(n^{-2}).$$

The second term on the right-hand side of (1) is dominated by $2(\mathsf{E}\,\|\hat{\mu}-\bar{\mu}\|^2)^{1/2}(\mathsf{E}\,\|\bar{\mu}-\mu\|^2)^{1/2} \leq O(n^{-1})$. Putting these results together completes the proof of part (a).

The proof of part (b) is similar. Rewrite

$$\hat{\mathscr{R}} - \mathscr{R} = (\hat{\mathscr{R}} - \check{\mathscr{R}}) + (\check{\mathscr{R}} - \bar{\mathscr{R}}) + (\bar{\mathscr{R}} - \mathscr{R}), \tag{2}$$

where $\check{\mathscr{R}}$ and $\bar{\mathscr{R}}$ are integral operators with kernels

$$\check{\rho}(s,t) = \frac{I(s,t)}{\sum_{i=1}^{n} U_i(s,t)} \sum_{i=1}^{n} U_i(s,t) Z_i(s) Z_i(t),$$

and $\bar{\rho}(s,t) = I(s,t)r(s,t)$. The first term on the right-hand side of (2) reflects the effect of estimation of the mean. By direct computation, we see that

$$\mathsf{E}\,\|\hat{\mathscr{R}} - \check{\mathscr{R}}\|_2^2 = \mathsf{E}\int_{[0,1]^2} I(s,t)\{\hat{\mu}_{st}(s) - \mu(s)\}^2\{\hat{\mu}_{st}(t) - \mu(t)\}^2 ds dt$$
$$= \mathsf{E}\int_{[0,1]^2} \frac{I(s,t)}{(\sum_{i=1}^{n} U_i(s,t))^4} \biggl(\sum_{i=1}^{n} U_i(s,t) Z_i(s)\biggr)^2 \biggl(\sum_{i=1}^{n} U_i(s,t) Z_i(t)\biggr)^2 ds dt.$$

Developing the sums in the integrand and using the independence of the functions and observation indicators and the Cauchy–Schwarz inequality, we can show that the above quantity is dominated by

$$n^{-2}\int_{[0,1]^2} \mathsf{E}\biggl(\frac{n^2 I(s,t)}{(\sum_{i=1}^{n} U_i(s,t))^2}\biggr)\{(\mathsf{E}\,Z_1(s)^4\,\mathsf{E}\,Z_1(t)^4)^{1/2} + \rho(s,t)^2\} ds dt \leq O(n^{-2}),$$

where the last inequality is due to the fact that the first expectation in the integrand is bounded by $\delta_2^{-2} + n^2 \sup_{(s,t)\in[0,1]^2} P(n^{-1}\sum_{i=1}^{n} U_i(s,t) \leq \delta_2)$, which can be shown by manipulations similar to those in part (a). Next, analogously to part (a) we obtain for the second and third term on the right-hand side of (2) that

$$\mathsf{E}\,\|\check{\mathscr{R}} - \bar{\mathscr{R}}\|_2^2 \leq n^{-1}\biggl\{\delta_2^{-2} + n^2 \sup_{(s,t)\in[0,1]^2} P\biggl(n^{-1}\sum_{i=1}^{n} U_i(s,t) \leq \delta_2\biggr)\biggr\}\,\mathsf{E}\,\|Z_1 \otimes Z_1 - \mathscr{R}\|_2^2$$
$$= O(n^{-1})$$

(here $\otimes$ denotes the tensor product) and $\mathsf{E}\,\|\bar{\mathscr{R}} - \mathscr{R}\|_2^2 \leq O(n^{-2})$. Combining these bounds we obtain the assertion of part (b).

2.2. Proof of Proposition 2

Lemma 4.2 of Bosq (2000) and the inequality between the operator norm and Hilbert–Schmidt norm yield that $|\hat{\lambda}_j - \lambda_j| \leq \|\hat{\mathscr{R}} - \mathscr{R}\|_\infty \leq \|\hat{\mathscr{R}} - \mathscr{R}\|_2$ for all j. The first result then follows from part (b) of Proposition 1. For the second part, Lemma 4.3 of Bosq (2000) gives the inequality

$\|\hat{\varphi}_j - \hat{s}_j\varphi_j\| \leq a_j\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty$, where a_j is a constant depending on the eigenvalue spacings. Note that this lemma is formulated in Bosq (2000) for fully observed functions but an inspection of the proof shows that the inequality holds for any two compact linear operators in place of $\hat{\mathscr{R}}, \mathscr{R}$. This inequality, the dominance of the Hilbert–Schmidt norm over the operator norm and part (b) of Proposition 1 complete the proof.

2.3. Proof of Proposition 3

Rewrite

$$\hat{\beta}_{ijM_i}^{(\alpha_n)} - \beta_{ijM_i} = (\hat{\beta}_{ijM_i}^{(\alpha_n)} - \tilde{\beta}_{ijM_i}) + (\tilde{\beta}_{ijM_i} - \beta_{ijM_i})$$

and use Theorem 1 to obtain the first part of the proposition. Compute

$$v_{ij}^2 = \operatorname{var}(\tilde{\beta}_{ijM_i} - \beta_{ijM_i}) = \langle \varphi_{jM_i}, \mathscr{R}_{M_iM_i}\varphi_{jM_i}\rangle - \langle \varphi_{jM_i}, \mathscr{R}_{M_iO_i}\mathscr{R}_{O_iO_i}^{-1}\mathscr{R}_{O_iM_i}\varphi_{jM_i}\rangle.$$

The convergence in probability of $\langle \hat{\varphi}_{jM_i}, \hat{\mathscr{R}}_{M_iM_i}\hat{\varphi}_{jM_i}\rangle$ to $\langle \varphi_{jM_i}, \mathscr{R}_{M_iM_i}\varphi_{jM_i}\rangle$ is a direct consequence of Propositions 1 and 2. The last term in the expression for v_{ij}^2 and the corresponding term in the estimator $\hat{v}_{ij}^2$ equal $\langle \tilde{a}_{ij}, \mathscr{R}_{O_iO_i}\tilde{a}_{ij}\rangle$, $\langle \hat{a}_{ij}^{(\alpha_n)}, \hat{\mathscr{R}}_{O_iO_i}\hat{a}_{ij}^{(\alpha_n)}\rangle$, respectively. In their difference

$$\langle \hat{a}_{ij}^{(\alpha_n)}, (\hat{\mathscr{R}}_{O_iO_i} - \mathscr{R}_{O_iO_i})\hat{a}_{ij}^{(\alpha_n)}\rangle + (\langle \hat{a}_{ij}^{(\alpha_n)}, \mathscr{R}_{O_iO_i}\hat{a}_{ij}^{(\alpha_n)}\rangle - \langle \tilde{a}_{ij}, \mathscr{R}_{O_iO_i}\tilde{a}_{ij}\rangle),$$

the convergence of the second term to zero was shown in the proof of Theorem 1. For the first term we compute

$$\begin{aligned} |\langle \hat{a}_{ij}^{(\alpha_n)}, (\hat{\mathscr{R}}_{O_iO_i} - \mathscr{R}_{O_iO_i})\hat{a}_{ij}^{(\alpha_n)}\rangle| &\leq \|\hat{\mathscr{R}}_{O_iO_i} - \mathscr{R}_{O_iO_i}\|_\infty \|\hat{a}_{ij}^{(\alpha_n)}\|^2 \\ &\leq O_P(n^{-1/2})\alpha_n^{-2}\|\hat{\mathscr{R}}_{O_iM_i}\|_\infty^2 \\ &\to 0. \end{aligned}$$

This completes the proof of the consistency of $\hat{v}_{ij}^2$. The remaining assertions are obvious.

2.4. Proof of Proposition 4

We can rewrite $\hat{X}_{iM_i}^{(\alpha_n)} - X_{iM_i} = (\hat{X}_{iM_i}^{(\alpha_n)} - \tilde{X}_{iM_i}) + (\tilde{X}_{iM_i} - X_{iM_i})$. Due to Theorem 2, the L^2-norm of the first term on the right-hand side converges to 0 in probability. The second term is the limiting stochastic process. The consistency of the covariance estimator can be proven like in the proof of Proposition 3. The assertion for the Gaussian case follows immediately from the fact that the limiting process is a linear function of X_i.

References

Bosq, D. (2000). *Linear Processes in Function Spaces*. Springer, New York.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York.