Selecting the Number of Functional Principal Components

For selecting the number K of functional principal components (FPCs) in the **PACE** program, one may use the input argument *selection_k* in the function **setOptions()**.

1 Automatic Selection of Number K of Functional Principal Components

(1) Pseudo-AIC or Pseudo-BIC criteria:

Let Y_{ij} be the jth observation of the random function $X_i(\cdot)$, made at a random time T_{ij} and ε_{ij} the additional measurement errors that are assumed to be i.i.d. and independent of the random coefficients ξ_{ik}, where $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, $k = 1, 2, \ldots$. Then the model we consider is

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij} = \mu(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(T_{ij}) + \varepsilon_{ij}, \;\; T_{ij} \in \mathcal{T},$$

where $\mathrm{E}\varepsilon_{ij} = 0$, $\mathrm{var}(\varepsilon_{ij}) = \sigma^2$.

Write $\widetilde{\mathbf{X}}_i = (X_i(T_{i1}), \ldots, X_i(T_{in_i}))^T$, $\widetilde{\mathbf{Y}}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, $\boldsymbol{\mu}_i = (\mu(T_{i1}), \ldots, \mu(T_{in_i}))^T$, $\boldsymbol{\phi}_{ik} = (\phi_k(T_{i1}), \ldots, \phi_k(T_{in_i}))^T$.

(i) Estimated marginal pseudo-Gaussian log-likelihood of $\widetilde{\mathbf{Y}}_i$:

$$\hat{L}_1 = \sum_{i=1}^{n} \left\{ -\frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log(\det\hat{\Sigma}_{Y_i}) - \frac{1}{2}(\widetilde{\mathbf{Y}}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\Sigma}_{Y_i}^{-1}(\widetilde{\mathbf{Y}}_i - \hat{\boldsymbol{\mu}}_i) \right\}$$

where the (j, l) element of $(\hat{\Sigma}_{Y_i})_{j,l} = \hat{G}(T_{ij}, T_{il}) + \hat{\sigma}^2\delta_{jl}$ and $\hat{G}(T_{ij}, T_{il}) = \sum_{k=1}^{K} \hat{\lambda}_k \hat{\phi}_k(T_{ij})\hat{\phi}_k(T_{il})$.

The marginal criteria are $AIC_1(K) = -2\hat{L}_1 + 2K$ and $BIC_1(K) = -2\hat{L}_1 + K\log(N)$, where $N = \sum_{i=1}^{n} n_i$, and the marginal choices are the minimizers over K.

In the **PACE** program, these choices are obtained via *selection_k* = 'AIC1' and *selection_k* = 'BIC1', respectively.

(ii) Estimated conditional pseudo-Gaussian log-likelihood of $\widetilde{\mathbf{Y}}_i|\hat{\xi}_{ik}$:

$$\hat{L}_2 = \sum_{i=1}^{n} \left\{ -\frac{n_i}{2}\log(2\pi) - \frac{n_i}{2}\log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}(\widetilde{\mathbf{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K}\hat{\xi}_{ik}\hat{\boldsymbol{\phi}}_{ik})^T(\widetilde{\mathbf{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K}\hat{\xi}_{ik}\hat{\boldsymbol{\phi}}_{ik}) \right\}$$

The corresponding conditional criteria are $AIC_2(K) = -2\hat{L}_2 + 2K$ and $BIC_2(K) = -2\hat{L}_2 + K\log(N)$, where $N = \sum_{i=1}^{n} n_i$, and the corresponding conditional choices are obtained as the minimizing values of K, obtained via *selection_k* = 'AIC2' and *selection_k* = 'BIC2', respectively.

Note: The marginal choices often lead to selections of smaller values of K and are often preferrable in practical applications.

(2) Fraction of Variance Explained (FVE):

The fraction of variance explained (FVE) is calculated as

$$FVE_J = \frac{\sum_{k=1}^{J} \hat{\lambda}_k}{\sum_{k=1}^{ngrid} \hat{\lambda}_k}, \quad J = 1, \ldots, ngrid$$

where $ngrid \times ngrid$ is the dimension of the discretized smoothed covariance matrix which serves as input for the numerical matrix spectral decomposition from which FPCS and eigenvalues are derived. When *selection_k* = 'FVE', the number K of included FPCs corresponds to the smallest J for which $FVE_J >$ *FVE_threshold*. The default for *FVE_threshold* is *FVE_threshold*=0.85. This setting provides the overall default for selecting K as it is fast and usually yields reasonable results. The return value FVE is the array of values of FVE_J.

2 User-defined Number of Included Functional Principal Components

Simply set *selection_k* to a positive integer that is no greater than *ngrid*.